REVISTA DE ESTUDOS DA LINGUAGEM

Faculdade de Letras da UFMG

ISSN
Impresso: 0104-0588
On-line: 2237-2083

Rev. Estudos da Linguagem	Belo Horizonte	v. 29	n. 2	p. 647-1638	abr./jun. 2021

REVISTA DE ESTUDOS DA LINGUAGEM

Universidade Federal de Minas Gerais
REITORA: Sandra Regina Goulart Almeida
VICE-REITOR: Alessandro Fernandes Moreira

Faculdade de Letras
DIRETORA: Graciela Inés Ravetti de Gómez
VICE-DIRETORA: Sueli Maria Coelho

Editor-chefe
Gustavo Ximenes Cunha (UFMG)

Editores convidados
Stella Esther Ortweiler Tagnin (USP)
Maria José Bocorny Finatto (UFRGS)
Guilherme Fromm (UFU)

Revisão e Normalização
Alda Lopes Durães Ribeiro
Gustavo Ximenes Cunha
Jairo Venício Carvalhais Oliveira

Secretaria
Gustavo Ximenes Cunha (UFMG)

Editoras-associadas
Ana Larissa Adorno Maciotto Oliveira (UFMG)
Carla Viana Coscarelli (UFMG)
Helcira Maria Rodrigues de Lima (UFMG)

Revisão de Língua Inglesa
Ana Larissa Adorno Marciotto Oliveira (UFMG)
Junia de Carvalho Fidelis Braga (UFMG)
Mara Passos Guimarães (UFMG)
Marisa Mendonça Carneiro (UFMG)

Editoração eletrônica
Alda Lopes Durães Ribeiro

REVISTA DE ESTUDOS DA LINGUAGEM, v.1 - 1992 - Belo Horizonte, MG, Faculdade de Letras da UFMG

Histórico:
1992 ano 1, n.1 (jul/dez)
1993 ano 2, n.2 (jan/jun)
1994 Publicação interrompida
1995 ano 4, n.3 (jan/jun); ano 4, n.3, v.2 (jul/dez)
1996 ano 5, n.4, v.1 (jan/jun); ano 5, n.4, v.2; ano 5, n. esp.
1997 ano 6, n.5, v.1 (jan/jun)

Nova Numeração:
1997 v.6, n.2 (jul/dez)
1998 v.7, n.1 (jan/jun)
1998 v.7, n.2 (jul/dez)

1. Linguagem - Periódicos I. Faculdade de Letras da UFMG, Ed.

CDD: 401.05

ISSN: Impresso: 0104-0588
On-line: 2237-2083

REVISTA DE ESTUDOS DA LINGUAGEM

V. 29 - Nº 2 - abr.-jun. 2021

Indexadores

Diadorim [Brazil]

DOAJ (Directory of Open Access Journals) [Sweden]

DRJI (Directory of Research Journals Indexing) [India]

EBSCO [USA]

JournalSeek [USA]

Latindex [Mexico]

Linguistics & Language Behavior Abstracts [USA]

MIAR (Matriu d'Informació per a l'Anàlisi de Revistes) [Spain]

MLA Bibliography [USA]

OAJI (Open Academic Journals Index) [Russian Federation]

Portal CAPES [Brazil]

REDIB (Red Iberoamericana de Innovación y Conocimiento Científico) [Spain]

Sindex (Sientific Indexing Services) [USA]

Web of Science [USA]

WorldCat / OCLC (Online Computer Library Center) [USA]

ZDB (Elektronische Zeitschriftenbibliothek) [Germany]

REVISTA DE ESTUDOS DA LINGUAGEM

Editor-chefe

Gustavo Ximenes Cunha (UFMG, Belo Horizonte/MG, Brasil)

Editoras-associadas

Ana Larissa Adorno Maciotto Oliveira (UFMG, Belo Horizonte/MG, Brasil)
Carla Viana Coscarelli (UFMG, Belo Horizonte/MG, Brasil)
Helcira Maria Rodrigues de Lima (UFMG, Belo Horizonte/MG, Brasil)

Conselho Editorial

Alejandra Vitale (UBA, Ciudad Autónoma de Buenos Aires, Argentina)
Didier Demolin (Université de la Sorbonne Nouvelle Paris 3, Paris, França)
Ieda Maria Alves (USP, São Paulo/SP, Brasil)
Jairo Nunes (USP, São Paulo/SP, Brasil)
Scott Schwenter (OSU, Columbus, Ohio, Estados Unidos)
Shlomo Izre'el (TAU, Tel Aviv, Israel)
Stefan Gries (UCSB, Santa Barbara/CA, Estados Unidos)
Teresa Lino (NOVA, Lisboa, Portugal)
Tjerk Hagemeijer (ULisboa, Lisboa, Portugal)

Comissão Científica

Aderlande Pereira Ferraz (UFMG, Belo Horizonte/MG, Brasil)
Alessandro Panunzi (Unifi, Florença, Itália)
Alina M. S. M. Villalva (ULisboa, Lisboa, Portugal)
Aline Alves Ferreira (UCSB, Santa Barbara/CA, Estados Unidos)
Ana Lúcia de Paula Müller (USP, São Paulo/SP, Brasil)
Ana Maria Carvalho (UA, Tucson/AZ, Estados Unidos)
Ana Paula Scher (USP, São Paulo/SP, Brasil)
Anabela Rato (U of T, Toronto/ON, Canadá)
Aparecida de Araújo Oliveira (UFV, Viçosa/MG, Brasil)
Aquiles Tescari Neto (UNICAMP, Campinas/SP, Brasil)
Augusto Soares da Silva (UCP, Braga, Portugal)
Beth Brait (PUC-SP/USP, São Paulo/SP, Brasil)
Bruno Neves Rati de Melo Rocha (UFMG, Belo Horizonte/MG, Brasil)
Carmen Lucia Barreto Matzenauer (UCPEL, Pelotas/RS, Brasil)
Celso Ferrarezi (UNIFAL, Alfenas/MG, Brasil)
César Nardelli Cambraia (UFMG, Belo Horizonte/MG, Brasil)
Cristina Name (UFJF, Juiz de Fora/MG, Brasil)
Charlotte C. Galves (UNICAMP, Campinas/SP, Brasil)
Deise Prina Dutra (UFMG, Belo Horizonte/MG, Brasil)
Diana Luz Pessoa de Barros (USP/UPM, São Paulo/SP, Brasil)
Edwiges Morato (UNICAMP, Campinas/SP, Brasil)
Emília Mendes Lopes (UFMG, Belo Horizonte/MG, Brasil)
Esmeralda V. Negrão (USP, São Paulo/SP, Brasil)
Flávia Azeredo Cerqueira (JHU, Baltimore/MD, Estados Unidos)
Gabriel de Avila Othero (UFRGS, Porto Alegre/RS, Brasil)
Gerardo Augusto Lorenzino (TU, Filadélfia/PA, Estados Unidos)
Glaucia Muniz Proença de Lara (UFMG, Belo Horizonte/MG, Brasil)
Hanna Batoréo (UAb, Lisboa, Portugal)
Heliana Ribeiro de Mello (UFMG, Belo Horizonte/MG, Brasil)
Heronides Moura (UFSC, Florianópolis/SC, Brasil)
Hilario Bohn (UCPEL, Pelotas/RS, Brasil)
Hugo Mari (PUC-Minas, Belo Horizonte/MG, Brasil)
Ida Lucia Machado (UFMG, Belo Horizonte/MG, Brasil)
Ieda Maria Alves (USP, São Paulo/SP, Brasil)
Ivã Carlos Lopes (USP, São Paulo/SP, Brasil)
Jairo Nunes (USP, São Paulo/SP, Brasil)

Jairo Venício Carvalhais Oliveira (UFMG, Belo Horizonte/MG, Brasil)
Jean Cristtus Portela (UNESP-Araraquara, Araraquara/SP, Brasil)
João Antônio de Moraes (UFRJ, Rio de Janeiro/ RJ, Brasil)
João Miguel Marques da Costa (Universidade Nova da Lisboa, Lisboa, Portugal)
João Queiroz (UFJF, Juiz de Fora/MG, Brasil)
José Magalhaes (UFU, Uberlândia/MG, Brasil)
João Saramago (Universidade de Lisboa, Lisboa, Portugal)
José Borges Neto (UFPR, Curitiba/PR, Brasil)
Kanavillil Rajagopalan (UNICAMP, Campinas/SP, Brasil)
Laura Alvarez Lopez (Universidade de Estocolmo, Stockholm, Suécia)
Leo Wetzels (Free Univ. of Amsterdam, Amsterdã, Holanda)
Laurent Filliettaz (Université de Genève, Genebra, Suiça)
Leonel Figueiredo de Alencar (UFC, Fortaleza/CE, Brasil)
Livia Oushiro (UNICAMP, Campinas/SP, Brasil)
Lodenir Becker Karnopp (UFRGS, Porto Alegre/RS, Brasil)
Lorenzo Teixeira Vitral (UFMG, Belo Horizonte/MG, Brasil)
Luiz Amaral (UMass Amherst, Amherst/MA, Estados Unidos)
Luiz Carlos Cagliari (UNESP, São Paulo/SP, Brasil)
Luiz Carlos Travaglia (UFU, Uberlândia/MG, Brasil)
Marcelo Barra Ferreira (USP, São Paulo/SP, Brasil)
Marcia Cançado (UFMG, Belo Horizonte/MG, Brasil)
Márcio Leitão (UFPB, João Pessoa/PB, Brasil)
Marcus Maia (UFRJ, Rio de Janeiro/RJ, Brasil)
Maria Bernadete Marques Abaurre (UNICAMP, Campinas/SP, Brasil)
Maria Cecília Camargo Magalhães (PUC-SP, São Paulo/SP, Brasil)
Maria Cecília Magalhães Mollica (UFRJ, Rio de Janeiro/RJ, Brasil)
Maria Cândida Trindade Costa de Seabra (UFMG, Belo Horizonte/MG, Brasil)
Maria Cristina Figueiredo Silva (UFPR, Curitiba/PR, Brasil)
Maria Luíza Braga (PUC/RJ, Rio de Janeiro/RJ, Brasil)
Maria Marta P. Scherre (UNB, Brasília/DF, Brasil)
Micheline Mattedi Tomazi (UFES, Vitória/ES, Brasil)
Miguel Oliveira, Jr. (UFAL, Maceió, Alagoas, Brasil)
Monica Santos de Souza Melo (UFV, Viçosa/MG, Brasil)
Patricia Matos Amaral (UI, Bloomington/IN, Estados Unidos)
Paulo Roberto Gonçalves Segundo (USP, São Paulo/SP, Brasil)
Philippe Martin (Université Paris 7, Paris, França)
Rafael Nonato (Museu Nacional-UFRJ, Rio de Janeiro/RJ, Brasil)
Raquel Meister Ko. Freitag (UFS, Aracaju/SE, Brasil)

Roberto de Almeida (Concordia University, Montreal/QC, Canadá)
Ronice Müller de Quadros (UFSC, Florianópolis/SC, Brasil)
Ronald Beline (USP, São Paulo/SP, Brasil)
Rove Chishman (UNISINOS, São Leopoldo/RS, Brasil)
Sanderléia Longhin-Thomazi (UNESP, São Paulo/SP, Brasil)
Sergio de Moura Menuzzi (UFRGS, Porto Alegre/RS, Brasil)
Seung- Hwa Lee (UFMG, Belo Horizonte/MG, Brasil)
Sírio Possenti (UNICAMP, Campinas/SP, Brasil)
Suzi Lima (U of T / UFRJ, Toronto/ON - Rio de Janeiro/RJ, Brasil)
Thais Cristofaro Alves da Silva (UFMG, Belo Horizonte/MG, Brasil)
Tommaso Raso (UFMG, Belo Horizonte/MG-Brasil)
Tony Berber Sardinha (PUC-SP, São Paulo/SP, Brasil)
Ubiratã Kickhöfel Alves (UFRGS, Porto Alegre/RS, Brasil)
Vander Viana (University of Stirling, Stirling/Sld, Reino Unido)
Vanise Gomes de Medeiros (UFF, Niterói/RJ, Brasil)
Vera Lucia Lopes Cristovao (UEL, Londrina/PR, Brasil)
Vera Menezes (UFMG, Belo Horizonte/MG, Brasil)
Vilson José Leffa (UCPel, Pelotas/RS, Brasil)

Sumário / Contents

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 661-671, 2021

CC BY

Linguística de Corpus: conquistas e desafios

Corpus Linguistics: achievements and challenges

Stella Esther Ortweiller Tagnin
Universidade de São Paulo (USP), São Paulo, São Paulo / Brasil
seotagni@usp.br
https://orcid.org/0000-0002-5517-2710

Maria José Bocorny Finatto
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul / Brasil
mariafinatto@gmail.com
https://orcid.org/0000-0002-6022-8408

Guilherme Fromm
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil
guifromm@ufu.br
https://orcid.org/0000-0001-5654-0135

Quando pensamos em quase todas as áreas de descrição linguística e na Linguística Aplicada, no século XXI, a abordagem/metodologia da Linguística de *Corpus* (doravante LC) é peça central no desenvolvimento de vários estudos. Afinal, ela se propõe como uma abordagem empírica, na qual os *corpora* estão disponíveis para validações dessas pesquisas e futuros desdobramentos das mesmas. Os estudos baseados em *corpora* (e toda a tecnologia envolvida neles) não param de se multiplicar e de solidificar esta tendência de empirismo e verificação de dados reais de língua. Entre esses estudos, temos os que lidam com *corpora* de forma indireta (*corpus based*), como comprovação e exemplificação de pressuposições levantadas pelo pesquisador. E há também os de forma direta (*corpus driven*), nos quais o próprio *corpus* serve de ponto de

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.661-671

partida para o pesquisador levantar hipóteses, muitas nunca cogitadas. Já são quase sessenta anos (tomando como base a publicação do primeiro grande *corpus*, *Brown Corpus*, em 1964) que trabalhos baseados em *corpora* vêm se destacando na grande área da Linguística.

De parcos *corpora* e trabalhos neles baseados até os mega-*corpora* (na base de bilhões de palavras) disponíveis na Internet e a infinidade de pesquisas que deles podem ser geradas, esta edição da RELIN propôs escrutinar passado, presente e futuro desta abordagem e metodologia (ou teoria, para alguns) que trabalha com dados empíricos advindos de exemplos reais do uso da linguagem. Como resultado, nosso rol de textos engloba várias áreas do conhecimento linguístico: o relacionamento de teorias linguísticas com a metodologia da LC, estudos sobre fraseologismos, tradução, gramática, linguagens de especialidades e ensino. Além dessas áreas, destacamos os textos que trabalham com a própria metodologia em si (especialmente quanto à compilação de *corpora*) e os textos que trabalham a interface da LC com a Linguística Computacional e o Processamento de Língua/Linguagem Natural (PLN).

Comecemos, então, pelos textos que nos remetem à própria compilação e exploração de variados *corpora*. Com o enorme número de textos disponíveis na internet, seria de se acreditar que compilar um *corpus*, de qualquer área, fosse tarefa fácil. Ledo engano. É justamente essa profusão de material que demanda do pesquisador uma criteriosa seleção, após a qual os textos ainda terão de passar por vários procedimentos antes de poderem ser investigados por ferramentas computacionais de análise linguística. Os artigos que se seguem comprovam e ilustram os percalços inerentes à compilação de *corpora* especializados.

Damasceno-Morais, em "O carpinteiro e a madeira: a constituição de *corpora* jurídicos em perspectiva etnometodológica", discute a construção de um *corpus* cuja finalidade é descrever e analisar como desembargadores num tribunal brasileiro de Segunda Instância atuam numa situação argumentativa. O *corpus*, denominado TRIBUNAL, é constituído a partir de áudios com exemplos reais de uso da linguagem empregada em deliberações de magistrados em processos de danos morais. Embora o foco do artigo seja a descrição do processo de compilação do *corpus*, e não a análise dos dados, que foram a essência de sua tese de doutorado apresentada na França, o autor já adianta que foi possível confirmar que o discurso jurídico está longe de ser frio e asséptico.

Em "Diseño de *corpus* específicos para el estudio histórico gramatical: el caso de las construcciones con clítico feminino", Arellano preconiza o uso de um *corpus* específico para o estudo histórico do clítico la na variante argentina do espanhol. Após discutir as desvantagens de utilizar *corpora* gerais para esse estudo e apontar resultados não satisfatórios obtidos com o CORDE, contrasta-os com uma pesquisa feita com um *corpus* composto por peças teatrais argentinas extraídas de uma antologia que abarca a totalidade do século XIX e a primeira metade do século XX. Essa escolha baseia-se no pressuposto de que o teatro melhor retrata a linguagem oral. Após evidenciar suas vantagens, conclui salientando outras áreas que podem se beneficiar desse tipo de *corpus*, não só para estudos históricos como também sincrônicos.

O objetivo precípuo da pesquisa relatada por Waquil, "Um *corpus* de Estudos de Gênero: por quê, como e para quê?", ainda em andamento, é construir um glossário com termos do campo de Estudos de Gênero no Brasil, com o intuito de contribuir para uma comunicação especializada mais precisa na área. Para isso, alia os pressupostos da Teoria Comunicativa da Terminologia com a metodologia da Linguística de *Corpus*. O artigo, conforme o título anuncia, detém-se na justificativa da construção de um *corpus* especializado (por quê), composto por artigos de duas publicações representativas da área (como) para construir um glossário com termos e contextos definitórios (para quê).

Araújo, em "Construindo *corpora* bilíngues quimbundo-português-quimbundo", debruça-se sobre a problemática da construção de *corpora* bilíngues que contemplem o quimbundo do Libolo, uma região do Kwansa Sul, e o português angolano a fim de melhor documentar fenômenos de seu reconhecido contato linguístico. Metodologicamente, faz uma aproximação entre a Linguística de *Corpus* e a Linguística Africana e propõe a construção de *corpora* escritos e orais, bilíngues e/ou paralelos. Para os *corpora* escritos, sugere a plataforma do *Corpus* Tycho Brahe e para os de fala o método de transcrição proposto pelo projeto C-Oral-Angola. O autor apresenta vários exemplos de falas e discute problemas de transcrição desses *corpora*, em especial em relação à identificação dos fenômenos de *codeswitching* e empréstimos, e argumenta que esses *corpora* poderão retratar a real situação de contato dessas línguas, além de fornecer dados objetivos para embasar a hipótese de um *continuum* afro-brasileiro do português.

Já Quadros e Sousa apresentam, em "Brazilian Sign Language *corpus*: Acre Libras Inventory", a proposta teórico-metodológica de um *corpus* de Língua Brasileira de Sinais (Libras) a ser desenvolvido no âmbito do projeto Inventário de Língua Brasileira de Sinais na Região do município de Rio Branco, no estado do Acre. O projeto faz parte do Inventário Nacional da Diversidade Linguística, que cataloga as línguas faladas no Brasil visando disponibilizar informações sobre seu patrimônio linguístico e prover políticas para a proteção dessas línguas. Os autores detalham a metodologia utilizada para a coleta, transcrição e análise dos dados. Detêm-se, em especial, na coleta dos dados, que implica em registrar uma língua baseada numa complexa produção de gestos, olhares, movimentos corporais e outros não encontrados em textos escritos. Os informantes são selecionados de acordo com rígidos critérios para garantir a autenticidade dos dados. Participam de um diálogo em Libras, conduzido por um pesquisador, gravado com quatro câmeras para proporcionar diferentes perspectivas. O processo de anotação do *corpus* e suas dificuldades, assim como a organização dos dados e a disponibilização *on-line* também são discutidos. Os autores esperam que a sistematização do processo de criação do *corpus* possa contribuir para consolidar a teoria e a prática em relação à língua de sinais no Brasil.

Este número da revista traz um interessante bloco de artigos que trata da correlação entre o Processamento da Linguagem Natural (PLN) e a Linguística de *Corpus* (LC). São trabalhos que lidam com a tradução automática, com temas como a sumarização automática de diferentes documentos, com a descrição gramatical e de gêneros textuais. E, mesmo em tempos em que a obtenção de um *corpus* – entendido apenas como uma coleção de documentos em formato digital - tornou-se algo muito facilitado por conta da Internet, temos um artigo que, justamente, destaca aspectos envolvidos no esforço desse empreendimento.

Nesse quadro, o artigo "Using machine translator as a pedagogical resource in English for specific purposes courses in the academic context", de Borsatti e Riess, traz uma proposta de uso pedagógico de tradução automática, feita por ferramentas *on-line*, em cursos de inglês para fins específicos. A intenção é avaliar a eficiência dessa tecnologia como suporte para a leitura de textos científicos em inglês como L2/ LE. Assim, a ferramenta Google Translate é posta à prova como um recurso que pode ser bastante útil.

Por sua vez, o artigo "An investigation of linguistic problems in automatic multi-document summaries", de Dias, Di Felippo, Rassi, Cardoso, Nóbrega e Pardo, trata de problemas de sumários – ou sínteses - de textos gerados automaticamente, salientando-se que esses sumários partem de diferentes documentos acerca de um tema comum. Os autores ponderam que a anotação manual de extratos tende a gerar subsídios para as tarefas automáticas de detecção e correção de problemas linguísticos com vistas à produção desses sumários. Esse procedimento pode torná-los não só mais informativos, isto é, com maior cobertura do conteúdo do material de origem, como também dotá-los de melhor estruturação linguística.

Na sequência, o trabalho "Procedimentos para construção do *Corpus* da Computação da Língua Inglesa (CoCLI) e cálculo do esforço na construção manual de *corpora*", de Oliveira, descreve os procedimentos metodológicos da pesquisa intitulada "ToGatherUp: um protótipo de ferramenta para a construção de *corpora*". A proposta do trabalho é verificar em que medida recursos como esse podem reduzir o tempo e o esforço despendidos pelo pesquisador em projetos de elaboração manual de *corpora*. Esse é um artigo que toca em um ponto fundamental e que contribui para todos os que lidam com a compilação de *corpus*.

O artigo "Inteligibilidade e convencionalidade em textos de divulgação da área médica em português brasileiro", de Carvalho e Rebechi, faz um cotejo de dados indicativos de inteligibilidade e de convencionalidade em textos de divulgação da área médica em português. O propósito é verificar a potencial adequação desses textos ao público brasileiro. São examinados textos escritos originalmente em inglês e suas traduções para o português, reunidos em um *corpus* comparável. As autoras verificaram que tanto os textos escritos originalmente em português quanto aqueles traduzidos não se mostram totalmente adequados para o leitor-alvo brasileiro de textos de divulgação médica. As autoras ponderam que a quebra da convencionalidade, identificada nos textos traduzidos, pode dificultar ainda mais a compreensão do leitor médio.

O artigo "Propriedades linguísticas da redação do Enem: uma análise computacional", de Bertucci, descreve algumas propriedades linguísticas recorrentes em textos nota 1000 do Enem. Foram examinadas 95 redações que alcançaram a nota máxima nos anos de 2014, 2018 e 2019, com auxílio do *software* Tropes, uma ferramenta computacional

de análise lexical que verifica as recorrências de categorias e repertório vocabular. O trabalho pretende contribuir com os estudos de *corpora*, dedicados aos gêneros escolares e com o ensino, prestando a importante contribuição de apresentar um novo *software* para a nossa comunidade de estudos.

No artigo "Sujeito oculto às claras: uma abordagem descritivo-computacional", Freitas e Souza nos trazem estudos descritivos e computacionais relacionados ao tema do sujeito oculto. A proposta do trabalho é identificar esses sujeitos, retirá-los dos textos e reintroduzi-los, verificando-se o impacto disso sobre a estruturação sintática e o reconhecimento automático em analisadores automáticos de linguagem. Esse trabalho mostra um caminho interessante e original para apoiar a descrição desses elementos gramaticais.

Por fim, nesse bloco, o artigo "O papel do *corpus* de estudo no aprimoramento descritivo da complementaridade informacional multidocumento", de Souza, pretende reconstruir um percurso metodológico no que se refere ao estudo em *corpus* das relações CST (*Cross-document Structure Theory*), que lida com textos jornalísticos do Português. O autor nos apresenta o *corpus* CSTNews, um conjunto de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST. A partir do seu ensaio, o autor nos mostra como avançar nesse tipo de estudo, que lida com diferentes textos escritos sobre um mesmo tema/tópico, buscando alternativas para representar o seu conteúdo a partir do processamento automático.

Nosso número apresenta dois artigos que trabalham a relação *corpora*/teoria. Embora a LC não seja uma metodologia que funcione com absolutamente todas as teorias linguísticas (como o Gerativismo, por exemplo), é importante destacar que a mesma pode funcionar bem com teorias consideradas antigas, não só com as mais novas, que já pressupõem a LC como base (que é o caso da Linguística Sistêmica-Funcional). Além de ser aplicado como abordagem/metodologia, o uso de *corpora* também gera novas perspectivas e novas teorias derivadas, como vemos nos textos a seguir.

Em "Pragmática de *Corpus*: o que é e onde estamos", Santos e Miranda, conforme o título já evidencia, apresentam uma nova área de estudos. Partindo de um histórico da Linguística de *Corpus* e da Pragmática, discutem como a intersecção dessas duas áreas fez emergir a Pragmática de *Corpus* e como uma se beneficia da outra. Uma revisão

da literatura discute seus aspectos teórico-metodológicos e seus desafios. Os autores introduzem as abordagens forma-função e função-forma e investigam, a título de exemplificação, com base em dois *corpora* de fala, as funções do marcador pragmático *kind of*, no discurso oral de brasileiros universitários no Brasil e na Irlanda, utilizando a abordagem forma-função.

Seguindo a ideia de artigos que trabalham com teorias linguísticas, o texto "Linguística de *Corpus* aplicada à Semântica de Frames: investigando conceptualizações pró-escolha no debate da Sugestão Legislativa n.º 15/2014", de Santos e Chishman, traça um paralelo entre a Semântica de Frames, de Fillmore, e a metodologia da Linguística de *Corpus*. Com um *corpus* composto por textos retirados de transcrições de audiências públicas (numa discussão sobre a validade do aborto), as autoras analisam o material coletado através de um programa computacional (QSR NVivo, com direcionamento *top-down*) e uma plataforma *on-line* para análise de *corpora* (Sketch Engine, com direcionamento *bottom-up*), com o intuito de investigar redes de significado na confluência entre uma teoria e uma metodologia. Como resultado, são apresentados modelos de *frames* (retirados das transcrições) dos defensores da proposta, que contêm: uma definição do frame, seus elementos (com as respectivas definições), os termos evocadores e exemplos retirados do *corpus*.

Toda uma área de análise linguística voltada para o trabalho com unidades de sentido que ultrapassam uma única palavra tem um grande destaque no século XXI: a Fraseologia. Do encontro frequente (ou seja, regular) de palavras gramaticais e lexicais, essa área engloba estudos desde unidades compostas por duas palavras (os chamados bigramas) até textos inteiros (como uma fábula). Três textos deste número trabalham com essa abordagem de análise.

O artigo de Novodvorski e Bevilacqua, "*De marcar la cancha a una canchereada* na metaforização da política pelo futebol: análise de unidades fraseológicas especializadas em *corpus* jornalístico argentino", apresenta uma análise de termos e unidades fraseológicas especializadas, do âmbito do futebol, em processos de metaforização com o domínio alvo da política. A partir de um *corpus* jornalístico monolíngue em espanhol rio-platense, da coluna Humor Político do jornal argentino Clarín, os autores identificaram construções que atestam a metaforização da política (um campo abstrato) pelo futebol (um campo mais concreto).

A dificuldade de usar colocações em artigos acadêmicos escritos em língua inglesa é o tópico de "Analysing the behaviour of academic collocations in a *corpus* of research-papers: a data-driven study", de Pinto, Camargo, Serpa e Silva. A investigação baseia-se no *corpus* Brazilian Academic *Corpus* of English (BrACE) composto por artigos científicos extraídos da plataforma SciELO, escritos por pesquisadores brasileiros, totalizando 906.000 palavras em oito áreas. As colocações usadas com maior frequência pelos pesquisadores brasileiros foram comparadas com as de maior frequência em artigos escritos por autores anglófonos. Para essa comparação, foram usadas três conhecidas listas de colocações acadêmicas e um *corpus* de inglês acadêmico, o *The Oxford Corpus of Academic English* (OCAE), com 71.372.972 palavras. Os resultados apontaram principalmente um subuso dessas colocações pelos pesquisadores brasileiros. As autoras concluem discutindo a validade de fazer generalizações a partir de seus dados, em vista do reduzido tamanho de seu *corpus,* mas contemplam novas pesquisas para validar esses dados. Ao final, apresentam sugestões para aprimorar o ensino do inglês acadêmico.

Em "'Quero que vocês me acompanhem nessa jornada': análise da emergência de metáforas em narrativas sobre o câncer de mama a partir de estratégias de Linguística de *Corpus*", Salgado, Vanin, Gomes e Presotto analisam um tema médico sensível sob a perspectiva de como o mesmo é apresentado nos textos disponibilizados em blogs sobre o assunto. Para tanto, usam da metodologia da LC para analisar terminologicamente as estratégias de *coping*, a forma como a paciente enfrenta a crise pela qual está passando, e uma de suas possíveis manifestações textuais, que é o uso de metáforas (como demonstração de suas experiências subjetivas). Partindo de um *corpus* bastante representativo, as autoras apresentam a descrição da metodologia e resultados a partir de um recorte de estudo, tomando como ponto de partida as palavras-chave, as consequentes metáforas que delas derivam e subjazem à narrativa individual de um *blog* específico (ou seja, um recorte do *corpus* total) e indicam a importância da sensibilização, por parte das equipes médicas, na subjetividade de como cada paciente enfrenta a doença.

Além dos estudos lexicais, os primeiros a serem abordados pela LC, estudos gramaticais (no sentido mais *hardcore*, com as subáreas mais tradicionais da Linguística) continuam sendo elaborados usando *corpora* como base. O texto a seguir é um exemplo.

O texto de Schwindt, Gaggiola e Petry, intitulado 'Frequência e distribuição de plurais irregulares no *Corpus* Brasileiro", descreve a questão dos plurais irregulares de substantivos e adjetivos no português brasileiro, a partir da análise dos exemplos disponibilizados através dos *types* do *Corpus* Brasileiro e tomando a Plataforma R como ferramenta de análise do *corpus*. Os autores procuram entender possíveis regras de formação de plural em três tipos de sílabas finais (predominantemente nomes oxítonos terminados em vogal+u/l e ão no singular), para demonstrar alguns padrões de regularidade na formação de plurais menos comuns (que fogem, no caso, das terminações predominantes vogal+is e ões) na língua portuguesa.

Trabalhos em Terminologia, uma das primeiras áreas a se beneficiar da metodologia da LC, junto com os Estudos de Tradução, continuam a ser elaborados. Podemos perceber, no entanto, que esses trabalhos estão sendo aliados a novas subáreas (como a gramática tradicional e o ensino) e teorias para o desenvolvimento de estudos mais complexos, como vemos a seguir.

"Uma proposta de coextensividade entre termo técnico, grupo nominal e item lexical no português brasileiro: um estudo com base em ferramentas da linguística de *corpus* sob o arcabouço de teoria sistêmico-funcional", o texto proposto por Rodrigues, Ferreguetti e Pagano, baseado nas propostas da Teoria Sistêmico-Funcional da Halliday e focado no fenômeno da coextensividade. Para o estudo, foi compilado um *corpus* (textos acadêmicos sobre Diabetes Mellitus tipo II, com 133.232 *tokens*), em português brasileiro e, como *corpus* de referência, foi usado o CALIBRA (também com textos acadêmicos), compilado usando princípios da a tipologia do contexto de cultura. Através do uso do software AntConc, foram analisadas três palavras-chave (autocuidado, diabetes, saúde) e os respectivos clusters/ngrams das mesmas, mostrando as possibilidades do uso das teorias de Pearson no tratamento de terminologias.

Prado, em "The Pragmatics of Aeronautical English: an investigation through *Corpus* Linguistics", objetiva identificar, num *corpus* oral, elementos que caracterizam a fluência e a interação no inglês aeronáutico, duas das habilidades da Escala de Proficiência Linguística da ICAO para avaliar a proficiência em inglês de pilotos e controladores de tráfego aéreo de modo que possam atuar em operações internacionais. Sua pesquisa investiga padrões de três palavras no *corpus* Radio Telephony

Plain English *Corpus* – RTPEC – com 130 áudios transcritos totalizando 110.737 palavras. O foco do estudo é o desempenho desses profissionais em situações anormais, quando a Fraseologia Padrão não é suficiente e recorrem ao chamado Plain English para negociar a solução do problema. Seus resultados indicam que se faz necessária uma discussão dos conceitos pedagógicos relativos ao conteúdo do inglês aeronáutico em sala de aula, privilegiando uma conscientização pragmática e uma tolerância cultural.

Em "Analyzing the use of personal pronouns in aeronautical communications through CORPAC (*Corpus* of Pilot and Air Traffic Controller Communications)", Pacheco mostra o processo de compilação e análise de um *corpus* de especialidade, o CORPAC (com 36 mil tokens, por enquanto), e uma análise menos comum do que se espera nesse tipo de *corpus*: a frequência de uso de pronomes pessoais (que, teoricamente, deveriam ser evitados na fraseologia área, para se evitar ambiguidade). No momento, o *corpus* está sendo alimentado com a transcrição de vídeos de treinamento para pilotos sobre situações de emergência e analisado com o WordSmith Tools (lista de palavras, seleção de pronomes pessoais e clusters deles derivados). O objetivo do trabalho com pronomes no *corpus* de especialidade é contribuir para o treinamento, elaboração de currículo e testes para a linguagem de aviação em língua inglesa (*Aviation English*).

Continuando no tema da aeronáutica e as questões de comunicação entre controle de tráfego aéreo (ATC) e pilotos (comunicações por radiotelefonia), Peixoto e Tosqui-Lucks, no seu texto intitulado "Weather events in air traffic control standards and communication: discourse patterns and implications for language teaching and assessment", se voltam para outra faceta que pode gerar problemas de comunicação: a terminologia sobre eventos meteorológicos a ser usada nos diálogos entre controladores e pilotos. Desta vez, o público-alvo escolhido são os futuros controladores de voo em suas três grandes áreas de atuação: torre, controle de aproximação e controle de área. Foram analisados 11 termos relacionados à meteorologia num *corpus* de aprendizes (dividido em três sub*corpora*) com a ajuda do AntConc; os textos usados para o *corpus* de referência foram os manuais (brasileiros e internacionais) sobre fraseologismos na área. Resultados indicam que os padrões terminológicos dependem muito do contexto e que os cursos oferecidos no Brasil são eficazes para o ensino da terminologia da área.

A LC vem se aproximando, no século XXI, das questões de ensino, sejam elas relacionadas à aprendizagem da própria abordagem/metodologia, sejam elas relacionadas ao uso de *corpora* como base para desenvolvimento de material educacional. Os próximos dois textos demonstram essas possibilidades.

"*Corpus* linguistics and continuous professional development: participants' prior knowledge, motivations and appraisals", texto de Viana e Lu, pretende nos mostrar, através de uma pesquisa com 28 respondentes de um curso livre no Reino Unido, como a Linguística de *Corpus* está sendo aplicada na prática pedagógica e nas pesquisas em diversas áreas. Dois grupos, tradicionalmente, são abordados como os principais beneficiários da LC: professores de língua e tradutores; os autores traçam um panorama (através de uma análise bibliográfica) bastante abrangente das vantagens e desafios no uso de *corpora* que essas duas áreas apresentam. Na pesquisa com os alunos do curso livre, foram levantadas várias razões para o interesse dos mesmos no trabalho com LC (especialmente no que concerne ao ensino de línguas e seu papel como metodologia), suas expectativas para com o curso e as prováveis barreiras que enfrentariam.

Para terminar este número temático da RELIN, o texto de Bocorny e Welp, intitulado "O desenho de tarefas pedagógicas para o ensino de Inglês para Fins Acadêmicos: conquistas e desafios da Linguística de *Corpus*", trabalha com a ideia de internacionalização e publicação de artigos acadêmicos em língua inglesa. Para tanto, as autoras sugerem princípios para a elaboração de tarefas pedagógicas (e o consequente uso na área de Inglês para Fins Acadêmicos) a partir de um *corpus* (CODISAE, aqui num recorte) composto por textos em inglês (língua franca da academia) de áreas de especialidade (no caso, mais especificamente, da Física), analisando suas introduções. Como sequência, essas tarefas serão disponibilizadas num Ambiente Virtual de Aprendizagem.

Esperamos que a leitura desse número, através dos diversos tipos de análise e reflexões apresentados, seja muito proveitosa, tanto para os pesquisadores iniciantes quanto para os mais experientes na abordagem e na metodologia da Linguística de *Corpus*. Por fim, mas não menos importante, deixamos o registro do nosso agradecimento ao editor e colega Gustavo Ximenes, à equipe de trabalho da RELIN, aos pareceristas e a todos que, direta ou indiretamente, ajudaram a concretizar mais esta edição.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 673-709, 2021

O carpinteiro e a madeira: a constituição de *corpora* jurídicos em perspectiva etnometodológica

The carpenter and the wood: the constitution of legal data from an ethnomethodological perspective

Rubens Damasceno-Morais
Universidade Federal de Goiás (UFG), Goiânia, Goiás / Brasil
r.damasceno.morais@uol.com.br
http://orcid.org/0000-0001-6245-6394

Resumo: Este artigo propõe-se a relatar uma experiência de pesquisa com *corpora* complexos, a fim de compartilhar o processo e procedimentos de elaboração de um banco de dados instituído precipuamente para pesquisa doutoral, empreendida na Université Lumière Lyon II/França, no laboratório ICAR, cuja especialidade é, justamente, o trabalho com a análise de *corpora* em diversos níveis de extensão e complexidade. A partir de uma perspectiva etnometodológica (MONDADA, 2008; OCHS, SCHEGLOFF, THOMPSON, 1996; SCHEGLOFF, 1999; TRAVERS, 2001; TRAVERSO, 2007), numa imersão em território jurídico (CORNU, 2005; DUPRET, 2006; LATOUR, 2004), a pesquisa ora relatada buscou descrever e analisar como os magistrados realizam a gestão do desacordo, em situações, muitas vezes, acentuadamente erísticas. Sem nos distanciarmos dos estudos teóricos acerca dos preceitos de metodologia de trabalhos acadêmicos em geral (GIL, 2002; MOTTA-ROTH; HENDGES, 2010; SALOMON, 2014), constituímos um banco de dados balizados pela noção de *situação argumentativa*, uma noção da retórica antiga retomada por Plantin (1993, 1995, 1996, 2016), a qual põe em destaque situações de conflito de opiniões, em contextos argumentativos vários. A partir da exaustiva e intricada transcrição dupla dos dados (BAUDE, 2006; BLANCHE-BENVENISTE, 2008; KERBRAT-ORECCHIONI, 2006), a pesquisa culminou na confirmação de que o discurso jurídico está longe de ser frio e asséptico e que as interações argumentativas naquele contexto se analisadas no calor das deliberações têm muito a nos ensinar sobre o argumentar em contexto institucional. Isso pode ser conferido em quatro capítulos analíticos cujo planejamento e execução ora trazemos a lume, a partir do estudo do direito em ação, isto é, em situação de interação, por meio

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.673-709

de deliberações de magistrados em processos de danos morais, num tribunal brasileiro de Segunda Instância.

Palavras-chave: etnometodologia; *corpora*; argumentação; tribunal; transcrição de dados orais.

Abstract: This article proposes to report a research experience with complex corpora, on the aim of sharing the backstage of elaborating a database instituted mainly for doctoral research, undertaken at the Université Lumière Lyon II/France, in the ICAR laboratory, whose specialty is precisely work with corpora analysis at different levels of extension and complexity. From an ethnomethodological perspective (MONDADA, 2008; OCHS, SCHEGLOFF, THOMPSON, 1996; SCHEGLOFF, 1999; TRAVERS, 2001; TRAVERSO, 2007), in an immersion in legal territory (CORNU, 2005; DUPRET, 2006; LATOUR, 2004) , the research reported here sought to describe and analyze how magistrates manage disagreement, in situations that are often eristic. Without distancing ourselves from theoretical studies about the precepts of methodology of academic works in general (GIL, 2002; MOTTA-ROTH; HENDGES, 2010; SALOMON, 2014), we formed a database based on the notion of *argumentative situation*, a rhetorical notion retaken up by Plantin (1993, 1995, 1996, 2016), which highlights situations of conflict of opinion, in various argumentative contexts. From the exhaustive and intricate double transcription of the data (BAUDE, 2006; BLANCHE-BENVENISTE, 2008; KERBRAT-ORECCHIONI, 2006). The research culminated in the confirmation that the legal discourse is far from being cold and aseptic and that argumentative interactions in that context, if analyzed in the heat of deliberations, have much to teach us about arguing in an institutional context. This can be seen in four analytical chapters whose planning and execution now we bring to light, from the study of law in action, that is, in a concrete situation, from the deliberations of magistrates in moral damages cases, in a Brazilian court of Second Instance.

Keywords: ethnomethodology; *corpora*; argumentation; court; transcription of oral data.

Recebido em 24 de agosto de 2020

Aceito em 09 de outubro de 2020

Introdução

Há mais ou menos vinte anos, os estudos com base em *corpora* orais movimentaram o cenário dos estudos em ciências da linguagem (BAUDE, 2006, p. 25), sobretudo em território das interações verbais. Em se tratando de território jurídico, as pesquisas de interações argumentativas há muito clamam por um olhar atento, devido à dificuldade da coleta de

dados em situações reais de interação, como tentaremos esmiuçar neste artigo, ao observarmos as peculiaridades da construção de discurso e contradiscurso em *episódios argumentativos* orais (PLANTIN, 2016, p. 77), como explicaremos a seguir.

No contexto do estudo das interações em geral, viu-se surgir, ainda nos últimos anos, uma busca maior por fatos ligados a situações de interação linguística do dia a dia ou situações institucionais, em contexto de trabalho, em que se tentavam retratar atividades diversas (o padeiro atendendo a um cliente numa padaria; o vendedor numa transação comercial; conversas informais ao lado da máquina de café etc.). Segundo Schegloff, "a interação apresenta-se como o foco principal da vida em sociedade" (1999, p. 141), e os pesquisadores interessados pelas interações passaram a consagrar suas análises ao exame dos mecanismos contingentes da fala e outros recursos semióticos, como o olhar, a prosódia, o movimento do corpo, os gestos e seus efeitos sobre a constituição emergente e dinâmica do contexto.

No domínio da Análise da Conversação ou Análise da Conversa Etnometodológica (ACE),[1] por exemplo, tratou-se de descrever a organização social de uma interação verbal e ainda o trabalho de categorização ou de descrição que uma relação interacional implicava, por meio de pesquisas que enxergavam a gramática como um modo de interação social. E os autores prosseguem afirmando que a gramática passou a ser vista como mantenedora de um laço estreito com as interações. Assim, passou-se mesmo a sugerir que as interações eram mais do que um recurso, elas faziam parte da própria gramática. Dito de outra forma, "a gramática foi vista como intrinsecamente interacional" (OCHS, SCHEGLOFF, THOMPSON, 1996, p. 38). Para Schegloff (1999, p. 142), ainda, "a relação profunda e íntima entre a 'linguagem' e o falar-em-interação foi tal que, para entender a linguagem, era necessário observar os contextos de interação que certamente a circundavam". Nesse sentido, a preocupação do pesquisador em fazer uma boa utilização para uma melhor compreensão (mais "profunda e íntima") dos dados – muitas vezes arduamente coletados – é

[1] Análise da Conversa Etnometodológica (ACE) é a expressão convencionalmente utilizada no Brasil.

> análoga ao interesse do carpinteiro ou do escultor de madeiras pela madeira com que trabalha. Por essa razão, se queremos de fato compreender o nosso objeto de pesquisa, devemos conhecer suas características intrínsecas, da mesma forma que *um carpinteiro deve conhecer a natureza da madeira com que trabalha.* (OCHS *et al,* 1996, p. 18, destaque nosso).

É precisamente dessa relação de "intimidade", de cuidado do pesquisador para com os seus dados que vamos aqui tratar. Por esta razão, achamos pertinente propor a descrição desta experiência com *corpora* complexos para este número especial sobre Linguística de *Corpus*, sobretudo porque, como já ressaltamos, a criação do banco de dados TRIBUNAL (forma como nomeamos o *corpus* selecionado para a pesquisa) representou efetivamente enorme desafio, motivado por um trabalho de descrição linguística de dados empíricos advindos de exemplos reais de uso da linguagem empregada em deliberações entre magistrados num tribunal brasileiro de Segunda Instância.[2] Nesse sentido, esclarecemos que os *corpora* mobilizados para a pesquisa ora apresentada são de complexidade não negligenciável, por três razões precípuas. Vamos a elas:

Primeiramente, porque empreendemos uma investigação no universo jurídico sem termos formação jurídica. Desse modo, tivemos de decifrar partes daquele ritual para não corrermos o risco de analisar os dados de forma ingênua ou simplesmente equivocada, por se tratar de território extremamente técnico, tanto pela questão terminológica jurídica quanto pelas idiossincrasias do ritual de deliberação, que escapam a um olhar meramente espectador ou diletante. Segundo, porque tivemos de traduzir todas as deliberações em linguagem técnica jurídica do português para uma linguagem técnica jurídica francesa uma vez que a tese seria defendida, inicialmente, para um júri que não compreende nem fala a língua portuguesa, o que também não é tarefa simples, visto a extensão dos *corpora* que mobilizamos para a investigação proposta. E, por fim, porque tivemos de fazer a transcrição da língua falada tanto para o português quanto para o francês. Desse modo foi necessária dupla transcrição, cuidando-se para que os *gaps* naturais da língua falada fossem compreendidos na versão escrita dos dados, nas duas línguas

[2] Os dados proporcionaram-nos quatro capítulos de minuciosa descrição e análise linguística (DAMASCENO-MORAIS, 2013) os quais relataremos de forma sumária neste artigo.

(português e francês), numa linguagem marcadamente técnica, mas que, no geral, também trazia características da conversação,[3] o que não facilita o trabalho do pesquisador.

Para não nos distanciarmos da metáfora do carpinteiro, ressaltamos que gerar, selecionar e seccionar um *corpus* é trabalho que não se distancia do ofício desse profissional da madeira, pois, assim como ele, devemos conhecer a fundo a matéria com a qual trabalharemos para sabermos exatamente onde/como lixar, limar, montar-desmontar-remontar todo um arsenal de dados que, se mal geridos, podem transformar uma pesquisa num grande fiasco acadêmico,[4] se se parte de um "background etnográfico" (OCHS *et al*, 1996) mal-ajambrado.[5] Nesse sentido, organizar uma pesquisa que lide com *corpora* extensos exige, se não muita dedicação, ao menos muito zelo com a matéria-prima a se trabalhar, sob o risco de se fazerem análises meramente intuitivas e, nesse sentido, desprovidas de relevância científica ou mesmo condenáveis metodologicamente. Importante ainda esclarecer que, em sentido lato, entendemos "pesquisa" como "um processo planejado de investigação que consiste em três momentos: 1) levantamento de perguntas, hipóteses ou problemas, 2) coleta dos dados, 3) análise e interpretação dos dados" (MOTTA-ROTH; HENDGES, 2010, p. 111).

O objetivo deste artigo é, por fim, apresentar, de forma sucinta, o processo e procedimentos de elaboração de um banco de dados, isto é, a

---

[3] No Brasil, a ACE convencionou usar o termo "conversa". Neste trabalho utilizaremos indiferentemente "conversação" ou "conversa".

[4] Apesar de não trazermos um exemplo da área de Linguística, importa aqui ilustrarmos a situação com um "bom" exemplo de pesquisa malsucedida. Como este artigo trata, justamente, de metodologias de pesquisa, achamos válido refletir, com um exemplo recente, que, independentemente da área, se a metodologia não é bem estabelecida, o trabalho corre sérios riscos de ser questionado. O "fiasco" ao qual nos referimos é o recente estudo contestado pela comunidade científica a respeito de pesquisa sobre a eficácia do uso da hidroxicloroquina. A pesquisa foi rejeitada via carta aberta porque a metodologia adotada na pesquisa e a integridade dos dados apresentavam falhas. A esse propósito, ver: *Cientistas questionam em carta aberta estudo sobre a hidroxicloroquina na The Lancet* (https://www.uol.com.br/vivabem/noticias/afp/2020/05/29/cientistas-questionam-em-carta-aberta-estudo-sobre-a-hidroxicloroquina-na-the-lancet.htm?cmpid=copiaecola) – Acesso: 19 ago. 2020.

[5] O que Ochs *et al* (1996) chamam de "background etnográfico" refere-se a uma análise mais aprofundada de contexto e, ainda, descrições mais detalhadas do objeto de pesquisa.

metodologia, elaborada especificamente para pesquisa de tese doutoral, defendida em 2013 na Université Lumière Lyon II, no laboratório Interactions, Corpus, Apprentissages, Représentations – ICAR,[6] cuja especialidade é, justamente, o trabalho com a análise de *corpora* diversos. Em tradução livre, o título do trabalho ora descrito é: "O preço da dor – gestão de desacordos entre magistrados, em um tribunal brasileiro de Segunda Instância",[7] trabalho ainda não publicado integralmente em português (DAMASCENO-MORAIS, 2013).

2 Inspiração etnometodológica em território jurídico

As pesquisas acerca do contexto jurídico e interações verbais começaram a pulular por volta dos anos oitenta. No início, o foco era mais a tentativa de perscrutar a forma como o trabalho era realizado pelos operadores da lei, a partir do que se chamava de "*tâches professionnelles*" (TRAVERS, 2001, p. 355) ou a descrição minuciosa da rotina de um trabalho realizado por alguém. O primeiro estudo de peso acerca do direito em contexto judiciário, no domínio da Análise da Conversação, chama-se *Order in Court : The Organisation of Verbal Interaction in Courtroom Settings* e foi realizado por Maxwell Atkinson et Paul Drew (1979). Sua pesquisa tomou como *corpus* audiências realizadas em um tribunal na Irlanda do Norte, no final dos anos sessenta.

Para Travers, os estudos de inspiração etnometodológica em contexto judiciário "fundam-se na análise minuciosa de gravações, feitas no intuito de se explicar/compreender de que modo as pessoas emprega[va]m recursos culturais e comunicacionais em audiências judiciárias" (TRAVERS, 2001, p. 359). No campo da Etnometodologia ainda se leva muito em consideração os meandros de uma interação ordinária, as justificativas ali apresentadas, os atores sociais que da interação participam, as atividades que realizam (e como as realizam), para "descrever as interações sociais" (GARFINKEL, 1967 *apud*

[6] O trabalho foi realizado na *École Doctorale Lettres, Langues, Linguistique & Arts* (ED 484), no laboratório ICAR (Interactions, Corpus, Apprentissages, Représentations – UMR 5191), e defendido na Université Lyon II, França.

[7] No original: *Le prix de la douleur: Gestion des désaccords entre magistrats, dans un tribunal brésilien de seconde instance*. A tese foi orientada por Christian Plantin e Véronique Traverso, e o ritual de defesa contou com os seguintes membros de banca: Christian Plantin, Véronique Traverso, Barbara Villez, Marianne Doury, Wander Emediato.

PLANTIN, 2016, p. 270) e a compreensão da inteligibilidade mútua das ações, a partir de expectativas sociais ou até de normas morais colocadas em prática ao longo de uma interação.

Nesse sentido, a pesquisa que ora documentamos buscou fôlego na perspectiva etnometodológica para composição do *corpus*, transcrição e análise de dados, pois, apesar de não focar em uma situação corriqueira do dia a dia, empreendemos as análises por meio de gravações de áudio realizadas por um tribunal para que, a partir das transcrições de tal material, fosse elaborado todo o trabalho de dissecção, que passa certamente pelas etapas de definição, coleta, seleção e análise. Nesse sentido, cientes de que o nosso objetivo precípuo era gerir uma pesquisa acadêmica, o que implica a "busca de resposta a problemas propostos" (GIL, 2002, p. 17), num exercício "científico por excelência" (SALOMON, 2014, p. 217), empreendemos a descrição minuciosa de interações verbais em território jurídico, como forma de perscrutar a lacuna que separa as leis formais (a letra de lei) e as maneiras pelas quais as decisões judiciárias são tomadas, *ao vivo e em cores*, isto é, na prática, algo não muito recorrente, sobretudo – e como já ressaltamos – devido às dificuldades de acesso a dados de natureza tão restrita, por rituais minuciosamente organizados em âmbito institucional.[8]

Desse modo, flagrar o momento em que uma decisão é tomada, no calor de um debate, de uma interação com forte teor emocional e argumentativo, é uma forma bastante instigante de compreender um contexto, e não somente jurídico. De fato, sabe-se que os tribunais são "o lugar privilegiado para se observar o direito em ação, por meio dos diversos eventos que acontecem ao mesmo tempo, tais como processos civis e penais, interrogatórios, escuta de testemunhas etc." (DUPRET, 2006, p. 424). Trata-se, em fim de contas, de "um microuniverso [o jurídico] em que os atores sociais estão engajados em um processo de comunicação linguística, caracterizado em sua maioria pelo falar em interação", como mostraram as análises apresentadas na tese cujo breve relato fazemos aqui.

[8] Apesar dessa dificuldade de acesso, tivemos autorização por escrito do tribunal que nos forneceu os dados, desde que garantíssemos o sigilo dos participantes. Em realidade, na tese defendida não se pode identificar nem local nem datas de seleção do *corpus*.

3 O estudo do direito em ação

Tomando por base os dados de cunho etnográfico coletados em um tribunal brasileiro, a pesquisa que ora relatamos brevemente é o resultado da descrição de procedimentos de magistrados atuantes na Segunda Instância (relatores, revisores e vogais),[9] no momento de deliberações conflituosas, em que se evidenciam divergências de opinião, em julgamentos acerca de danos morais. Segundo Kerbrat-Orecchioni (1995, p. 8), as trocas realizadas entre participantes de interações trilogais (com três participantes), pouco importando o tipo de contexto, pode proporcionar boas surpresas ou, no mínimo, apresentam estruturas interacionais interessantes. E não por acaso escolhemos os dados que apresentaremos, isto é, dados de situação de interação oral, muitas vezes polêmicas, em que o teor emocional se faz notar por mecanismos paraverbais como tom de voz, hesitações, palavras pronunciadas pela metade, gaguejamento etc. bastante típicos da linguagem oral, sobretudo sob um "enfoque interacionista" (KERBRAT-ORECCHIONI, 2006, p. 11; MARCUSCHI, 2003).

Esse tipo de dado é bastante comum em trabalhos ditos etnográficos, os quais "sem se envergonhar de sua ignorância" (LATOUR, 2004, p. 205), lançam-se a descrever situações desconhecidas "na forma como elas realmente acontecem", a partir de metodologias diversas. Nesse sentido, quando afirmamos que nos enredamos em uma teia complexa de dados, sentimo-nos como um espeleólogo ao explorar uma caverna virgem (SERRANO, 2011, p.15), ou como alguém imerso em um labirinto em busca de uma saída, como poetiza Serrano:

> Assim como Teseu, o pesquisador se encontra perdido e enclausurado em seu próprio labirinto. O *método* será o fio de Ariadne que o ajudará a encontrar a saída, mostrando-lhe o caminho. Sem sua ajuda [do método], a pesquisa será para sempre um dédalo indecifrável, de portas trancadas e de muros invencíveis, uma algaravia sem sentido se for produzida a partir das arremetidas desconcertantes da intuição, um emaranhado de objetivos laxos, de esboços difusos e de aspirações entrecruzadas (SERRANO, 2011, p. 102).

[9] No ritual jurídico, um magistrado "vogal" é quem tem a primazia de votar numa deliberação.

O objetivo da pesquisa ora relatada foi descrever e analisar como os magistrados fazem a gestão do desacordo, em situações, muitas vezes, acentuadamente erísticas. O *corpus* para utilização na pesquisa específica da tese foi constituído a partir da noção de *situação argumentativa*, uma noção da retórica antiga retomada por Plantin (1993, 1995, 1996, 2016), a qual põe em destaque situações de conflito de opiniões, em diversos contextos argumentativos. Desse modo, é importante destacar, para esta pesquisa selecionamos apenas situações em que há conflito de opiniões – ou *estases* – as quais foram por nós descritas e analisadas a partir de pressupostos dos estudos etnometodológicos e que prestam atenção ao "estudo do fazer e dizer em contexto" (DUPRET, 2006, p. 17).

Na pesquisa empreendida interessamo-nos pelo efeito que as regras procedimentais do ritual em 2ª Instância geram na interação; isto é, observamos de que forma a ação dialogal/trilogal se desenvolve ao longo de uma deliberação, em contexto jurídico. Estivemos atentos ainda à maneira pela qual os magistrados lidam com as formalidades e restrições do contexto e o reflexo disso na interação argumentativa que se desenvolve ao longo das deliberações. A partir disso, propusemos algumas análises dos excertos coletados, nos dados que nomeamos de *Corpus* TRIBUNAL, como forma de tentarmos sistematizar os meios e métodos utilizados pelos interactantes (magistrados) no momento de gerirem os conflitos de opinião surgidos, ao vivo e em cores, na dinâmica de alguns julgamentos.

Para Dupret (2006), um estudo jurídico de inspiração etnometodológica "nos permite prestar atenção à construção de fatos, ao comportamento dos interactantes no contexto judiciário, à forma como lidam (os interactantes) com as restrições impostas pela formalidade do ambiente etc." (DUPRET, 2006, p. 91). Esse foi, em resumo, o propósito do trabalho, no qual nos predispusemos a investigar "o direito em ação" (MARTÍNEZ, 2007, p. 5). E tal missão coaduna-se com um dos princípios do interacionismo, o qual busca compreender "a linguagem em ação" (TRAVERSO, 2007, p. 17). Tal método de pesquisa prioriza, enfim, a análise de textos orais que resultem de situações de troca verbal, em contextos de contato face a face, "produzidas por mais de uma pessoa em interação" (SCHEGLOFF, 1996, p.10).

Em verdade, acreditamos que o tipo de pesquisa que realizamos é vital para o desenvolvimento de uma sociologia da realidade jurídica mesclada a uma análise linguística cujo estudo, apesar de

interessantíssimo, é muito pouco pesquisado por campos diferentes do estritamente jurídico. Desse modo, uma análise do mundo das leis – muitas vezes bastante controverso – e visto do campo dos estudos interacionais, argumentativos e retóricos pode ser uma boa maneira de adentrarmos cortes e tribunais, mesmo sem sermos juristas. E isso não significa empreendermos uma pesquisa "perfunctória" (como se diz no jargão jurídico), isto é, superficial, meramente intuitiva ou desprovida de rigor metodológico. Aliás, este artigo serve exatamente para mostrar como se pode, a partir de uma perspectiva etnometodológica, compreender um domínio intricado e dali extrair resultados que façam sentido, com a ajuda de metodologia responsável e o menos subjetiva/intuitiva possível.

4 A função retórica da metodologia e a dissecção dos dados

Os tipos de pesquisa mais comumente citados na literatura especializada são: exploratória, descritiva e explicativa. Essas apresentam delineamentos do tipo experimental, estudo de caso, bibliográfico, entre outros. A população e amostra referem-se com o tipo e a extensão da amostra ou com o universo da pesquisa. A coleta de dados refere-se às técnicas a serem usadas, instrumentos e observação em momentos específicos da pesquisa. Quanto à análise de dados, procede-se, geralmente, quantitativa ou qualitativamente (GIL, 2002; MARTINS JÚNIOR, 2012; SALOMON, 2014, entre muitos outros). Como vemos, os aspectos metodológicos de uma pesquisa são vários. A seguir, vamo-nos ater à *função retórica* da metodologia, isto é, "narrar os procedimentos de coleta (fonte, tamanho da amostra, critérios para a coleta) e análise de dados e descrever os materiais que levam à obtenção de resultados, com maior ou menor detalhamento" (MOTTA-ROTH; HENDGES, 2010, p. 115), para que se possa compreender a jornada aqui consubstanciada.

Quando se fala em metodologia, corre-se o risco de se cair na cilada do classificacionismo anódino, isto é, rotular intuitivamente uma pesquisa sem muita certeza de que tipo de metodologia se está a empreender (fato corriqueiro, a bem da verdade). Nesse sentido, deve-se considerar que essa questão "é mais complexa, pois existem diversas maneiras de classificar métodos e *pouco consenso* entre os teóricos sobre qual o número e o nome exato dos métodos" (MOTTA-ROTH; HENDGES, 2010, p. 114, destaque nosso). Nesse sentido, ao se falar em metodologia, é fundamental (e independentemente do rótulo utilizado)

que se deixem claríssimos os recursos materiais e procedimentos adotados, buscando bem informar o auditório dos passos executados na pesquisa (participantes, tipos de amostras, instrumentos utilizados, programas computacionais etc.), sobretudo porque a "integridade intelectual" (GIL, 2002, p. 18), exige "rigor na pesquisa" (COSCARELI; MITRE, 2007, p. 74), a qual deve ser deslindada em sua complexidade pelo pesquisador, como numa espécie de prestação de contas, na hora de divulgação dos resultados da pesquisa, exatamente como nos propomos a fazer nesta chamada especial de publicação.

Nesse sentido, as informações que julgamos importantes para este momento "retórico" dizem respeito às fontes de pesquisa, que, no caso presente, foram fontes primárias (filmagens *in loco* realizadas pelo próprio pesquisador) e dados secundários (coleta de áudios de um banco de dados mantido pelo tribunal). No trabalho realizado na elaboração de tese, apenas utilizamos os áudios do banco de dados do próprio tribunal. As filmagens, acórdãos publicados e afins serão utilizados em pesquisas futuras. Para o escrutínio dos dados (pré-)selecionados, em um longo processo, foi necessário um profundo mergulho no extenso material coletado, numa relação com os dados do tipo 'carpinteiro → madeira', para que de fato pudéssemos examinar os mecanismos de gestão do desacordo utilizados pelos magistrados em momentos de *estase argumentativa*, seja durante a qualificação de um fato como dano moral, seja no momento de definir o montante a ser pago, em caso de ilicitude comprovada em uma ação.

Para consecução das análises, procuramos ainda seguir os passos propostos por Traverso (2007, p. 23), para análise de interações dialogadas. Nesse sentido, as etapas seguidas neste processo foram:

1º passo: Escolha da situação a ser analisada

2º passo: Observação

3º passo: Coleta dos dados

4º passo: Transcrição

5º passo: Análise.

Em se tratando de *corpora* complexos, acreditamos que o segundo passo acima descrito é um dos mais importantes, pois é o momento em que o pesquisador adentra a "caverna escura". Assim como o espeleólogo,

ali se vai tentar descrever e entender o ambiente em exploração. Nesse sentido, não é um momento de mera "observação" dos dados, mas um momento de observação "exaustiva", pois, tudo que ali se nota pode auxiliar no momento de se elegerem categorias analíticas (momento dos mais difíceis de uma pesquisa). Nesse sentido eu diria que *tudo o que ali se nota é notável*. Uma ocorrência que se repete pode ser entendida como um possível padrão e que poderá ser descrito de maneira minuciosa ao longo da pesquisa. Esse momento de observação exaustiva merece um bloco de notas do tipo "anotações *brainstorming*" em que o pesquisador anotará tudo que, de uma forma ou de outra, chamou a sua atenção e que poderá se tornar uma importante categoria de análise nas etapas posteriores. Lembrando: tudo o que se nota é notável.

A partir dessas observações exaustivas o pesquisador estará apto a acessar a próxima etapa, isto é, coletar e recortar o que lhe interpelara, na exploração inicial da caverna escura. Nesse sentido, será simplesmente impossível recortar de forma responsável os dados sem esse mergulho. Então, aconselha-se que se faça uma boa exploração do terreno para que, no momento de empreender as análises, o *corpus* não se torne um corpo estranho. Caberá ao pesquisador falar com desenvoltura sobre os seus dados porque, em final de contas, a autoridade ali será o próprio pesquisador (e ninguém mais). Nesse sentido, não se pode falar com perícia sobre algo que não se conhece ou que se conhece *en passant*. É importante destacar que a diferença entre um bom e um mau carpinteiro é justamente o modo como esse profissional manuseia sua matéria-prima, a intimidade que demonstra ao discorrer sobre ela. E cada madeira tem suas idiossincrasias (fibrosidade, resistência, densidade, coloração etc.). Ora, então por que seria diferente com os dados da minha pesquisa?

Para consecução do trabalho ora relatado, primeiro selecionamos uma situação (os julgamentos em Segunda Instância); em seguida, após escuta e reescuta exaustiva de tudo que fora coletado, leituras de julgados (registros de deliberações, leitura de acórdãos etc.), selecionamos o que de fato faria parte do *corpus* das análises. As transcrições aconteceram na etapa seguinte e as análises renderam quatro capítulos, um para cada categoria notada e devidamente anotada na fase inicial de observação. O Quadro 1 recapitula esse processo.

QUADRO 1 – Recorte dos dados

RECORTE DOS DADOS – *Corpus* TRIBUNAL	
Etapa 1	Definição de situação a ser analisada (= deliberações em Segunda Instância)
Etapa 2	Observação exaustiva (= escuta de julgamentos; leitura de acórdãos etc.)
Etapa 3	Coleta dos dados (= seleção e recorte do que faria parte da tese)
Etapa 4	Transcrição (= deliberações previamente gravadas entre juízes)
Etapa 5	Análise (= a partir de categorias analíticas identificadas na Etapa 2)

Fonte: Elaboração do autor

Importante observar a imbricação de todas essas etapas. A Etapa 5 só foi possível graças a uma Etapa 2 exaustiva. Ali, visitei arquivos mal iluminados do tribunal, literalmente escalei prateleiras empoeiradas para buscar processos já julgados e arquivados, mergulhei em votos publicados na internet, filmei e gravei julgamentos (tudo com devida autorização), escutei e reescutei mais de cem áudios (julgamentos) para, ao final disso tudo, selecionar o que tinha despertado meu interesse e que mais se adequava à problemática da pesquisa, isto é, a gestão do desacordo entre magistrados num tribunal de Segunda Instância.

Importante destacar que se trata de longo e árduo processo. Somente na quinta etapa é que efetivamente pude ter uma visão panorâmica do material coletado. Nos quatro capítulos analíticos que a análise desses dados nos permitiu elaborar, esforçamo-nos para destacar a importância da dimensão institucional e sua influência no sistema de turnos de fala dos magistrados, a partir da descrição da forma de organização e a conduta dos participantes ao longo das interações (os magistrados *experts*), nos momentos de conflito de opinião. Nossa preocupação ali foi, sobretudo, descrever a dinâmica de atuação dos magistrados no momento de definição do *pretium doloris*, ou seja, no

momento de se definir um valor financeiro a um ilícito, considerado um dano moral. Nesse sentido, interessamo-nos ainda pela organização institucional do procedimento, pelas escolhas lexicais, pelas rotinas empregadas pelos participantes, pelo efeito da falta de assimetria entre os participantes dos debates etc. Em suma, buscamos respostas às questões de pesquisa. Tivemos o cuidado de procurar apresentar e explicar, a partir de uma base teórica oriunda dos estudos da argumentação e da retórica, a dinâmica empregada, ao longo das deliberações, a fim de compreendermos um tênue aspecto da prática judiciária e o modo de gestão do desacordo naquele contexto, presente no *corpus* TRIBUNAL.

Apresentaremos neste artigo alguns números correspondentes à etapa de constituição dos dados orais pertencentes ao *corpus* TRIBUNAL, isto é, relativos a essas etapas de dissecção dos dados. Inicialmente foram selecionados 263 julgamentos em 2ª Instância, advindos todos dos arquivos eletrônicos do tribunal que nos autorizou a visitar seus processos judiciais. O critério para a seleção dos processos foi, nesta primeira etapa, o *ano* de cada julgamento (a partir do ano 2000) e o *tema* do caso julgado. Lemos cada um dos 263 resumos, no intuito de fazermos a triagem dos casos que de fato poderiam nos interessar, de acordo com o objeto da pesquisa. Somente nos interessaram aqueles julgados sobre *dano moral* e cujo teor parecia mais polêmico, pois, não nos podemos esquecer, o curso da pesquisa eram os conflitos em deliberações entre desembargadores. Claro que tudo isso já fora delimitado pela problemática (a gestão do desacordo). Nesse sentido, é bom prestarmos atenção, uma problemática não precisa ser um problema no sentido de "dor de cabeça", mas um "problema" no sentido de *aporética,* isto é: "pergunta científica, ou seja, a formulação correta dos problemas ou dos objetos a investigar" (SALOMON, 2014, p. 282), uma espécie de *norte* para a pesquisa e que não pode jamais ser dissociado das etapas de seleção, recorte e análise dos dados.

Todo esse trabalho de triagem foi feito antes de sairmos à cata de cada julgamento (em forma de processo) e de cada registro em áudio, etapa que nos permitiria conhecer em profundidade o teor de cada julgamento, pois até ali apenas tínhamos lido os resumos, mas não tínhamos nem ouvido o julgamento nem buscado os processos nos arquivos. Na verdade, esse enxugamento foi necessário, uma vez que os arquivos do tribunal eram imensos (tanto o arquivo de processos físicos, isto é, em papel, quanto o arquivo de áudios); desse modo, seria

contraproducente sair à cata de quase 300 processos, o que poderia tomar meses de trabalho. Por isso, tivemos de fazer essa seleção criteriosa, para irmos em busca somente dos casos que de fato pudessem trazer algum interesse para a pesquisa, segundo os critérios preestabelecidos. A seleção já tinha começado e precisava ser racional, sob o risco de esmorecermos frente a *corpora* muito extensos.

Assim, mais 100 resumos foram excluídos por motivos diversos: tema repetido, dúvidas acerca do real interesse do documento para a pesquisa, grau de polemicidade do assunto etc. Na verdade, estávamos em busca de casos de dano moral que tivessem origem em processos com teor aparentemente mais emocional (por exemplo, solicitação de compensação financeira em casos de sofrimento psicológico alegados pela parte ou a "dor do espírito", em vez de casos de mero dano material como o ressarcimento por uma batida de carro, por exemplo). Isso porque o escopo do nosso projeto buscava, num primeiro momento, analisar argumentação e emoções, o que, por si, justifica nossas escolhas nesse momento de seleção dos julgamentos. Desse modo, os casos que não se encaixaram em tais critérios foram sendo eliminados, mesmo se, num primeiro momento, tenham despertado nosso interesse. E esse é um momento de muita atenção, porque, num primeiro momento, tudo parece interessante, mas como não existe especialista *X-tudo*, precisamos delimitar, recortar, lixar, limar os dados e estarmos atentos para não nos enredarmos numa teia que poderá se tornar um eterno labirinto.

Isso porque uma pesquisa não precisa ser um vetor de sofrimento, muito pelo contrário, a pesquisa precisa ser prazerosa desde o começo. Por isso, o manuseio dos dados (e voltamos mais uma vez à metáfora do lenhador) é vital para o sucesso da empreitada. E mais, não nos devemos esquecer de que uma pesquisa nunca exaurirá as possibilidades de exploração dos dados. É por essa razão que devemos estar sempre atentos à problemática, à resposta que queremos encontrar e que nos possibilitará falar com autoridade e destreza sobre pelo menos um aspecto notório da pesquisa empreendida. A baixa qualidade de muitas pesquisas advém de análises sem o necessário aprofundamento, um tanto quanto pasteurizadas, até mesmo lugares-comuns, frutos, muitas vezes, de lampejos de nossa própria empolgação com nossos dados, sem, muitas vezes, apresentarmos uma investigação de fato metódica e consistente. Devemos ficar vigilantes e evitarmos o caminho fácil das generalizações apressadas, as quais, inclusive, têm cunho falacioso.

Para melhor compreensão do contexto no qual a pesquisa que ora relatamos se situa, é importante esclarecer que todos os casos julgados em 2ª Instância são automaticamente gravados em áudio e arquivados. Esse procedimento faz parte da rotina dos julgamentos no tribunal que abrigou nossa pesquisa. Desse modo, após minuciosa busca e desapego total de alguns dados que já tinham sido coletados, acabamos por selecionar pouco mais de cem processos. Como dito, preferimos os casos que pareciam mais polêmicos, isto é, aqueles casos em que a intensidade da *estase* parecia ser maior. Em outros termos, somente os casos de desacordo de opiniões mais denso entre os magistrados foram selecionados, em julgamentos acerca do *pretium doloris*. Esse foi um critério baseado na problemática da pesquisa.

O Quadro 2 resume os casos que subsistiram a todas as etapas de triagem e, após minuciosa observação, passaram efetivamente a fazer parte do corpus da pesquisa ora relatada.

QUADRO 2 – *Corpus* TRIBUNAL

***Corpus* TRIBUNAL**	
27 casos selecionados	**Dissecção dos dados**
Caso_59	8 excertos
Caso_18, Caso_60	6 excertos
Caso_15	4 excertos
Caso_3, Caso_8, Caso_17, Caso_20	3 excertos
Caso_1, Caso_4, Caso_7, Caso_22, Caso_47, Caso_61, Caso_62	2 excertos
Caso_9, Caso_11, Caso_16, Caso_25, Caso_26, Caso_32, Caso_33, Caso_40, Caso_41, Caso_43, Caso_46, Caso_48	1 excerto

Fonte: Elaboração do autor

O Quadro 2 sintetiza um processo de seleção a partir de mergulho em vasto *corpora* (sentenças, áudios, filmagens, observações de audiências etc.) e que durou pelo menos 1 ano e meio. Ali, das centenas de julgamentos lidos, áudios atentamente escutados, filmagens realizadas, tivemos de compreender de quais dados efetivamente necessitaríamos

para, finalmente, fazermos a seleção daqueles que fariam parte do *corpus* TRIBUNAL: apenas 27 julgamentos em áudio, todos julgados apenas em 2ª Instância, a partir do ano 2000. Nesse sentido, selecionamos todos os casos que traziam alegação de danos morais pelas partes autoras dos processos. Após inúmeras idas e vindas, julgamos que esse recorte seria o ideal para nos ajudar a responder às questões de pesquisa. Despedimo-nos ali do labirinto e traçamos a rota reta que nos levaria às respostas que buscávamos.

Uma particularidade dos dados orais coletados deve-se ao fato de que, como já aludimos acima, todos os julgamentos são gravados automaticamente e em seguida arquivados pela própria instituição. Os excertos que finalmente fizeram parte do recorte final (QUADRO 2) são dados secundários, isto é, não foram elaborados diretamente por nós. Eles foram registrados independentemente desta pesquisa e sem qualquer interferência de nossa parte, uma vez que todas as deliberações são automaticamente registradas em áudio, naquela corte de justiça, sendo imediatamente arquivadas, após cada sessão de deliberação. E essa característica dos dados torna-os bastante instigantes; as interações ali registradas são espontâneas, na medida em que os interactantes, ao longo das deliberações, não ficaram sob a mira do gravador ou da câmera de um pesquisador alheio àquele ritual.[10]

Nesta pesquisa, optou-se por se manter anônima toda e qualquer referência que pudesse identificar as pessoas participantes das sessões. De acordo com Baude, a anonimização dos dados "é importante para a vida privada das pessoas envolvidas e para a legalidade dos dados coletados pelos pesquisadores" (BAUDE, 2006, p. 67). Temos consciência, em realidade, de que a anonimização dos dados é uma condição necessária para a preservação da instituição na qual os dados foram gerados. Para Latour, quando trabalhamos com dados de um tribunal, por exemplo, confrontamo-nos com um problema de método e de deontologia, no momento de publicarmos os resultados da pesquisa. Ainda de acordo com o autor, que desenvolveu trabalhos no meio jurídico, é necessário "ocultar

[10] Em outros momentos da pesquisa, munimo-nos de câmeras e filmamos várias sessões de deliberação e também de julgamento. Ali ficou patente a forma não tão natural com que as pessoas reagiam, talvez intimidadas pela câmera e pela minha presença. Não obstante, os dados que nós mesmos registramos (vídeo e áudio) não foram selecionados para a pesquisa aqui relatada; mas fazem parte do *corpus* TRIBUNAL.

nomes de pessoas, de lugares, de rituais cujas etapas são analisadas, sem, contudo, permitir que tal atitude descaracterize o que se tenta mostrar na pesquisa" (LATOUR, 2004, p. 7-9).

Desse modo, nas transcrições realizadas, apagamos toda e qualquer possibilidade de identificação (nomes, sobrenomes, apelidos etc.), dados pessoais (endereços, locais de nascimento, números identificadores etc.), referências a lugares (topônimos, instituições, serviços etc.). Em realidade, nem o tribunal – ou a cidade em que esse se situa – foi informada. Nem mesmo de quantos tribunais os dados foram retirados apresentamos informações seguras, de propósito. Referimo-nos sempre a "um" tribunal; nada impede que tenhamos montado o *corpus* TRIBUNAL a partir de casos julgados em diversas cidades. A única informação precisa que fornecemos é que os dados foram coletados em tribunal brasileiro, julgados a partir do ano 2000. Também não indicamos ano exato de julgamento, para inviabilizar – ou dificultar – qualquer possibilidade de identificação.

Tais medidas foram necessárias para que se pudesse manter o anonimato acerca das identidades dos magistrados que participaram das deliberações. Desse modo, em vez de divulgarmos os números reais dos processos, os nomes de autores e de réus, optamos por identificar cada caso com o nome "Caso", seguido de um traço "_" e de um número aleatório. Também optamos por nomear ficticiamente cada caso, para facilitar sua identificação, ao longo das análises, as quais, muitas vezes, serão comparativas. O Comitê de Ética/CEP foi acionado e autorizou a divulgação dos resultados da pesquisa, a qual, devemos lembrar, foi realizada integralmente fora do Brasil. Ainda, reiteramos que tivemos autorização do próprio tribunal para coleta e análise de dados, desde que anonimizados.

Propomos, a seguir, no Quadro 3, os 27 casos com os números e codinomes que lhes demos e pelos quais foram identificados ao longo de toda a tese.

QUADRO 3 – Casos selecionados

Número aleatório	**Nome fictício atribuído a cada julgamento**
Caso_1	*Caso dos fetos*
Caso_3	*Caso do passe estudantil*
Caso_4	*Caso da empresa de telefonia*
Caso_7	*Caso da fila*
Caso_8	*Caso da manchete ofensiva*
Caso_9	*Caso do veículo amassado*
Caso_11	*Caso da cédula falsa*
Caso_15	*Caso da publicidade de tabaco*
Caso_16	*Caso do voo cancelado*
Caso_17	*Caso da síndrome da dor*
Caso_18	*Caso do contrato extinto*
Caso_20	*Caso do email difamatório*
Caso_22	*Caso da criança deficiente*
Caso_25	*Caso da malha fina*
Caso_26	*Caso da aluna que cola*
Caso_32	*Caso da biblioteca*
Caso_33	*Caso do posto de gasolina*
Caso_40	*Caso da perna quebrada*
Caso_41	*Caso do remédio para emagrecimento*
Caso_43	*Caso do envelope vazio*
Caso_46	*Caso do apagão aéreo*
Caso_47	*Caso da bagagem extraviada*
Caso_48	*Caso do coletivo*
Caso_59	*Caso do falso HIV*
Caso_60	*Caso do erro médico*
Caso_61	*Caso do estuprador estuprado*
Caso_62	*Caso do concerto de paletó*

Fonte: Elaboração do autor

Importante destacar que numeramos os 27 casos de forma aleatória, apenas para facilitar a referência a esses julgados nos quatro capítulos analíticos. Do mesmo modo, demos um nome fictício para cada

julgamento, com base no teor do julgamento, também para facilitar para o leitor a referência a esses 27 casos. Como se trata de *corpus* longo, esse cuidado foi fundamental para que o leitor não se perdesse ao longo das análises e pudesse acompanhar o raciocínio empreendido ao longo dos capítulos analíticos.

5 Os desafios da dupla transcrição

Chamaremos de 'transcrição' o que Blanche-Benveniste (2008, p. 278) define como "tradução ortográfica das palavras pronunciadas nos áudios". O trabalho de transcrição supõe, certamente, um trabalho de interpretação e de escolhas do transcritor, pois ali se procede à transposição de um código oral para um código escrito. É importante destacar que esse exercício de "pré-análise" dos dados (TRAVERSO, 2002, p. 79), isto é, a transcrição, além de não ser o principal objetivo da pesquisa, é visto meramente como uma etapa, indispensável, para que possamos proceder às análises que serão apresentadas. De todo modo, acreditamos que a transcrição não passa de um reflexo (bastante pálido) das falas transcritas, sendo impossível transpor toda a complexidade de uma fala, de todo um contexto para uma transcrição, quando trabalhamos com interações complexas entre várias pessoas, como é bem o caso na pesquisa empreendida.

Como já mencionado, as análises elaboradas dizem respeito a dados orais, e não a dados escritos, apesar de também termos coletado dados escritos, mas que não fizeram parte da pesquisa aqui relatada. Em realidade, enxergamos as transcrições ali realizadas simplesmente como uma forma de colocar sob os olhos do leitor o que exaustivamente *escutamos*, quando de nossa imersão nos dados selecionados. Temos consciência ainda de que "uma transcrição não passa de uma forma de representação" (GADET, 2008, p. 38); que ela é "essencialmente instável" (MONDADA, 2008, p. 81). Desse modo, em nenhum momento tivemos a preocupação de fazer transcrições perfeitas, pois, como diz Blanche-Benveniste:

> É impossível tratar a língua escrita como uma representação transparente da língua oral; ou seja, não se pode fazer coincidir as unidades das duas representações da língua, nem tampouco alinhar os fenômenos prosódicos do oral com as idiossincrasias do escrito, traduzida em alíneas, frases, sinais de pontuação (...) nesse sentido, *transcrever é empobrecer.* (2008, p. 192, destaque nosso).

Em realidade, considerando-se que não existe um sistema unificado de transcrições (TRAVERSO, 2007, p. 24), procuramos dar conta da tecnicidade dos dados e das incertezas da transcrição – o que é normal quando se trata de dados orais – para criarmos uma versão escrita o mais próxima possível do que escutamos (dados orais), sempre tentando conciliar fidelidade à escuta com a transcrição feita. E neste momento, sabemos, nossa responsabilidade é de fato enorme, pois o que transcrevemos, de certa forma, trará a um público leigo um pouco do que se passa em uma sessão de deliberação em 2ª Instância. No entanto, precisamos já deixar claro, o que tentamos trazer a lume a partir das transcrições não deve, em nenhum momento, servir como rótulo para o que de fato acontece em todos os tribunais do Brasil. Que não se caia em simplificações. Isso porque, o que conseguimos descrever[11] representa muito pouco do que ocorre em sessões de deliberação, isso em relação às interações verbais e argumentativas, e não necessariamente à forma como se aplicam as leis, pois, como já justificamos, o trabalho que transcrevemos e descrevemos em nenhum momento pretendeu emitir juízos de valor à forma como as leis são aplicadas. Nesse sentido, não se pretendeu dizer se os magistrados fazem corretamente ou incorretamente seu trabalho. Seria inclusive leviano de nossa parte pretendê-lo, visto não sermos juristas. O nosso olhar buscou sobretudo, a partir das transcrições e observações empreendidas, compreender como se faz a gestão do desacordo em momentos de conflitos de opinião (mais precisamente, em momentos de *estase argumentativa*, como já explicamos) e isso foi dito claramente desde as primeiras páginas da pesquisa, sob o risco de as análises tornarem-se telhado de vidro.

Desse modo, não se devem buscar análises jurídicas nas análises que empreendemos, sob o risco ou de se fazerem interpretações indevidas, ou de se gerar frustração no leitor eventual da pesquisa. Buscamos, desde o começo, evitar rótulos ou interpretações deslocadas dos magistrados ou do *modus operandi* jurídico, mesmo se, segundo Traverso: "as escolhas feitas por um transcritor ou por um tradutor [em caso de

[11] Importante destacar que, apesar de termos "enxugado" os dados, esse "pouco" foi exaustivamente analisado e com base em critérios metodológicos claros. Nesse sentido, sentimo-nos à vontade para afirmar que as conclusões a que chegamos não foram intuitivas nem superficiais, visto o rigor que apresentamos no manuseio dos dados e, que, inclusive, foi destacado no *rapport* de tese, quando da defesa do trabalho.

línguas e culturas diferentes] possa ajudar a construir uma imagem dos locutores, conferindo-lhes certos atributos (...) podendo reforçar estereótipos" (TRAVERSO, 2002, p. 96). Em resumo, a constituição do *corpus* TRIBUNAL buscou simplesmente atender aos objetivos a que nos propusemos (GADET, 2008, p. 45), os quais já foram apresentados na introdução deste breve artigo.

O processo de transcrição realizou-se em mão dupla, isto é, em uma primeira versão em português e, num segundo momento, em uma tradução do português para o francês jurídico. Para cumprir essa intricada etapa, optamos por adotar as convenções de transcrição do Laboratório ICAR, no qual a pesquisa teve início e foi executada. Naquele laboratório, estivemos insertos na Célula de Corpus Complexos (CCC) – que busca o desenvolvimento de ferramentas e práticas que estejam diretamente ligadas à produção de *corpora* pluridisciplinares e multimodais e em que se consideram "complexos" os *corpora* que envolvam vídeos, sons, textos, traços e imagens cuja exploração necessite de um estudo meticuloso.[12] No Quadro 4, apresentamos algumas das convenções do laboratório ICAR.

QUADRO 4 – Convenções de transcrição – Laboratório ICAR

/	entonação ascendente	( )	transcrição com dúvidas
\	entonação descendente	&	ausência de intervalo entre dois turnos
(.)	pausa breve	=	continuação de turno de fala
(..)	pausa mediana	LAla	ênfase
(...)	pausa longa	:	alongamento
(0.6)	pausa em segundos	-	interrupção
[ ]	sobreposição de vozes	° °	voz baixa
xxx	segmento incompreensível	# #	fala acelerada
((riso))	comentário		

Fonte: elaboração do autor.

Para apresentarmos rapidamente esse desafio de transcrição dupla, mostraremos a seguir um dos excertos transcritos (ao qual retornaremos mais à frente, para falarmos do recorte longitudinal dos dados) primeiro

[12] Para mais informações, conferir a página do laboratório ICAR, na versão original em francês: http://icar.cnrs.fr/recherche/recherche-thematiques_et_axes_transversaux/

na versão em português (os dados estão, originalmente em português, é importante lembrar) e, em seguida, sua tradução/transcrição para o francês. Vamos ao excerto:

QUADRO 6 – Transcrição e tradução de dados

***Corpus* TRIBUNAL/40:2010 - 2min22seg**

Caso da perna quebrada

1	**REL**	senhor presidente eh que:((limpa a garganta)) se investe contra
2		(.) eh: uma sentença que julgou improcedente o pedido por
3		indeniza-é de indenização por dano moral e estético promovida
4		por ((identificação)) o autor admitido nos quadros da extinta
5		((identificação)) cargo de auxiliar de educação de vigilância
6		com a função de viGIa alega que estava de plantão na escola
7		((identificação)) quando foi rendido por três homens portando
8		armas de fogo que o imobilizaram e quebraram a sua perna direita
9		causando lhe deformidade permanente tanto que foi aposentado por
10		invalidez\ ele culpa o estado pelo evento danoso e que deixou de
11		zelar pela segurança de alunos e servidores do estabelecimento
12		de ensino\ pretende indenização de duzentos mil reais (mudança
13		de tom)) #a-aqui senhor presidente pra resumir a senten:ça levou
14		em conta que sendo ele vigiLANTE ele estaria com a-eh eh sob
15		o ris:co constante desses eventos eu estou estabelecendo aqui
16		senhor presidente um:-uma diferença entre vigiLANTE e aquele
17		guarda que realmente estaria ali para a SEGURAN:ça do
18		estabelecimento e de pessoas e assim eu estou PROVENDO o recurso
19		entendendo que é do estado a responsabilidade pelo que aconteceu
20		eh: REformando a sentença e julgando procedente os pedi-o
21		pedido\ condenando o réu ((identificação)) a pagar ao autor
22		vinte mil reais pelos danos morais e dez mil reais pelos danos
23		estéticos corrigidos monetariamente e-etc etc\ condeno ainda ao
24		pagamento de custas e de honorários de MIL reais é assim que
25		estou portanto dando provimento ao recurso do autor senhor
26		presidente\

27	**1V**	o meu voto é com o eminente relator
28		desembargador ((à 2V))/
29	**2V**	com o eminente relator
30	**1V**	°ta razoável esse valor não ta desembargador ((à REL))/°
31	**REL**	°((mudança de tom)) ta\ né°&=
32	**1V**	&°ta razoável/°
33	**REL**	=°apenas a PERNA\ coita:do ele: teve um sofrimento violento né
34		quando se-per:- mas acho
35		que\
36	**1V**	a [apelação
37	**REL**	[vossa excelência tava aumentan:do\ ou diminuin:do\ como é que
38		ta-/ (.) eu to dis[POSto a discutir
39	**1V**	[°nao não\
40		°heim/ não não eu ta:va:°&
41	**REL**	&°ta-°\
42	**1V**	eh: apelação provida unânime

Fonte: Elaboração do autor

Agora, vejamos o mesmo excerto transposto para o francês (a partir da linha 19, pois foi esse o trecho efetivamente analisado):

19	**RAP**	et de cette façon je REÇOIS la demande
20		car à mon avis c'est l'état qui doit être poursuivi pour cette
21		affaire euh: en RÉvisant la sentence et recevant la demande \ je condamne le défendeur ((identification)) au
22		paiement de
23		vingt mille reais en tant que dommages et intérêts et dix mille reais pour le préjudice esthétique et sa compensation monétaire
24		etc etc\ je condamne encore au
25		payement des dépens du procès de MILLE reais c'est comme ça
26		que je considère la demande de l'auteur recevable monsieur le président\

27	**C1**	je vote favorablement au éminnent rapporteur
28		monsieur ((parle à C2))
29	**C2**	je ratifie le vote du éminnent rapporteur
30	**C1**	*°c'est correct ce montant n'est ce pas votre honneur ((parle au* RAP))/°
31	**RAP**	*°((changement de ton de voix)) oui\ n'est ce pas°&=*
32	**C1**	&°oui c'est correct/°
33	**RAP**	=seulement la JAMBE\ le pau:vre lui il: a eu <u>très</u> mal n'est-ce
34		pas quand on l'a frappé mais je crois
35		que\
36	**C1**	le [recours ordinaire
37	**RAP**	[vous pourriez augmenter\ ou bais:ser\ alors qui/
38		(.) je suis [PRÊt à en discuter
39	**C1**	[°non non°\
40		*°comment/ non non j'étais juste:°&*
41	**RAP**	&°d'accord-°\
42	**C1**	euh: recours accueilli à l'unanimité

A partir da transcrição (QUADROS 5 e 6), tivemos também de nos preocupar com a identificação dos locutores, em função de seus papéis ao longo das deliberações. Na transcrição em português utilizamos os seguintes identificadores: **REL** (= RELator), **1V** (= 1º Vogal), **2V** (2º Vogal), tudo isso de forma impessoal e anônima.[13] Desse modo, não nos interessa a identificação real dos relatores ou dos magistrados que atuaram como vogais ao longo dos 27 julgamentos selecionados; bastou-nos apenas identificar a *função* que tal interactante exerceu nas várias sessões de julgamento aqui transcritas. Essa é mais uma forma de preservamos a identidade de todos os envolvidos nas 27 deliberações minuciosamente recortadas para a pesquisa aqui relatada.

Uma sessão de deliberação reúne três ou mais magistrados, de acordo com a natureza do caso em julgamento. Desse modo, no início de cada julgamento, o magistrado relator (REL) expõe seu voto, já

[13] Aqui não há necessidade de detalharmos a tradução para o francês, visto não ser esse o foco deste artigo.

redigido antes da deliberação (no entanto, nada o impede de mudar seu voto durante a discussão do caso). Em seguida, é a vez do primeiro vogal (1V) – ou do revisor (REV) – pronunciar seu voto; por fim, o segundo vogal (2V), e que não conhece o caso tão a fundo como o relator (ou o revisor, caso haja), se pronunciará para que a votação chegue ao final, isto é, para que haja um veredito (por unanimidade ou por maioria). Após a fala de 2V, o presidente da sessão oficializará o resultado, proclamando-o, no melhor estilo austiniano.

Por fim, é importante esclarecer que, nas transcrições realizadas, não se seguiu a pontuação convencional, pois o intuito foi ressaltar algumas características rítmicas, temporais e prosódicas das falas transcritas e que nos pareceram relevantes para descrevermos a gestão do desacordo em deliberações em 2ª Instância, problemática da pesquisa realizada. Tal escolha buscou, mesmo que de forma tênue, aproximar o mais possível a versão transcrita dos votos da realidade falada, com seus cortes abruptos, pausas, gaguejos, sensação de frase mal formulada, incompletude.

6 O recorte transversal e o longitudinal dos dados

Às análises que compõem o trabalho de tese cuja metodologia ora relatamos associamos o método *transversal* e o método *longitudinal* ao longo da leitura que faremos dos dados. Nesse sentido, para Traverso:

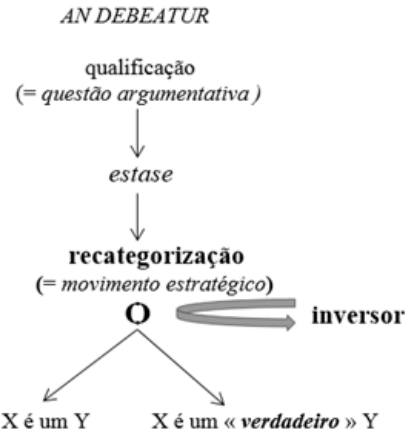
> a *análise transversal* consiste em estudar um fenômeno, previamente identificado, em diferentes situações (em diferentes interações) pertencentes a um mesmo corpus, ligado à busca de respostas a uma mesma problemática geral (...) a *análise longitudinal* busca dar conta de uma interação em uma só sequência, num mesmo momento, considerando o seu início, meio e fim. (TRAVERSO, 2007, p. 27, itálico nosso).

No primeiro caso (análise *transversal*), examinamos um mesmo fenômeno identificado no *corpus* em diferentes excertos/momentos, em diversas situações diferentes, isto é, a regularidade de um fenômeno observado em julgamentos diferentes, com magistrados diferentes, em datas diferentes. No trabalho que apresentamos, a análise transversal foi bastante profícua, pois, por meio dela, identificamos alguns fenômenos interessantes, sobre os quais discorreremos rapidamente ainda neste artigo.

6.1 O recorte transversal

A esse respeito, publicamos dois artigos por meio dos quais mostramos a ocorrência de um mesmo fenômeno, a que chamamos de "recategorização argumentativa".[14] Para tal, tivemos de compreender como os magistrados argumentavam e expunham seus pontos de vista durante uma deliberação. Ao descrever o fenômeno, chegamos à Figura 1.

FIGURA 1 – Processo de recategorização argumentativa – análise transversal de dados



Fonte: Elaboração do autor

Foi graças a um recorte transversal de dados que pudemos chegar à Figura 1, a qual só nos foi possível conceber após a escuta exaustiva de 27 deliberações, dentre as quase 100 previamente selecionadas no início da pesquisa. Tal figura busca explicar que, no *corpus* TRIBUNAL, os processos em segunda instância via de regra suscitam, na primeira etapa de cada julgamento (chamada *an debeatur* pelos juristas), uma *questão argumentativa* (PLANTIN, 2005) do tipo: "Tal ação deve ser considerada como lícita ou como ilícita?". Desse modo, no momento de apresentar seus argumentos, e em caso de conflito de opiniões (*estase*), notamos que os magistrados lançam mão do processo do que chamamos de *recategorização argumentativa*, para fundamentar seus julgamentos

[14] Vide artigos em que apresentamos idiossincrasias do fenômeno que chamei de "recategorização argumentativa", por meio da apresentação de dois estudos de caso (DAMASCENO-MORAIS, 2014a, 2014b).

destoantes ao longo de algumas das deliberações analisadas, como buscamos evidenciar.

Como tentamos explicar, tal movimento estratégico (que chamamos de *recategorização*) só é possível graças à prerrogativa que têm os juristas de (re)interpretarem um mesmo fato de formas diferentes, chegando, dessa forma, a conclusões muitas vezes antagônicas, sem, para tanto, ferir o código, as leis. E o símbolo O, aqui chamado de *inversor*, está ali justamente para mostrar o momento exato em que essa divergência de interpretação aparece ao longo dos julgamentos que fazem parte do *corpus* TRIBUNAL, o qual utilizamos como fonte de pesquisa para a elaboração deste e de outros trabalhos, alguns ainda nem iniciados. Desse modo, no momento da classificação, um mesmo fato pode ser considerado "um X" ou um "verdadeiro X", dependendo da forma como cada magistrado interpretará os fatos, no momento da qualificação de uma ação. E esse momento de qualificação pode, obviamente, trazer conteúdo retórico não negligenciável, como demonstramos minuciosamente na tese em questão.

6.2 O recorte longitudinal

Um exemplo de recorte *longitudinal* pode ser representado por um julgamento visto em sua integralidade (e não recortes de um mesmo fenômeno identificados em julgamentos diferentes). Essa abordagem metodológica pode ser vista, por exemplo, quando explicamos o papel das emoções em um julgamento, analisando apenas um julgamento, do começo ao final. Ali pretendemos entender e explicar as maneiras pelas quais as emoções podem estar entrelaçadas nos discursos jurídicos argumentativos. A partir da transcrição integral de um breve julgamento (ie, do início ao fim) em um Tribunal de Justiça do Brasil tivemos a oportunidade de observar e descrever um pouco dos componentes racionais e emocionais de um julgamento entre desembargadores. "O caso da perna quebrada", como cognominamos aquele julgamento, permite examinar como os juízes definem o valor da indenização a ser paga nos casos de dano moral. Esse julgado foi mostrado anteriormente, no Quadro 5, neste artigo.

Ali indicamos que não apenas de argumentos técnicos se compõe uma decisão; a subjetividade também é importante nesse contexto jurídico. Nessa etapa da pesquisa, confirmamos o que juristas e filósofos

do campo da argumentação, como Cornu (2005), Feteris (1999), Garapon (2001, 2008), Robrieux (2010), Perelman (1999), Stamakis (1995), entre outros, já haviam notado: juízes não são frias máquinas que julgam cegamente. A nossa análise pôde mostrar que uma sentença é uma mistura de regras jurídicas e experiência pessoal dos magistrados, em certa medida. E o recorte feito no julgamento não foi como no caso anterior, isto é, transversal. Em outras palavras, a metodologia que empregamos para fazer o recorte dos dados é diretamente proporcional ao que buscamos responder ou compreender na pesquisa empreendida.

Ainda sobre o ritual, de acordo com o tipo de caso em julgamento, a sessão de deliberação reúne em geral três magistrados. No início de cada julgamento o desembargador relator expõe o caso, geralmente narrando os acontecimentos (o que aconteceu, o resultado do julgamento em Primeira Instância, o tipo de contestação etc.) e, ao final da exposição, apresenta o seu voto (se favorável ou contrário à sentença proferida em Primeira Instância), que já fora preparado, por escrito, antes da sessão. No entanto, o fato de o voto estar escrito não garante que a decisão do relator será acatada pelo grupo; tampouco impede que, em detrimento do tipo de debate realizado, o próprio relator mude o seu voto, durante a deliberação, que ocorre oralmente. Nesse sentido, o primeiro compromisso do relator é, desse modo, ter estudado a fundo o caso, minuciosamente, pois será a partir da narração dos fatos por ele feita e da justificação de seu voto que os demais magistrados votarão.

O produto de cada julgamento chama-se *acórdão*; de forma bastante elementar, pode-se dizer que um acórdão, em Segundo Grau de instrução, equivale ao documento "sentença", em Primeiro Grau de jurisdição. Certamente, pode-se simplificar ainda mais a questão afirmando-se que um acórdão é uma sentença. Não nos podemos esquecer de que, em Segunda Instância, os magistrados julgam casos já julgados por colegas de profissão; seu trabalho consiste, em resumo, em verificar a validade das decisões proferidas por um juiz solitário em Primeira Instância. Faz parte ainda das atribuições do relator de um processo resumir todos os argumentos que foram utilizados por autor e réu, e, ainda, relatar a justificativa que o magistrado utilizara à época do primeiro julgamento (em Primeira Instância) para justificar a sentença outrora proferida. A principal tarefa do desembargador relator será, então, justificar sua decisão, a qual, como vimos, pode ir de encontro à (ou ao encontro da) sentença inicialmente proferida em Primeira Instância.

Após o anúncio do voto do relator/REL, o desembargador que naquele caso atua como primeiro vogal/1V será o segundo magistrado a se pronunciar (sempre após o relator). Isso significa que 1V terá duas possibilidades: ele será a favor ou contra o voto de REL. Em seguida será a vez de 2V (o magistrado que naquele caso atua como segundo vogal). Geralmente o segundo vogal não conhece o processo tão minuciosamente como o relator e, muitas vezes, ele (2V) profere o seu voto com base apenas na exposição dos fatos feita por REL. De qualquer modo, 2V será o último a se pronunciar e, normalmente, após sua fala, uma decisão terá sido tomada; e essa decisão será formalmente proferida pelo presidente da turma na qual a deliberação está sendo realizada. Interessante destacar que pode acontecer de um dos três magistrados (REL, 1V, 2V) acumularem, no julgamento, a função de REL ou 1V ou 2V com a função de presidente da turma, pois se trata de atribuições diferentes.

Em síntese, uma sessão de deliberação em Segunda Instância traz, geralmente, três interactantes: o relator/REL, o primeiro vogal/1V e o segundo vogal/2V (como vimos, o papel de presidente da sessão geralmente é acumulado por um desses três ou pode ser exercido por um quarto magistrado, dependendo da situação). Alguns processos têm um revisor/REV (em vez de dois vogais); no entanto, quando isso acontece, geralmente o trio é formado por REL + REV + V (apenas um vogal, em vez de dois). Ao final de cada caso julgado, os magistrados, por maioria ou por unanimidade, têm a obrigação, seja de concordar com a decisão do juiz de Primeira Instância – o primeiro a julgar o caso em análise –, seja de reformar a sentença daquele magistrado (= discordar), e, nesse caso, terão mudado a sentença original.

Como dissemos, o julgado ao qual estamos a nos referir para ilustrar o recorte longitudinal, foi apresentado anteriormente no Quadro 5. Por falta de espaço, não temos como nos estender na explicação da análise, neste momento. Não obstante, a análise completa pode ser lida em artigo publicado em inglês (o que, é forçoso admitir, dificultou ainda mais o trabalho, porque o excerto teve de ser transposto para o inglês).[15]

[15] Vide Damasceno-Morais (2016).

7 Dominação, poder e lobos

No caso da análise *transversal*, em que identificamos a "recategorização argumentativa" (FIGURA 1), buscamos apontar as categorias *a priori*, isto é, as reações dos interactantes, descrevendo minuciosamente o contexto antes da ocorrência do fenômeno observado; durante sua ocorrência e, ainda, depois da ocorrência da chamada recategorização observada e descrita, ao longo das deliberações entre os desembargadores. Procuramos, ainda, adotar uma postura de ir-e-vir incessante na escuta dos áudios coletados, na iminência de melhor compreender o que ouvíamos. No segundo caso (análise *longitudinal* / Quadro 5), ilustrado pela transcrição do julgamento da "perna quebrada", buscamos analisar uma interação em sua integralidade, isto é, sem fatiamento de excertos, o que nos possibilitou melhor compreender o contexto de ação que antecede e o que sucede o surgimento da *estase* (o desacordo) entre os interactantes. Assim, a mescla dos recortes *transversal* e do *longitudinal* nos permitiu ter uma visão mais versátil do *corpus* com base em escolhas metodológicas claras e detalhadas.

A busca de regularidades linguísticas, por meio de marcas textuais, nos momentos de *estase* argumentativa, nos ajudou a compreender os procedimentos de construção metódica e de estruturação da interação em nível global e local, entre os interactantes. As duas maneiras de recortar os dados permitiram-nos identificar como se estabelecem as relações entre os magistrados em deliberação;[16] pudemos, ainda, perceber a dimensão simbólica de falas, frases, frases entrecortadas, suspiros, pausas (etc.) ocorridas entre os interactantes nos momentos de interação estásica.

Mesmo que o objetivo aqui não seja dizer se os magistrados fazem corretamente o seu trabalho no plano jurídico ou pessoal (como já explicamos), acreditamos que é possível, a partir das análises que aqui relatamos, apresentar novas perspectivas de estudos nesta seara, suscitando a reflexão não só do público diletante, mas também dos próprios profissionais que atuam na área jurídica. E exatamente por isso tentaremos nos abster de juízos de valor (no sentido de avaliar os

[16] Certamente aqui não temos tempo de explicitar minuciosamente todas essas ocorrências e regularidades, as quais ocuparam quatro capítulos de análise da tese aqui citada. A ilustração que fazemos de forma breve neste artigo tem a única função de mostrar como a metodologia nos permitiu encontrar resultados, a partir de análise exaustiva de dados complexos, como já explicamos.

resultados dos julgamentos observados, emitindo algum tipo de (des) contentamento acerta de algum voto proferido), uma vez que nos apoiamos numa perspectiva metodológica basicamente descritiva, a partir de alguns preceitos da Etnometodologia, como já explicado. Restringimo-nos, assim, à escuta e reescuta dos dados, os quais nos permitiram detectar, descrever e analisar fenômenos que nos pareceram significativos.

Os dados selecionados fazem parte de um contexto em que a interação verbal se realiza entre magistrados que, em um eixo vertical, estão no mesmo nível hierárquico. Propositalmente, evitamos a coleta de dados que tivessem interações entre advogados e magistrados como foco; evitamos ainda situações de testemunhas em interação com juízes; réus com advogados etc., como forma de evitar a armadilha de análise de situações assimétricas e que, muitas vezes, descambam num discurso estereotipado sobre dominação, poder e lobos. Buscamos evitar análises que, não poucas vezes, propõem apenas a confirmação das "certezas" iniciais, isto é, as análises que enxergam a sociedade sempre como uma cruel selva de pedra, injusta e infamemente capitalista, que, sobretudo em situação institucional, geralmente é representada por um juiz autoritário e superpoderoso diante do acusado, um patrão canastrão, um médico anti-ético, um gerente inescrupuloso, uma multinacional impiedosa, que, implacável e inexoravelmente, são algozes de um sistema brutal e letal.[17]

Foi para evitar esse tipo de armadilha de pesquisas com conclusões pré-fabricadas que escolhemos uma situação de interação *simétrica*, trilogal, em que os interactantes estão num mesmo patamar institucional e social (todos são desembargadores), portadores igualmente de notório conhecimento técnico e mesmo grau de respeitabilidade (o que é um pré-requisito para o cargo, isto é, uma "reputação ilibada", no jargão jurídico). Obviamente, e para não soarmos *naïfs*, sabemos que, mesmo nesse tipo de situação *simétrica*, como a que selecionamos para o corpus TRIBUNAL, os liames e amarrilhos; os encadeamentos profissionais; o trato; as (in)compatibilidades pessoais entre os magistrados, pessoas da lei, constituem laços complexos, pois, não nos devemos esquecer, trata-se de grupos de pessoas que trabalham juntas, deliberam em conjunto e dividem um tipo de intimidade que mistura o respeito profissional e a questão deontológica a certa cumplicidade, coleguismo, condescendência e até mesmo certa transigência em relação aos votos elaborados por eles,

[17] A esse respeito, ver Damasceno-Morais (2005).

em sua rotina de trabalho. E isso, certamente, está longe de representar o ser imaculado e imparcial alegorizado pela justiça com a venda nos olhos, como problematizamos na tese que deu origem a este breve relato. Não por acaso, a construção da justiça, coletiva – e por isso intricada e complexa –, é um terreno fértil para que mais pesquisas ali mirem a lupa e para que desmitifiquemos os seres de toga (CORNU, 2005; KREUZBAUER, 2007; LATOUR, 2004; MARTINEAU, 2010; POSNER, 2008; STAMAKIS, 1995 entre outros). Mas isso já é assunto para outro artigo. Por ora, justifico apenas a escolha de situações simétricas para o corpus que construímos para o trabalho aqui relatado.

Últimas considerações

Vimos nesta contribuição para a edição especial sobre "Linguística de *Corpus*: conquistas e desafios" – e que esperamos possa de fato contribuir para alguma reflexão – uma oportunidade não de teorizar sobre metodologias diversas,[18] mas de relatar um caso prático de composição, organização, recorte e análise de dados complexos, a partir da investigação de uma problemática que, aqui, se traduz pela descrição e análise da gestão do desacordo entre desembargadores numa Corte de justiça em Segunda Instância. O aspecto metodológico da pesquisa ora relatada testemunha as dificuldades inerentes e específicas à pesquisa empreendida e pode ser vista como um exemplo prático das dificuldades e desafios de se empreender uma pesquisa com *corpora* complexo.

Acreditamos que a importância maior desta experiência é registrar a forma como "o gozo da descoberta" (SALOMON, 2014, p. 154), apesar de ser um combustível necessário para levar um pesquisador a explorar um mundo desconhecido (lembremo-nos do espeleólogo na caverna), não é suficiente. Nesse sentido não basta o "amor aos dados" (como o amor do carpinteiro pela madeira), mas a construção de uma eficaz e clara metodologia de (de)composição e análise de dados,[19] sem

[18] Segundo Coscarelli e Mitre (2007, p. 74), a missão do pesquisador não é teorizar sobre metodologias, "não é ter respostas e soluções, e sim levantar perguntas interessantes".

[19] O *rapport* da tese defendida (espécie de ata circunstanciada da defesa) trouxe comentários avaliativos bastante encorajadores e entusiásticos acerca da metodologia adotada para composição do banco de dados TRIBUNAL, razão pela qual tomamos a liberdade de aqui relatar essa experiência acadêmica neste número especial.

o que corremos o risco de ficar eternamente presos num labirinto do qual jamais retiraremos as informações de que necessitamos. Nesse sentido, se ousamos relatar essa empreitada acadêmica é por acreditarmos que tal experiência pode ser útil a outros pesquisadores que precisem mergulhar num universo desconhecido para descrevê-lo de forma metódica, condição *sine qua non* para a credibilidade da pesquisa, que pode ser longa e complexa, porque uma pesquisa não pode apenas mostrar, ela precisa demonstrar (SERRANO, 2011, p. 15).

Agradecimentos

Agradeço imensamente a Laura Silveira Botelho pela boa vontade na leitura (empolgada e empolgante) deste relato-recorte de pesquisa.

Referências

ATKINSON, J. M.; DREW, P. *Order in Court:* The Organization of Verbal Interaction in Judicial Settings. London: Macmillan, 1979. DOI: https://doi.org/10.1007/978-1-349-04057-5

BAUDE, O. *et al. Corpus oraux, guide des bonnes pratiques*. Orléans: CNRS Editions; Presses Universitaires Orléans, 2006.

BLANCHE-BENVENISTE, C. Les unités de language écrite et de langue parlée. *Cahiers de L'Université de Perpignan*, Perpignan, n. 37, p. 192-216, 2008.

CORNU, G. *Linguistique juridique*. Paris: Éditions Montchrestien, 2005.

COSCARELLI, C. V.; MITRE, D. *Oficina de leitura e produção de textos* (Livro do Professor). Belo Horizonte: Editora UFMG, 2007.

DAMASCENO-MORAIS, R. *O eminente discurso da queda iminente:* o telejornalismo econômico em foco. 2005. 143f. Dissertação (Mestrado em Linguística) – Faculdade de Letras, Departamento de Linguística, Línguas Clássicas e Vernácula, Universidade de Brasília, 2005.

DAMASCENO-MORAIS, R. *Le prix de la douleur:* gestion des désaccords entre magistrats, dans un tribunal brésilien de seconde instance, 2013. 491f. Tese (Doutorado em Ciências da Linguagem) – Faculdade de Ciências da Linguagem, Université Lumière Lyon 2, 2013.

DAMASCENO-MORAIS, R. La recatégorization comme procédé argumentatif dans le domaine juridique. *Argumentation & Analyse du Discours*, Te-Aviv, v. 13, p. 2-16, 2014a. DOI: https://doi.org/10.4000/aad.1808

DAMASCENO-MORAIS, R. Argumentar em campo jurídico e as possibilidades de inversão de uma decisão: o caso da depilação a laser. *EID&A - Revista Eletrônica de Estudos Integrados em Discurso e Argumentação*, Ilhéus, BA, v. 6, p. 153-170, 2014b.

DAMASCENO-MORAIS, R. Emotional Legal Arguments and a Broken Leg. *Ontario Society for the Study of Argumentation/OSSA*, Ontario, v. 1, p. 1-12, 2016.

DUPRET, B. *Le jugement en action*: ethnométhodologie du droit, de la morale et de la justice en Egypte. Paris: Droz, 2006.

FETERIS, E. T. *Fundamentals of Legal Argumentation:* A Survey Theories on the Justification Judicial Decisions. Netherlands: Kluwer Academic Publishers, 1999. DOI: https://doi.org/10.1007/978-94-015-9219-2

GADET, F. L'oreille et l'oeil à l'écoute du social. *Cahiers de L'Université de Perpignan*, Perpignan, n. 37, p. 35-48, 2008.GARAPON, A. *Bien juger.* Essai sur le rituel judiciaire. Paris: Éditions Odile Jacob, 2001.

GARAPON, A.; ALLARD, J.; GROS, F. *Les vertus du juge*. Paris: Dalloz, 2008.

GARFINKEL, H. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall, 1967.

GIL, A. C. *Como elaborar projetos de pesquisa*. São Paulo: Atlas, 2002.

KERBRAT-ORECCHIONI, C.; PLANTIN, C. *Le trilogue*. Lyon: Presses Universitaires de Lyon, 1995.

KERBRAT-ORECCHIONI, C. *Análise da conversação*: princípios e métodos. São Paulo: Parábola, 2006.

KREUZBAUER, G. Modelling Argumentation in Moral and Legal Discourse. *In*: CONFERENCE OF THE INTERNATIONAL SOCIETY FOR THE STUDY OF ARGUMENTATION, 6., 2007, Amsterdam. *Proceedings* […]. Amsterdam: Sic Sat & International Center for the Study of Argumentation, 2007. p. 827-834.

LATOUR, B. *La fabrique du droit*: une ethnographie du Conseil d'État. Paris: La Découverte, 2004.

MARCUSCHI, L. A. *Análise da conversação*. São Paulo: Ática, 2003.

MARTINEAU, F. *Petit traité d'argumentation judiciaire*. 4. ed. Paris: Praxis Dalloz, 2010.

MARTÍNEZ, E. G. *Flagrantes auditions:* échanges langagiers lors d'interactions judiciaires. Berne: Peter Lang, 2007.

MARTINS JÚNIOR, J. *Como escrever trabalhos de conclusão de curso*. 6. ed. São Paulo: Vozes, 2012.

MONDADA, L. Documenter l'articulation des ressources multimodales dans le temps: la transcription d'enregistrements videos d'interactions. *Cahiers de L'Université de Perpignan*, Perpignan, n. 37, p. 127-155, 2008.

MOTTA-ROTH, D.; HENDGES, G. R. *Produção textual na universidade*. São Paulo: Parábola Editorial, 2010.

OCHS, E.; SCHEGLOFF, E. A., THOMPSON, S. A. *Interaction Grammar.* Cambridge: Cambridge University Press, 1996. DOI: https://doi.org/10.1017/CBO9780511620874

PERELMAN, C. *Logique juridique:* Nouvelle rhétorique. Paris: Éditions Dalloz, 1999.

PLANTIN, C. *Lieux communs, topoi, stereotypes, cliches*. Paris: Éditions Kimé, 1993.

PLANTIN, C. L'argument du paralogisme. *Hermès: Cognition, Communication, Politique*, [*S.l.*], n. 15, v. 1, p. 245-262, 1995. DOI: https://doi.org/10.4267/2042/15170

PLANTIN, C. *L'argumentation.* Paris: Le Seuil, 1996. (Mémo)

PLANTIN, C. *L'argumentation.* Paris: PUF, 2005. (Que sais-je?)

PLANTIN, C. *Dictionnaire de l'argumentation:* une introduction aux études d'argumentation. Lyon: ENS Éditions, 2016.POSNER, A. R. *How Judges Think*. London: Harvard University Press, 2008.

ROBRIEUX, J. *La rhétorique et argumentation*. Paris: Armand Colin, 2010.

SALOMON, D. V. *Como fazer uma monografia.* 13. ed. São Paulo: Editora WMF Martins Fontes, 2014.

SCHEGLOFF, E. A. *Issues of Relevance for Discourse Analysis*: Contingency in Action, Interaction and Co-Participating Context. *In*: HOVY, E. H.; DONIA, R. S. (org.). *Computational and conversational discourse*: Burning issues – An interdisciplinary account. Berlin: Springer Editor, 1996. p. 3-35. DOI: https://doi.org/10.1007/978-3-662-03293-0_1

SCHEGLOFF, E. A. What next?: Language and Social Interaction Study at the Century's Turn. *Research on Language and Social Interaction,* [*S.l.*], v. 32, n. 1, p. 141-148, 1999. DOI: 10.1207/S15327973RLSI321&2_17

SERRANO, F. P. *Pesquisar:* a tese, um desafio possível no labirinto. São Paulo: Parábola, 2011.

STAMAKIS, C. *Argumenter en droit:* – une théorie critique de l'argumentation juridique. Paris: Publisud, 1995.

TRAVERS, M. Ethnomethodologie, analyse de conversation et droit. *Droit et Société*, [*S.l.*], n. 48, p. 349-369, 2001. DOI: https://doi.org/10.3917/drs.048.0349TRAVERSO, V. Transcrire l'interaction – Transcription et traduction des interactions en langue étrangère. *Cahiers de Praxématique*, [*S.l.*], n. 39, p. 77-99, 2002. DOI: https://doi.org/10.4000/praxematique.1836

TRAVERSO, V. *L'analyse des conversations.* Lyon: Armand Colin, 2007.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 711-737, 2021

Diseño de corpus específicos para el estudio histórico gramatical: el caso de las construcciones con clítico femenino

The creation of specific corpora for the historical study of grammar: the case of constructions with the feminine clitic

Nicolás Arellano
Universidad de Buenos Aires (UBA), Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Buenos Aires / Argentina
nicolas.a.arellano@gmail.com
http://orcid.org/0000-0002-5197-5428

Resumen: Este artículo busca analizar las ventajas de una aproximación a los datos lingüísticos históricos a partir de la confección de corpus específicamente diseñados. Para ello, en primer lugar, se presentan las principales limitaciones de los corpus generales de referencia, particularmente del CORDE (RAE) y *Corpus del Español* (BYU), no solamente en cuanto al acceso de sus motores de búsqueda, sino también a la disponibilidad de los textos que los componen. En segundo lugar, se hace uso de un caso en específico, el del origen y desarrollo de las construcciones con clítico femenino, para ilustrar la propuesta. A continuación, se contrasta esta propuesta de abordaje de datos con otras investigaciones que utilizan corpus generales. Se evidencia, así, que este modo de acceso a las emisiones lingüísticas históricas favorece el estudio de procesos gramaticales novedosos de interfaz que se circunscriben a ámbitos informales, populares, orales y diatópicamente marginales.

Palabras clave: locuciones idiomáticas; clítico femenino; corpus generales; corpus históricos; español rioplatense.

Abstract: The aim of this article is to analyze the advantages of an approach to historical linguistic data based on the creation of specifically designed corpora. For this, the main limitations of general reference corpora are presented in the first place; particularly of CORDE (RAE) and Corpus del Español (BYU). The limitations are not only presented with regard to the access on their search engines, but also to the availability of texts that are part of said corpora. Secondly, a particular case is utilized, which shows the origin and development of the constructions with the feminine clitic, so as to exemplify the proposal. Next, the mentioned proposal on the approach of

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.711-737

data is contrasted to other research that use general corpora. Thus, it is demonstrated that this way of accessing historical linguistic utterances benefits the study of novel grammatical interface processes that deal exclusively with informal, popular, oral, and dialectically peripheral fields.

Keywords: idioms; feminine clitic; general corpora; historical corpora; Río de la Plata Spanish.

Recebido em 7 de setembro de 2020

Aceito em 26 de outubro de 2020

1 Introducción

Con la afluencia del giro computacional de las ciencias sociales y humanas (DE MATTEIS, 2015), el acceso y desarrollo de bases de datos en línea, corpus electrónicos generales y de referencia, así como la multiplicidad de nuevos géneros y prácticas discursivas en Internet –correos electrónicos, mensajería instantánea, entradas en redes sociales, comentarios en periódicos, intervenciones en foros, por citar algunos–, han cobrado un lugar preponderante dentro de la investigación del lenguaje (VELA DELFA; CANTAMUTTO, 2015). Este escenario no es nuevo y se enmarca al mismo tiempo en una realidad histórica: las investigaciones filológicas y de las primeras décadas del siglo XX fueron precursoras del análisis lingüístico a partir de la evidencia empírica directa (BERBER SARDINHA, 2000). A partir de la década de 1960, con la incorporación de ciertas herramientas computacionales y la progresiva pérdida de la influencia de los métodos formales en los estudios gramaticales, el trabajo con datos lingüísticos auténticos comenzó a ganar cada vez más terreno (BIBER, 1990; ROJO, 2008). Si bien actualmente existe consenso en los beneficios de acercarse a las hipótesis lingüísticas de una manera que no excluya mutuamente a las dos corrientes (FILLMORE, 1992; LÜDELING; KYTÖ, 2008), esto es que tenga presente tanto fuentes directas como armado de oraciones sin necesariamente una constatación empírica, la obligación del acceso a los datos, a los "hechos del lenguaje" (WEISSER, 2016), se hace más evidente ante algunos tipos de fenómenos gramaticales.

Los avances, si bien importantes, con relación a los grandes corpus de referencia no se limitan en los alcances de las herramientas computacionales y de la lingüística de datos reales. De manera sostenida,

a partir de la década de 1990 (ROJO, 2008), la disponibilidad de la web como corpus lingüístico potencialmente infinito y el diseño y puesta en práctica de corpus específicos de discursos especializados han hecho avanzar la disciplina y la variedad de las investigaciones de forma notable. La lingüística de corpus resulta así una metodología en cierta manera joven, que permite al lingüista desarrollar un entendimiento más acabado de cómo es y cómo funciona el lenguaje a partir de la consulta cualitativamente significativa de muestras de lengua reales, provenientes de diversas fuentes, en grandes cantidades. Los corpus, que deben ser colecciones electrónicas organizadas de texto, seguir criterios de representatividad y ser guiados por parámetros puntuales, constituyen las herramientas específicas desde las cuales se accede a la verificación de una hipótesis en particular que motiva la investigación o desde las cuales se indaga sobre una problemática de la que aún no se tienen muchos datos o información (PARODI, 2008; SILVÉRIA OLIVEIRA, 2015).

La siguiente investigación se centra en el análisis de las herramientas de la lingüística de corpus que pueden aplicarse a fenómenos gramaticales novedosos e históricos que ilustran algún tipo de liminalidad entre los niveles de la lengua, sobre todo entre el morfosintáctico y el léxico. En particular, se analizan los inconvenientes que se desprenden de la utilización de los métodos tradicionales en la investigación del desarrollo histórico de las construcciones pronominales con clítico femenino –*arreglárselas*, *pegarla*, *susanearla*– (ARELLANO, 2020; MARE; CASARES, 2017; MASULLO; BÉRTORA, 2014), las cuales cuentan con una expansión marcada en la variedad rioplatense; de ellos también se ha propuesto su presencia en las variedades ibéricas (CIFUENTES HONRUBIA, 2018). En la sección 2, en primer lugar, resumimos la historia y abordamos las características de los corpus generales de referencia. En segundo lugar, focalizamos sobre las propiedades características de los corpus históricos en general y de los dos corpus históricos del español más importantes –*Corpus del Español* y CORDE– en particular. En la sección siguiente, presentamos el fenómeno de las construcciones con clítico femenino y analizamos críticamente una propuesta de investigación a partir de corpus generales. En la sección 4, mostramos las bases de la creación de una base de datos adaptada para la investigación de este fenómeno. En la sección 5, se contrastan los resultados obtenidos a partir de la utilización de ambos tipos de corpus. En la última sección, presentamos las conclusiones de la investigación.

2 Corpus generales: características y limitaciones

En cuanto a los corpus generales de referencia, es decir, las bases de datos lingüísticas en línea constituidas por millones de palabras, estos posibilitan el acceso a muestras de lengua de una manera sencilla y sin la obligación de obtener elicitaciones o recuperar contenido desde fuentes primarias o desconocidas. Al mismo tiempo, la sistematización de las producciones de los hablantes permite que otras variables metalingüísticas, a menudo ignoradas, puedan ser tenidas en cuenta a la hora de buscar el dato lingüístico necesario para la investigación. Los grandes corpus generales disponibles en línea no constituyen solamente una colección de textos accesibles digitalmente,[1] sino que representan fundamentalmente muestras de lengua anotadas según una variedad de categorías útiles para una multiplicidad de investigaciones y enfoques lingüísticos (DAVIES, 2009). En la mayoría de ellos, puede encontrarse información relacionada con la situación de comunicación, el género discursivo, la variedad del español y el lugar y momento de emisión,[2] por citar algunas características. Debido a la facilidad de acceso que representan los grandes corpus, muchas investigaciones del lenguaje han relegado la inclusión del "discurso real" a la búsqueda de material lingüístico en estas bases de datos. De manera inductiva, deductiva y cuantitativa, el lingüista reflexiona e hipotetiza a través de los datos presentes en los corpus y asume la existencia, productividad y frecuencia de los fenómenos a través de los resultados obtenidos.

---

[1] Conviene diferenciar, así, los archivos y las bibliotecas electrónicas (los primeros, depósitos de textos sin organización; los segundos, colecciones con algún tipo de criterio, sobre todo de género textual o no estrictamente lingüístico) de los corpus o los subcorpus, que constituyen una parte específica de un archivo o una biblioteca organizadas con un diseño y unos objetivos explícitos (BERBER SARDINHA, 2000).

[2] Nuevos proyectos lingüísticos orientados al análisis de discurso interaccional, las humanidades digitales y la sociolingüística han comenzado a incluir otras variables en la notación de sus bases de datos, al mismo tiempo que implementan medidas y decisiones para obtener el consentimiento de sujetos y/o anonimizar la identidad de los enunciadores. En este tipo de empresas se suele agregar la información tipológica (según el canal de comunicación, escrito u oral), metodológica (disponible o elicitado), genérica, de la situación de comunicación y de las características identitarias básicas de los interlocutores (DE MATTEIS, 2015; HUNSTON, 2008; VELA DELFA; CANTAMUTTO, 2015). Los datos que no se recuperan de corpus lingüísticos generales en la investigación fueron sistematizados siguiendo estos lineamientos lingüísticos y éticos.

Sin embargo, aunque la aparición de los grandes corpus lingüísticos representa sin dudas mejores posibilidades para el desarrollo de determinadas investigaciones, estos no dejan de traer aparejados ciertos inconvenientes y limitaciones, sobre todo en relación con la elección de los textos que los componen. Como dentro de sus objetivos normalmente se asume la compilación de la mayoría de los aspectos de una lengua o por lo menos una variedad de ella, generalmente se han señalado las incompatibilidades del estudio de la variación, debido a la neutralización que se deriva al intentar dar cuenta de muchos tipos de discursos simultáneamente (RISSANEN, 2008). Más allá de esta crítica, que se corresponde con las primeras instancias del desarrollo de la disciplina, en las que la cantidad de palabras y textos estaban sujetos a la posibilidad de alojar determinada cantidad de textos, se han presentado también otros problemas.

Por una parte, se estima que los corpus están influenciados en gran medida por contextos escritos (ROMAINE, 2008; WEISSER, 2016). Como consecuencia, los discursos orales se encuentran, con suerte, subrepresentados o directamente no son tenidos en cuenta (CLARIDGE, 2008). En el caso de aparecer, suelen estar relacionados con ámbitos cultos y formales. Muchos corpus cuentan con ejemplos de discursos públicos, disertaciones y clases orales, por citar algunos casos, que generalmente pueden ser actos orales de un discurso previamente escrito (HUNDT, 2008). Por otra parte, como consecuencia de la ausencia de textos del lenguaje hablado, también escasean muchos tipos de textos que, además de ser orales, suelen suceder en interacción y/o son de carácter popular o informal. Este tipo de discursos, al estar raramente disponibles por escrito, ser difíciles de transcribir e involucrar participantes que generalmente no son tenidos en cuenta por los investigadores, suelen ser abandonados. Esta falta afecta el trabajo en general de cualquier investigador de la lengua, pero sobre todo de un conjunto de áreas, entre ellas la gramática, la dialectología, la sociolingüística, la lexicografía y los estudios de cambio lingüístico.

En relación con esto último, efectivamente el panorama desfavorable se acrecienta cuando se lidia con corpus históricos, y no emisiones lingüísticas actuales, en los que el acceso al material de forma representativa se vuelve aún más complejo por razones evidentes. De por sí, los corpus históricos suelen ser más pequeños que los sincrónicos (CLARIDGE, 2008). Asimismo, la disponibilidad de registros orales

grabados incluso durante la primera mitad del siglo XX es realmente exigua. En caso de existir ejemplos, es posible no contar con toda la información lingüística y metalingüística necesaria para llevar a cabo algunas investigaciones, debido a la calidad del material audiovisual o a las dificultades en el rastreo de los participantes de las emisiones registradas, entre otras cuestiones. Los ejemplos anteriores al siglo XX se concentran especialmente en transcripciones de discursos orales, en los que también suelen abundar los registros más formales o los usuarios más cultos. Así, son las fuentes literarias como un conjunto las que ocupan gran parte del total de textos que pertenecen a las bases de datos históricas.

En la actualidad, se cuenta con dos grandes corpus históricos disponibles en español:[3] el Corpus Diacrónico del Español (CORDE),[4] administrado por la Real Academia Española; y el *Corpus del Español Genre/Historical*, desarrollado por Mark Davies y la Universidad de Brigham Young (BYU).[5] En ambos, los textos analizados van desde los primeros registros escritos con los que cuenta el español (en el siglo XI) hasta mediados del siglo XX. En términos de cantidad de palabras, resultan ambos proyectos de importante capacidad, incluso con un total de *tokens* superior al *Helsinki Corpus*, una de las bases de datos más extensas para el análisis histórico del inglés (DAVIES, 2009).

Para el caso del *Corpus del Español*, hasta el año 1900, está compuesto de 1670 textos, manuscritos o facsímiles, particularmente tomados de textos literarios disponibles de la Biblioteca Virtual *Miguel*

---

[3] Estos no constituyen los únicos dos corpus disponibles en línea, sino solo los más importantes en cuanto a cantidad de palabras, influencia y extensión en las investigaciones, debido a su gratuidad y las universidades e instituciones que los desarrollan. Para trabajar con corpus que fundamentalmente se ocupan de variedades ibéricas desde un punto de vista histórico, pero que se centren en variedades no formales, se recomienda consultar, por ejemplo, *PostScriptum* (Centro de Lingüística de la Universidad de Lisboa), que se destaca por ser un corpus histórico que incluye epístolas, cartas y géneros escritos informales. El Corpus Diacrónico y Diatópico del Español de América (CORDIAM), coordinado por la Academia Mexicana de Letras y dirigido por Concepción Company Company, es quizá la única excepción a la tendencia peninsular. Otros corpus históricos son relevados por Contreras Seitz (2009) y Enrique-Arias (2012).

[4] *Corpus diacrónico del español* (CORDE). Recuperado de: http://corpus.rae.es/cordenet.html. Acceso en: 6 set. 2020.

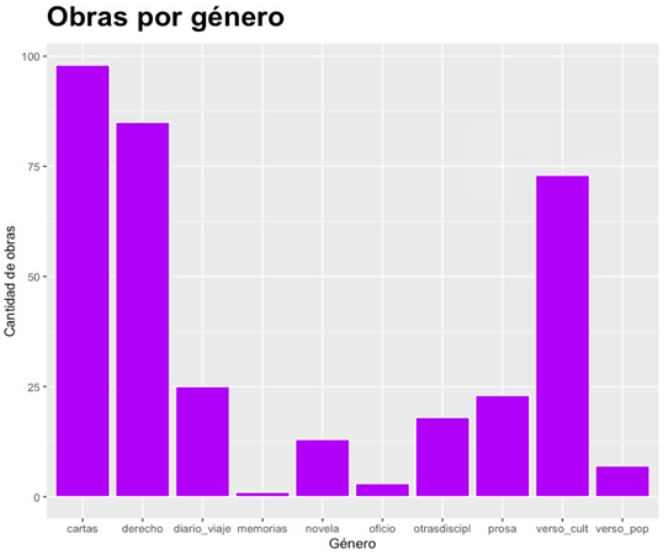
[5] *Corpus del Español Genre/Historical*. Recuperado de: https://www.corpusdelespanol.org/hist-gen/. Acceso en: 6 set. 2020.

de Cervantes y otros proyectos recopilatorios literarios electrónicos. Se destaca la presencia de 392 novelas para el período que abarca el siglo XIX (DAVIES, 2002). La presencia de otros registros, ya sea otro tipo de literatura, textos periodísticos u orales, comienza a realizarse a partir del siglo XX, por lo que el equilibrio que se sostiene mantener se corresponde únicamente con las décadas finales de la franja temporal elegida, esto es los últimos años del siglo XX, sobre todo la década de 1990. No obstante, entre los puntos débiles, se encuentra la utilización de fuentes orales, las cuales se basan principalmente en el ámbito del dominio público de personalidades importantes de Iberoamérica, por lo que se destacan entrevistas a representantes de partidos políticos, intervenciones en sesiones legislativas y discursos y notas periodísticas de diarios españoles de tirada nacional.

Al mismo tiempo, no se especifican de manera extensiva las variedades del español que se toman en cuenta en los textos del siglo XIX ni en otras épocas, por lo que no puede determinarse el porcentaje de textos latinoamericanos y si estos tienen relación con el criterio poblacional que sigue adelante el equipo de la Universidad de Brigham Young en la selección de textos para sus corpus sincrónicos. En estas bases, al contrario del corpus histórico, se constata una clara mayoría de textos y cantidad total de palabras de variedades latinoamericanas, en concordancia con el menor peso demográfico que aporta la península ibérica.

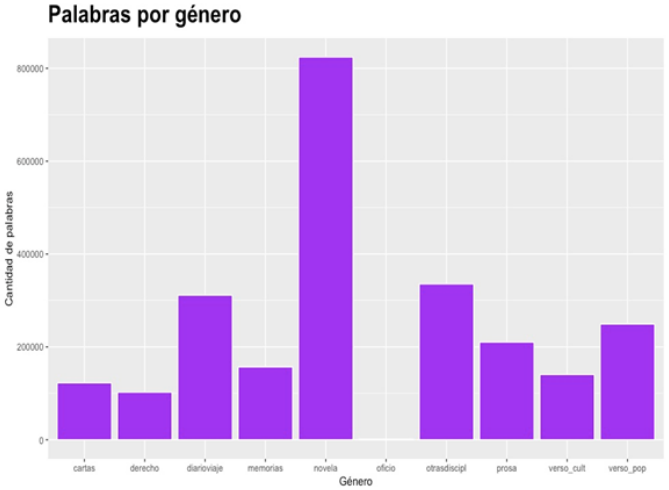
En relación con el Corpus Diacrónico del Español (CORDE), este se compone de 5500 "grandes obras" de la lengua española, con un 74% de datos provenientes de España, y un 25% correspondiente a América Latina (1% corresponde al judeoespañol), priorizando la delimitación de 6 grandes áreas diatópicas (ROJO, 2008; SÁNCHEZ SÁNCHEZ; DOMÍNGUEZ CINTAS, 2007). Cuenta con un total de 180 millones de palabras (PARODI, 2008). En principio, el peso de las obras literarias parece ser pequeño, si se tiene en cuenta que solamente el 44% de los textos se corresponde con literatura. De hecho, cuando se observa la segmentación según el tipo textual, género discursivo o área del conocimiento que compone el corpus, se observa que las cartas, los tratados de derecho, los versos cultos y el diario de viaje se corresponden con las cuatro entradas más frecuentes en el corpus (FIGURA 1). Sin embargo, cuando se analiza el peso por cantidad de palabras, la situación cambia drásticamente: el género novela aporta la mayoría de ellas (FIGURA 2).

FIGURA 1 – Gráfico de confección propia a partir del número de obras por género en Corpus Diacrónico del Español (CORDE) sección Argentina.



Fuente: Datos disponibles en: http://corpus.rae.es/cordenet.html

FIGURA 2 – Gráfico de confección propia a partir del número de palabras por género en Corpus Diacrónico del Español (CORDE) sección Argentina.



Fuente: Datos disponibles en: http://corpus.rae.es/cordenet.html

Al contrario del *Corpus del Español*, las posibilidades de búsqueda en el CORDE para investigaciones gramaticales más allá del léxico, sobre todo relacionadas con fenómenos morfosintácticos y semánticos, se ven reducidas por la ausencia de lematización y de anotación específica de clases de palabras más allá de las principales (BUENAFUENTES DE LA MATA; SÁNCHEZ LANCIS, 2012; DAVIES, 2009). A pesar de estas observaciones, aun así los corpus cuentan con más dificultades para estudios de tipo morfológico o sobre unidades perifrásticas (DAVIES, 2002). En este escenario, los estudios sobre afijos poco regulares, no monosémicos o con restricciones semánticas particulares (por ejemplo, en los que la adición de determinado afijo quede determinada al aspecto o la valencia del verbo) representan un problema adicional que la sintaxis de expresiones regulares no siempre puede saldar.

Este tipo de fenómenos no son poco frecuentes en las lenguas del mundo (BYBEE, 1985) ni en español, por lo que muchos procesos morfosintácticos, incluso de alta frecuencia, no pueden ser buscados con facilidad a través de algunos corpus. Hasta en las búsquedas más avanzadas siempre se presume la existencia de una equiparación única uno a uno entre una determinada palabra y su clase. Esta presuposición supone un problema para palabras homomórficas o con distinto criterio ortográfico, préstamos y, sobre todo, para palabras funcionales, que pueden ser categorizadas como artículos, partículas, pronombres y clíticos, todas nomenclaturas correctas dependiendo de los contextos gramaticales en los que aparecen o el sistema de notación que se siga. Esta particularidad resulta principalmente evidente para los clíticos, cuya proclisis o enclisis en combinación con determinados verboides o bien cuenta con más de una posibilidad de realización o bien no presenta restricciones gramaticales tan claras en algunos contextos.

De este modo, se evidencia la necesidad de un método de aproximación a los datos que pueda ser aprovechado desde el punto de los investigadores y la implementación de soluciones creativas más allá de los grandes corpus de referencia y de los textos electrónicos más fácilmente disponibles (BUENAFUENTES DE LA MATA; SÁNCHEZ LANCIS, 2012; ENRIQUE-ARIAS, 2012). A continuación, se presenta un fenómeno particular de la variedad rioplatense, de amplio desarrollo actual e histórico, que ilustra las problemáticas, limitaciones y características desarrolladas hasta ahora.

3 Construcciones verbales con clítico femenino: aproximación desde los corpus generales

Las locuciones verbales con clítico femenino *la* no anafórico, inherente o marginal constituyen un proceso novedoso de formación de palabras en el que el pronombre acusativo femenino *la* puede unirse a bases patrimoniales y neológicas del español. De esta manera, se pueden formar o bien construcciones más bien idiomáticas (1), en las que el significado de las bases difiere en cierta medida de la locución formada, o bien un tipo de unidad que sirve para intensificar (ALBELDA MARCO, 2004) el valor de las acciones descritas (2).

(1) Sepan que si **la pego**, les traigo a todos conmigo al éxito (Facebook, 2015).

(2a) ¿Hace falta **cancherearla**? (Facebook, 2019).

(2b) Sé que el disco **la flopeó** mal, pero es un buen disco (Twitter, 2014).

(2c) Quiero contarte que empecé a **mariekondearla** y saqué la mitad de la ropa (Twitter, 2017).

En el caso de (1), el pronombre *la* en la construcción *pegarla* (‘tener éxito, generalmente con suerte’) contribuye a la formación de una locución con un significado diferente al valor general asociado al lexema *pegar*. En los ejemplos de (2), el clítico aporta un matiz elativo que sirve para maximizar la forma de realización del evento descrito. Todos los casos de este grupo pueden funcionar –y, de hecho, se registra así– sin el pronombre *la*: *cancherear* (‘actuar como un canchero’, es decir, fanfarronear), *flopear* (del inglés *to flop* ‘fracasar comercialmente’) y *mariekondear* (‘actuar como Marie Kondo’, es decir, ordenar).

Este tipo de construcciones generalmente son proferidas por hablantes jóvenes y pueden encontrarse fácilmente en el discurso de las redes sociales o en las interacciones propias de la mensajería instantánea. Asimismo, también están presentes en discursos plenamente orales (ARELLANO, 2020).

Si bien los últimos estudios (ARELLANO, 2020; ARIAS, 2018; SILVA GARCÉS, 2017, por citar algunos) se han centrado sobre el valor sincrónico y el estado actual de las construcciones en la lengua, otras investigaciones de carácter más o menos reciente han señalado tangencialmente la pertenencia histórica del fenómeno. Para la variedad

argentina, Gobello (1991), Conde (2011, 2013) y Ghio y Albano (2013) han recuperado algunas construcciones de la primera mitad del siglo XX al establecer una relación entre este tipo de locuciones y el léxico lunfardo rioplatense. Se pueden observar también algunos ejemplos que se forman a partir de verbos patrimoniales en combinación con otros pronombres femeninos y reflexivos, que no son exclusivamente el clítico *la*.

(3) Milonguera, bullanguera, que **la va de** alma de loca (FONT, 1927 *apud* GHIO; ALBANO, 2013, p. 111)

(4) **enchufarla, saberla lunga** (CONDE, 2013, p. 100)

(5) El tipo **se las trae** (GHIO; ALBANO, 2013, p. 114)

(6) El indio no le da ni de comer a la mujer, que **se las tiene que rebuscar** sola (GOBELLO, 1991, §Rebuscar)

Pese a estas menciones, encontradas sobre todo en letras de tango y refranes, el carácter histórico del fenómeno de creación de construcciones verbales con clítico femenino no ha sido abordado desde una perspectiva filológica y diacrónica. Generalmente se asume así la presencia de este tipo de locuciones y no se toman en cuenta de manera sistemática otras descripciones de unidades y las distintas entradas en diccionarios generales que recuperan algunas unidades lexicalizadas (*DLE*, *DAMER*, entre otros).

En una primera aproximación, y como ya fue mencionado, surgen algunas complicaciones para la búsqueda y ampliación de los datos lingüísticos de este fenómeno en corpus generales. Los clíticos se homologan a los artículos definidos y los pronombres anafóricos de objeto directo. En efecto, pueden aparecer unidos ortográficamente a un verboide o separados cuando se relacionan con verbos conjugados o en construcciones con infinitivo o gerundio. Asimismo, los ejemplos se encuentran en gran medida en discursos del tipo fenoménico, en los que la incidencia del factor coloquial y oral está más presente. Los registros, géneros textuales y variedades dialectales en las que esperamos encontrar este tipo de ejemplos también escasean en representación.

De todos modos, existen investigaciones que hacen uso de los corpus lingüísticos clásicos para analizar la naturaleza de las construcciones con clítico femenino. Cifuentes Honrubia (2018) constituye uno de esos casos. En este trabajo, el investigador hace un relevamiento histórico de un considerable conjunto de construcciones que

combinan los pronombres *la*(*s*) y *se*, o en los que simplemente aparece alguna de las formas del clítico pronominal femenino, que abarca desde los primeros registros del español hasta la actualidad. Esto se lleva a cabo con el uso del CORDE[6] de manera general y el uso esporádico del CREA (*Corpus de la Real Academia Española*, el corpus sincrónico de la RAE), limitándose al estudio del español europeo (CIFUENTES HONRUBIA, 2018, p. 13). Aparecen, de todas maneras, un buen número de ejemplos de variedades latinoamericanas en su argumentación, especialmente a medida que los ejemplos encontrados se acercan al presente.

Más allá de las conclusiones que establece la investigación acerca del origen y desarrollo de la construcción, metodológicamente no está libre de obstáculos. Cifuentes Honrubia (2018, p. 13) reconoce las dificultades del estudio histórico "por la propia complicación del manejo de corpora". Así, el lingüista no limitó su trabajo al uso de los corpus mencionados, sino que, en una primera instancia, dio cuenta de una "labor de rastreo y acreditación de las distintas construcciones consideradas" por medio de la confección de una lista de construcciones que obtuvo "a través de trabajos lexicográficos y fraseológicos". Más allá de estas líneas, no se precisan otras reflexiones o indicaciones acerca de la obtención de los ejemplos históricos y literarios que registra en su investigación.

Las características, tanto del fenómeno como del motor de búsqueda del CORDE, contribuyen a la decisión de comenzar las investigaciones a partir de otras fuentes. Enrique-Arias (2012) particularmente señala dos limitaciones que se desprenden de la utilización directa de este tipo de corpus generales. La primera es la noción de perspectiva, es decir, relacionada con la forma de acceso a los datos y es de la que da cuenta Cifuentes Honrubia. Efectivamente, para conocer las formas de antemano que el investigador intenta buscar, se necesita contar con un acceso anticipado a través de gramáticas, diccionarios, lexemarios o algún tipo de documento o lista organizados con el criterio deseado. El otro problema es el de la comparabilidad. Este se relaciona con la idea de que los cambios lingüísticos se suceden en distintas etapas: "un estadío original anterior al cambio, una fase en la que triunfa la nueva estructura

[6] Por las similitudes en la conformación de ambos corpus, nos concentramos únicamente en el análisis derivado del Corpus Diacrónico del Español (CORDE). Al igual que lo que ocurre con este corpus, los resultados obtenidos a partir de la utilización de una base propia presentan pocas coincidencias con las unidades que se pueden obtener a través del motor de búsqueda del *Corpus del Español* (BYU).

y una etapa intermedia en la que coexisten el sistema innovador y el original" (ENRIQUE-ARIAS, 2012, p. 89).

En investigaciones que intentan dar cuenta del desarrollo de un fenómeno a lo largo de un considerable período de tiempo, resulta fundamental seguir un horizonte de equivalencia entre los distintos textos que conforman el total de la base de datos a desarrollar. De esta forma, se puede dar cuenta de una manera más acabada acerca del proceso de expansión y retracción de determinadas formas lingüísticas.

Con todo, la investigación con corpus debe cumplir en el mejor de los casos una doble misión: una primera, heurística, dedicada a conocer más acerca del fenómeno a través de las unidades formales que la representan; y una segunda, que permita ir desde la comprensión del fenómeno a la búsqueda de instancias que la verifiquen.

4 Base de datos propia: lineamientos generales

Puede pensarse –y con razón– que los corpus generales ofrecen una gran cantidad de ventajas que pueden ser difíciles de compensar. En principio, son herramientas que llevan años de desarrollo y exitosamente han contribuido a un sinfín de investigaciones. Además, cuentan con una selección criteriosa de géneros y contienen una gran cantidad de textos en su conformación. Sobre todo en relación con la representatividad, se cree que los corpus generales resuelven la cuestión del equilibrio de textos y formas a partir de la concentración de un gran volumen de palabras. No obstante, entendemos que no hay criterios objetivos para la determinación de la representatividad (PARODI, 2008). Es posible hablar de ella en términos relativos, alrededor de otros factores que no son la extensión. Según Berber Sardinha (2000), por ejemplo, los corpus compilados en escala pequeña por investigadores individuales terminan siendo más representativos, debido a que los corpus generales "no son necesariamente adecuados para la investigación de cualquier característica lingüística" (BERBER SARDINHA, 2000, p. 348; mi traducción).

Así, resulta claro que cualquier investigación histórica, y en especial aquellas que pretenden dar cuenta de la evolución de estructuras, construcciones y determinados comportamientos morfológicos o sintácticos, requiere de un tipo de aproximación heterogénea. Para ello, se debe priorizar la confección de bases de datos propias, a partir de distintas fuentes, incluso corpus generales de referencia, aunque estos deben ser

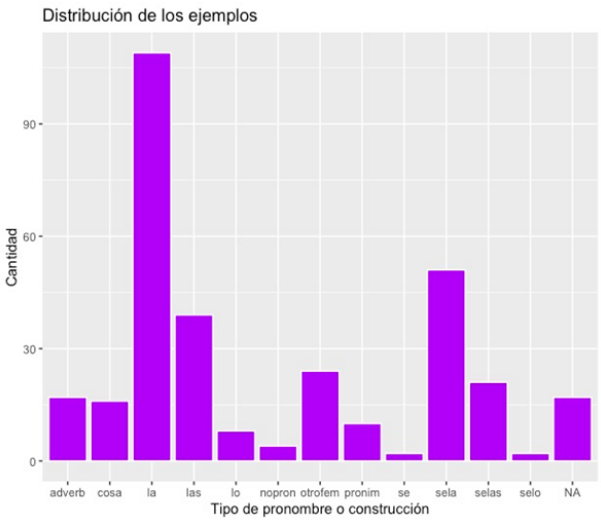
tenidos en cuenta de modo secundario. Las fuentes deben encontrarse y anotarse específicamente según la pertinencia del estudio que se quiere llevar a cabo. Si los estudios están destinados a una palabra o fenómeno en particular, es habitual hacer selecciones de texto que contengan el ítem; esto es particularmente cierto si el ítem es infrecuente (HUNSTON, 2008). Sin embargo, para la emergencia de nuevos fenómenos en el habla coloquial, las prioridades tienen que estar fijadas en la obtención de ejemplos de español hablado, informal y deben ser diatópicamente pertinentes.

Si bien es evidente la contradicción inherente que subyace a la idea de registros orales en investigaciones históricas anteriores al siglo XX, entendemos que esta empresa puede llevarse a cabo a través de registros y géneros textuales particulares que habiliten la presencia de interlocutores o personajes populares y que introduzcan lenguaje hablado (CLARIDGE, 2008). El género literario que mejor se adapta a estas exigencias es el teatro (CONDE, 2010). En general, se trata de obras de carácter realista en las que se procura reflejar la lengua coloquial (FONTANELLA DE WEINBERG, 1970). Sobre todo para la etapa que va desde finales del siglo XIX hasta mediados del siglo XX, el teatro argentino experimenta un auge que no es ajeno a la vasta producción de sainetes, el género teatral popular por excelencia. Esta expansión coincide con la época del aluvión migratorio y la consecuente modificación del entramado social y cultural, la transformación de las estructuras económicas, la urbanización acelerada y el comienzo de un proceso de industrialización en las metrópolis argentinas (BRAVO HERRERA, 2015; FLØGSTAD, 2014; FONTANELLA DE WEINBERG, 1970). En cuanto a lo lingüístico, se señala que es una etapa en la que existe "una tendencia general a una evolución de los tratamientos asimétricos hacia tratamientos simétricos" (FONTANELLA DE WEINBERG, 1970, p. 17); así, sostiene Fontanella de Weinberg, "se ha pasado a un trato igualmente recíproco pero cercano" (1970, p. 22).

Además, este período representa el momento de creación y expansión del lunfardo, argot porteño, cuyo ámbito predilecto de difusión es el lenguaje literario y periodístico y los tangos. El lunfardo no constituye una lengua en sí, sino un tipo de léxico, un sociolecto, un modo de expresión del *populus minutus* porteño (CONDE, 2013, p. 80). Representa un tipo de habla marginal, opuesta a la estandarizada, y por lo tanto con ella "se procura crear una situación comunicativa desjerarquizada" (CONDE, 2013, p. 78).

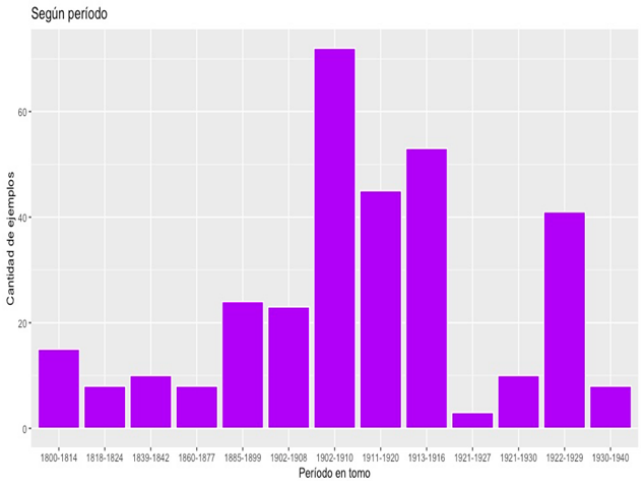
La base de datos histórica confeccionada y analizada en este estudio se compone de 303 oraciones, extraídas de los 15 primeros tomos de la *Antología de obras de teatro argentino*, compilada por la investigadora especialista en teatro Beatriz Seibel, y publicada por el Instituto de Teatro Argentino. La colección abarca un período de 140 años ininterrumpidamente, desde principios del siglo XIX hasta la década de 1940, y presenta obras pertenecientes a distintos subgéneros teatrales, retratando las tendencias estéticas de cada momento. En el 55% de las obras se encontraron construcciones pronominales. La edición se realiza manteniendo los manuscritos originales y está disponible de manera totalmente digital y gratuita. Estos últimos elementos resultan de fundamental importancia para un correcto análisis de texto a través de herramientas computacionales (CLARIDGE, 2008). Eventualmente, el análisis puede complementarse con otra porción de datos tomada de diccionarios y proyectos lexicográficos del lunfardo y letras de tango (CONDE, 2011; GOBELLO, 1991), que suman un total de otros 63 ejemplos. A continuación, se especifican los resultados por período según la especificación de cada tomo (FIGURA 3) y la distribución de los tipos de pronombres asociados a las construcciones (FIGURA 4).

FIGURA 3 – Distribución de ejemplos en base de datos histórica según período.



Fuente: Confección propia a partir de Seibel (2008).

FIGURA 4 – Distribución de ejemplos en base de datos histórica según tipo de construcción.



Fuente: Confección propia a partir de Seibel (2008)

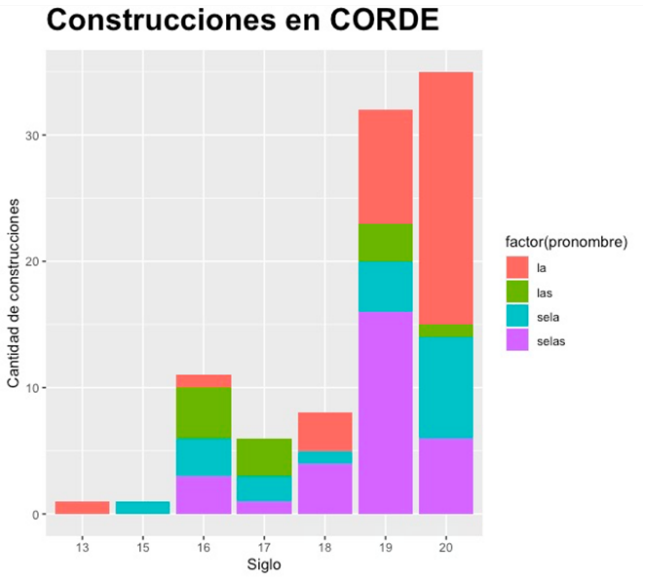
5 Contrastes entre corpus

En relación con el origen de las construcciones, Cifuentes Honrubia (2018) rastrea las primeras oportunidades en que la construcción aparece sin una clara referencia anafórica y, por lo tanto, no puede ser confundida por la combinación de un verbo y un objeto directo pronominalizado convencional. Además de los casos de *hacerla* (‘faltar a lo que se debía’) y *guardársela* (‘aplazar la venganza o castigo de una ofensa’), cuyos primeros ejemplos se ubican en los siglos XIII y XV, a partir de siglo XVI comienzan a detectarse nuevos lexemas y formaciones: 11 para este siglo, 6 para el XVII y otros 8 para el XVIII. A partir del 1800, el número va en aumento y se detectan 32 para el siglo XIX y 35 para el siglo XX (13 de ellos antes de 1950). Solamente 11 de las construcciones relevadas en todo el período analizado pierden frecuencia y no se localizan en textos recientes (FIGURA 6).

En relación con los tipos y la cantidad de pronombres utilizados, si bien el aumento del número de construcciones puede estar relacionado con la disponibilidad propia de las fuentes, y no específicamente con un aumento de la cantidad de construcciones habilitadas en la lengua en el

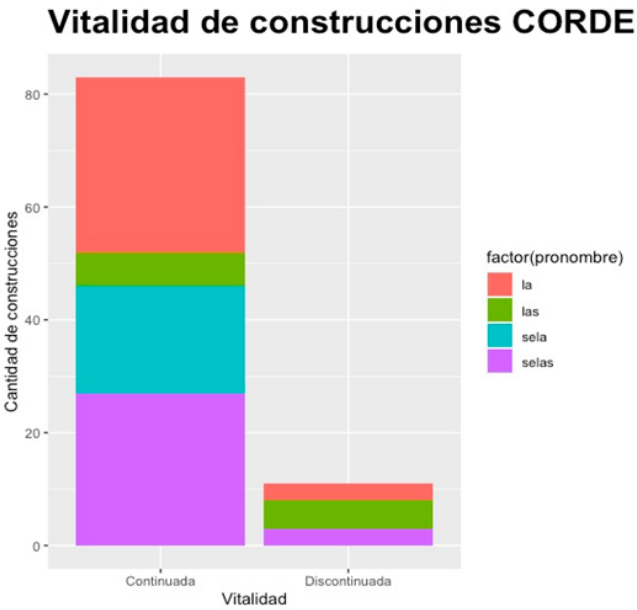
momento analizado, se evidencia una tendencia en el mantenimiento de la disponibilidad de los cuatro tipos de formantes (*se+la*, *se+las*, *las* o *la*), aunque con algunos cambios. En particular, se observa una leve propensión hacia una pérdida del clítico acusativo femenino plural *las* como formante, un aumento en proporción de la cantidad de lexemas construidos con el clítico femenino *la* y la consolidación de la unión de *se+la* como formante más frecuente (FIGURA 5), sobre todo a partir del siglo XIX (los datos del siglo XX corresponden hasta 1940, para poder comparar con la muestra del español argentino en un mismo período).

FIGURA 5 – Distribución de ejemplos según cantidad de construcciones novedosas por siglo y los pronombres que se utilizan como formantes.



Fuente: Confección propia a partir de la base de datos de Cifuentes Honrubia (2018).

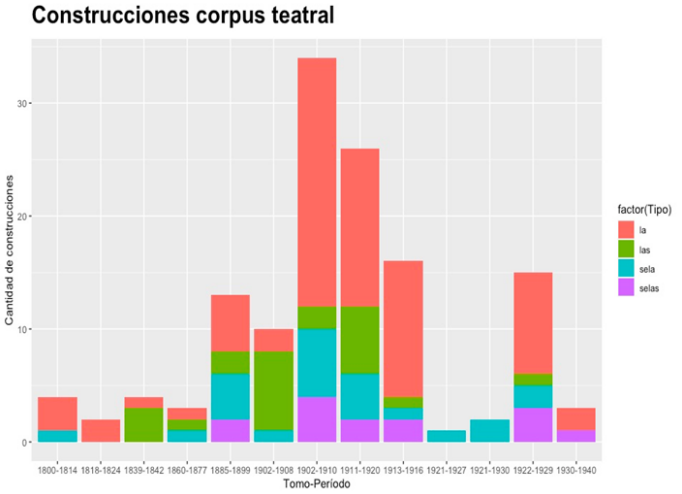
FIGURA 6 – Pérdida de vitalidad de las construcciones según los pronombres que se utilizan como formantes.



Fuente: Confección propia a partir de la base de datos de Cifuentes Honrubia (2018)

La situación difiere cuando se analiza la variedad argentina, a partir de las obras de teatro editadas por la antología de Beatriz Seibel. Mientras que para el siglo XIX se observan pocos ejemplos, de igual manera puede observarse la disponibilidad de múltiples formantes. La prevalencia de las construcciones con clítico pronominal *la*, y en menor medida con *se+la*, comienza a ser evidente desde principios de siglo XX y se extiende, aunque con una menor cantidad de ejemplos, hacia mediados del mismo siglo (FIGURA 7).

FIGURA 7 – Distribución de ejemplos para la base de datos de confección propia según cantidad de construcciones novedosas por período de tomo y los pronombres que se utilizan como formantes.



Fuente: Confección propia a partir de Seibel (2008).

La comparación lexemática también arroja diferencias interesantes. Solo algunas pocas construcciones aparecen en ambas bases y a menudo lo hacen con cambios, particularmente en la elección de pronombres o la necesidad de un predicativo secundario. Para el caso argentino, *haberlas* (‘disputar’), *tirarla* y *echarla*(s) (ambos ‘tener ínfulas de’) aparecen sin el pronombre *se*, al contrario de lo que sucede en el CORDE. En el resto de los casos, como en *dársela* o *dárselas* (‘tener ínfulas de’), *armarla* (‘generar un revuelo’) y *cantarlas* (‘decir lo que se piensa’) se observan exactamente las mismas construcciones, aunque en la base de datos teatral también aparecen algunos predicativos, como *dársela seca* y *dárselas todas* (‘pegar violentamente’) y *cantarlas claro*. Como consecuencia, existe una cantidad relativamente alta de construcciones que no se encuentran en la base opuesta. Si bien en el caso de la base de datos del español europeo no tenemos un acceso directo al total de ejemplos por cada construcción, todas las construcciones relevadas cuentan con una multiplicidad de instancias, que son mostradas por Cifuentes Honrubia (2018) en su argumentación. En este sentido, podemos asumir por lo menos una relativa frecuencia asociada a cada una de estas construcciones. De la misma manera, si nos atenemos a los

casos de la base de datos construida a partir de los ejemplos de obras de teatro argentinas, algunas construcciones con alta frecuencia no aparecen en pocas o ninguna oportunidad en la base de datos ibérica. En particular, se destacan *tenerla con* ('guardar rencor') y las diferentes variantes de *irla de —irlas de/con—* ('creerse algo que no se es'); todos estos lexemas han aparecido por lo menos 5 veces en el conjunto de obras argentinas.

Particularmente puede observarse una importante vitalidad del fenómeno si se tiene en cuenta el análisis de la base de datos de las obras de teatro argentinas en dos aspectos fundamentales. El primero se corresponde con el aumento de las construcciones relevadas en general y en especial con el aumento en proporción de las variantes con exclusivamente pronombre *la* en comparación con los ejemplos del mismo período y con años anteriores. En una segunda instancia, se destaca el tipo de bases verbales con las que el pronombre se combina. Así, se encuentran combinaciones con verbos intransitivos, incluso inacusativos o con un régimen pronominal distinto (*irla*, *palpitarla*, *protestarla*, *equivocarla*, *trabajarla*), ocasionalismos restringidos a la situación comunicativa en especial retratada en la obra o relacionados con el personaje en particular que lo enuncia (*epilogarla*, *chapariarla*), verbos con *-ear* (*gorjearla*) y otras bases verbales novedosas, formadas a partir de sustantivos (*balconearla*) o préstamos léxicos de otras lenguas o relacionadas al ámbito del lunfardo (*escabiarla*, *morfarla*). En todos los casos, la adición del pronombre *la* no contribuye idiomáticamente a la construcción, por lo que los nuevos significados se pueden derivar de los lexemas originales.

(7) Yo nací para **protestarla**. Lo tengo adentro, en la entraña. (*Los disfrazados, apud* SEIBEL, 2008, Tomo 8, p. 232)

(8) Pero conmigo... ¡la chapariolan! (*El conventillo de la paloma, apud* SEIBEL, 2008, Tomo 12, p. 185)

(9) ¿Y cómo **la gorjeás** cuando estás en la batea? ¿O cuando querés que cante el pajarito? (*El debut de la piba, apud* SEIBEL, 2008, Tomo 9, p. 140)

(10) Eso es pa otra gente, que toca el piano y se levanta a las once sin gana 'e **morfarla**… (*Los disfrazados, apud* SEIBEL, 2008, Tomo 8, p. 233)

(11) Vamos a **balconearla**. (*El rey del cabaret*, *apud* SEIBEL, 2008, Tomo 12, p. 107)

Gran parte de estos ejemplos se ven asimismo confirmados por otras bases de datos secundarias de la misma época, que contienen datos provenientes de letras de tango, lexemarios dedicados al lunfardo y otras obras literarias populares no teatrales. Así, los fenómenos en cada lado del Atlántico tienen desarrollos distintos, tanto en sus particularidades históricas como actuales. La explosión de principios del siglo XX en Argentina llega, aunque con intermitencias, hasta principios del siglo XXI con una nueva expansión del fenómeno hacia bases nominales más complejas y préstamos léxicos, como en los ejemplos de (2), que reponemos aquí con nuevas variantes.

(12) Si tengo pago, no **la fantasmeo**. (*Gas Montana Remix*, 2018)

(13) Pienso **buquearla** un rato con el día de la mujer, ahí voy. (Twitter, 2014)

(14) Te guardaste un añito, te teñiste y saliste a **susanearla**. (Twitter, 2014)

Con todo, en el caso de la variedad argentina, y a partir del nuevo corpus, se puede dar cuenta de una serie de características que están ausentes de otros análisis. En primer lugar, se detecta una mayor frecuencia y productividad del fenómeno, especialmente a partir de las primeras décadas del siglo XX. En segundo lugar, este aumento se ve acompañado por una tendencia a la regularización y morfologización del pronombre *la*, al mismo tiempo que se evidencia una pérdida progresiva de la carga idiomática que aporta el clítico. Asimismo, puede postularse una línea temporal de desarrollo del fenómeno, que brinda una posible explicación a la relación del proceso actual con uno histórico. Los ejemplos registrados en Argentina en diccionarios, obras del lunfardo y algunas investigaciones lingüísticas particulares no constituyen casos esporádicos, sino más bien instancias de realización de un fenómeno de amplia difusión en la lengua.

6 Conclusiones

En la introducción identificamos la importancia de las herramientas de la lingüística de corpus en las investigaciones actuales y la relación de esta metodología con otras subdisciplinas dentro de los estudios del lenguaje.

A partir de la indagación sobre la posibilidad de los corpus de referencia, en la sección 2 establecimos sus características más importantes. Nos concentramos en las limitaciones que han sido señaladas por la bibliografía y otras analizadas aquí. Particularmente, recuperamos las observaciones sobre las restricciones para el estudio de la dialectología, la sociolingüística, los estudios gramaticales y la historia de la lengua. En relación con esta última cuestión, se describieron las críticas que se realizaron específicamente sobre los tipos de datos lingüísticos que se recogen en los corpus históricos: falta de registros orales, informales e interactivos y alta dependencia de discursos cultos y literarios. Estas características se encuentran presentes en los dos corpus de mayor alcance del español: CORDE y *Corpus del Español*.

En la sección 3, presentamos las propiedades y antecedentes de estudio de las construcciones con clítico femenino, un proceso de interfaz entre morfología, sintaxis y léxico de amplia expansión en la actualidad en discursos orales e informales. En particular, se indicaron las pocas referencias de estudio histórico que presenta el fenómeno, pese a las referencias hechas a la presencia de estas locuciones en géneros textuales de carácter más popular.

A partir de estos datos y las limitaciones de los corpus generales introducidas en la sección 2, se presentan los lineamientos a favor de la confección de corpus de discursos específicos en la sección 4. Específicamente para el estudio de las construcciones con clítico femenino, se describe el armado de un corpus basado en obras teatrales populares rioplatenses a través de una antología de obras que ocupa la totalidad del siglo XIX y la primera mitad del siglo XX.

Finalmente, en la sección 5, se contrastan los resultados obtenidos del estudio histórico de las construcciones con clítico femenino a partir de la utilización del CORDE para el caso del español ibérico con el análisis de los datos del español de Argentina a partir del corpus histórico teatral. Si bien ambos análisis muestran resultados satisfactorios en tanto encuentran instancias en sus respectivas bases de datos del fenómeno analizado, se evidencian las ventajas de una investigación histórica específica de la variedad rioplatense. El proceso de morfologización, regularización y productividad del clítico femenino *la* en esta zona toma características particulares que no se muestran en el análisis del fenómeno europeo. En efecto, una gran porción de los lexemas, incluso de alta frecuencia, encontrados en el corpus teatral no cuentan con un correlato en CORDE

incluyendo o no las variedades americanas. También el origen y la escisión en el desarrollo del fenómeno en Argentina y en España pueden identificarse con cierta certeza a partir del corpus específico presentado.

Asimismo, una de las mayores ventajas de este tipo de aproximación más heterogénea es la de poder apreciar procesos de surgimiento, afianzamiento, expansión y retracción de determinado fenómeno lingüístico a través de la delimitación y diseño de un corpus específico controlado. De esta manera, la confección de determinadas bases de datos anotadas puede mejorar los estudios de cambio histórico y describir las características propias de fenómenos de gramaticalización y lexicalización del español. Este tipo de registros informales, además, pueden resultar útiles para la investigación de otro tipo de procesos gramaticales, que pueden estar circunscritos solamente al habla coloquial o tener poca incidencia en ámbitos de prensa y literatura. Por último, la confección de corpus específicos puede contribuir a la aportación de otras perspectivas a fenómenos ya investigados a partir de corpus generales, no solo a estudios históricos, sino también sincrónicos.

Referencias

ALBELDA MARCO, M. *La intensificación en el español actual*. 2004. 444f. Tesis (Doctorado en Filologia) – Departamento de Filología Española, Facultat de Filologia, Universitat de Valencia, 2004.

ARELLANO, N. Entre la morfología y la sintaxis: una aproximación a la creación de verbos con pronombre acusativo «la». *Forma y Función*, Bogotá, v. 33, n. 2, p. 81-108, 2020. DOI: https://doi.org/10.15446/fyf.v33n2.80194

ARIAS, J. J. Clítico inherente/marginal *la* en el español rioplatense: ¿de qué la va esta construcción? *Quintú Quimün*, Rio Negro, Argentina, n. 2, p. 74-103, 2018.

BERBER SARDINHA, T. Lingüística de corpus: histórico e problemática. *Delta:* Documentação de Estudos em Lingüística Teórica e Aplicada, São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005

BIBER, D. Methodological Issues Regarding Corpus-Bases Analyses of Linguistic Variation. *Literary and Linguistic Computing*, Oxford, v. 5, n. 4, p. 257-269, 1990. DOI: https://doi.org/10.1093/llc/5.4.257

BRAVO HERRERA, F. E. *Huellas y recorridos de una utopía*: la emigración italiana en la Argentina. 1. ed. Buenos Aires: Teseo, 2015.

BUENAFUENTES DE LA MATA, C.; SÁNCHEZ LANCIS, C. E. Procesos de gramaticalización y lexicalización a la luz de los corpus académicos. *In*: JIMÉNEZ JULIÁ, T. E.; LÓPEZ MEIRAMA, B.; VÁZQUEZ ROZAS, V.; VEIGA RODRÍGUEZ, A. (coord.). *Cum corde et in nova grammatica:* estudios ofrecidos a Guillermo Rojo. Santiago de Compostela: Servicio de Publicaciones e Intercambio Científico/ Universidad de Santiago de Compostela, 2012. p. 153-165.

BYBEE, J. *Morphology.* A Study of the Relation between Meaning and Form. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1985. DOI: https://doi.org/10.1075/tsl.9

CIFUENTES-HONRUBIA, J. L. *Construcciones con clítico femenino lexicalizado.* Madrid: Verbum, 2018.

CLARIDGE, C. Historical corpora. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics*. Berlin; New York: Walter de Gruyter, 2008. v. 1, p. 242-259.

CONDE, O. El lunfardo en la literatura argentina. *Gramma*, Buenos Aires, v. 21, n. 47, p. 224-246, 2010.

CONDE, O. *Diccionario etimológico del lunfardo*. Buenos Aires: Taurus, 2011.

CONDE, O. Lunfardo rioplatense: delimitación, descripción y evolución. *In*: VILA RUBIO, N. (ed.). *De parces y troncos.* Nuevos enfoques sobre los argots hispánicos. Lleida: Edicions de la Universitat de Lleida, 2013. p. 77-105.

CONTRERAS SEITZ, M. Hacia la constitución de un corpus diacrónico del español de Chile. *RLA. Revista de Lingüística Teórica y Aplicada*, Concepción, Chile, v. 47, n. 2, p. 11-134, 2009. DOI: https://doi.org/10.4067/S0718-48832009000200007

DAVIES, M. Un corpus anotado de 100.000.000 palabras del español histórico y moderno. *Procesamiento del Lenguaje Natural*, Jaén, Espanha, n. 29, p. 21-27, 2002.

DAVIES, M. Creating Useful Historical Corpora. A Comparison of CORDE, the Corpus del Español, and the Corpus do Português. *In*: ENRIQUE-ARIAS, A. (coord.). *Diacronía de las lenguas iberorrománicas*: nuevas aportaciones desde la lingüística de corpus. Madrid: Iberoamericana/Vervuert, 2009. p. 137-166. DOI: https://doi.org/10.31819/9783865278685-009

DE MATTEIS, L. Ejes para un debate sobre el uso ético de datos interaccionales escritos y obtenidos en línea. *In*: CANTAMUTTO, L. (ed.). *Actas de las I Jornadas de Humanidades Digitales*. Buenos Aires: Facultad de Filosofía y Letras/Universidad de Buenos Aires, 2015. p. 235-247.

ENRIQUE-ARIAS, A. Dos problemas en el uso de corpus diacrónicos del español: perspectiva y comparabilidad. *Scriptum Digital*, Barcelona, n. 1, p. 85-106, 2012.

FILLMORE, C. J. "Corpus Linguistics" or "Computer-Aided Armchair Linguistics". *In*: DIRECTIONS IN CORPUS LINGUISTICS. NOBEL SYMPOSIUM, 1991, Stockholm. *Proceedings* […]. Berlin; New York: Mouton de Gruyter, 1992. p. 35-60.

FLØGSTAD, G. *Forking Paths:* Category Change, Subfunction Variation, and Preterits in Porteño Spanish and beyond. 2014. 247f. Tesis (Doctorado en Filosofía) – University of Oslo, Oslo, 2014.

FONT, J. *Alma de loca* (letra de tango). Buenos Aires, 1927.

FONTANELLA DE WEINBERG, M. B. La evolución de los pronombres de tratamiento en el español bonaerense. *Thesaurus: Boletín del Instituto Caro y Cuervo*, Bogotá, v. 25, n. 1, p. 12-23, 1970.

GHIO, A.; ALBANO, H. 'Locuciones verbales' con pronombre personal átono 'la'/'las' en el español coloquial de Buenos Aires. *Gramma*, Buenos Aires, v. 24, n. 51, p. 102-116, 2013.

GOBELLO, J. *Nuevo diccionario de lunfardo*. Buenos Aires: Corregidor, 1991.

HUNDT, M. Text corpora. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics*. Berlin; New York: Walter de Gruyter, 2008. v. 1, p. 168-187.

HUNSTON, S. Collection Strategies and Design Decisions. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics*. Berlin, New York: Walter de Gruyter, 2008. v. 1, p. 154-168.

LÜDELING, A.; KYTÖ, M. Introduction. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics*. Berlin, New York: Walter de Gruyter, 2008. v. 1, p. iv-xi. DOI: https://doi.org/10.1515/9783110211429

MARE, M.; CASARES, M. F. *¡A lingüistiquearla!* Neuquén: Educo, 2017.

MASULLO, P.; BÉRTORA, H. Objetos acusativos expletivos en el español rioplatense. *In*: CONGRESO INTERNACIONAL DE LETRAS, VI., 2014, Buenos Aires. *Anales* [...]. Buenos Aires: Universidad de Buenos Aires, 2014. p. 195-205.

PARODI, G. Lingüística de corpus: una introducción al ámbito. *RLA. Revista de Lingüística Teórica y Aplicada*, Concepción, Chile, v. 46, n. 1, p. 93-199, 2008. DOI: https://doi.org/10.4067/S0718-48832008000100006

RISSANEN, M. Corpus linguistics and historical linguistics. *In*: LÜDELING, A.; KYTÖ, M. (Ed.). *Corpus Linguistics*. Berlin, New York: Walter de Gruyter, 2008. v. 1, p. 53-68.

ROJO, G. Lingüística de corpus y lingüística del español. *In*: CONGRESO DE LA ALFAL, XV., 2008, Montevideo. Actas [...]. Montevideo: ALFAL, 2008. Ponencia plenaria. p. 1-31.

ROMAINE, S. Corpus Linguistics and Sociolinguistics. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics*. Berlin, New York: Walter de Gruyter, 2008. v. 1, p. 96-112.

SÁNCHEZ SÁNCHEZ, M. S.; DOMÍNGUEZ CINTAS, C. El banco de datos de la RAE: CREA y CORDE. *Per Abbat: Boletín Filológico de Actualización Académica y Didáctica*, La Rioja, Espanha, n. 2, p. 137-148, 2007.

SEIBEL, B. (ed.). *Antología de obras de teatro argentino*. Desde sus orígenes a la actualidad. Buenos Aires: Instituto Nacional del Teatro, 2008. Tomos 1-15.

SILVA-GARCÉS, J. Clíticos marginales en verbos denominales en -ear. *Quintú Quimün*, Rio Negro, Argentina, n. 1, p. 34-60, 2017.

SILVÉRIA OLIVEIRA, F. Comparação linguística e perfilação gramatical sistêmica em um corpus combinado. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 23, n. 3, p. 727-768, 2015. DOI: http://dx.doi.org/10.17851/2237-2083.23.3.727-768

VELA DELFA, C.; CANTAMUTTO, L. Problemas de recogida y fijación de muestras del discurso digital. *Chimera: Romance Corpora and Linguistic Studies*, Madrid, v. 2, p. 131-155, 2015.

WEISSER, M. *Practical Corpus Linguistics*: An Introduction to Corpus-Based Language Analysis. 1. ed. Oxford: John Wiley & Sons, 2016. DOI: https://doi.org/10.1002/9781119180180

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 739-770, 2021

Um corpus de Estudos de Gênero: por quê, como e para quê?

A Gender Studies corpus: why, how and for what?

Marina Leivas Waquil
Universidade de São Paulo (USP), São Paulo, São Paulo / Brasil
marinawaquil@gmail.com
http://orcid.org/0000-0003-1773-5380

Resumo: Este trabalho apresenta um importante recorte de uma pesquisa que tem como objetivo contribuir com os estudos terminológicos, tradutológicos e sobre corpus ao analisar as unidades que representam e transmitem conhecimento especializado de uma área em crescente evolução acadêmica no Brasil e que discute demandas sociais urgentes, os Estudos de Gênero. Para isso, neste artigo, será exposta a etapa fundamental de qualquer pesquisa com corpus: a definição da área a ser analisada e a compilação de textos com base em critérios confiáveis e que deem conta de representar a área em questão. Assim, o objetivo central deste artigo é mostrar por quê, como e para quê se propôs relacionar a Linguística de Corpus com os Estudos de Gênero a partir de um corpus, apresentando, para tal, um histórico da área selecionada que justifica a análise proposta e sua caracterização como campo especializado. Além disso, destaca-se o referencial teórico que sustenta o trabalho e o corpus de estudo, compilado com base em critérios da Linguística de Corpus e composto pelos dois principais periódicos da área de Estudos de Gênero no Brasil, a *Revista Estudos Feministas* e a *Cadernos Pagu*. Conclui-se defendendo a importância de produzir pesquisas linguísticas e terminológicas que dialoguem com demandas sociais contemporâneas e urgentes.

Palavras-chave: Estudos de Gênero; Linguística de Corpus; Terminologia; *Revista Estudos Feministas*; *Cadernos Pagu*.

Abstract: This work presents an important part of a research that aims to contribute to terminological and translational studies as well as corpus studies, upon analyzing the units that represent and transmit specialized knowledge in a field of soaring academic evolution in Brazil and that discusses urgent social demands, Gender Studies. To do so, this article will expose a fundamental stage of any research regarding corpus: the

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.739-770

definition of the field to be analyzed and the clipping of texts based on reliable criteria that are able to represent such targeted field of study. Accordingly, the main intent of this article is to convey why, how and for what purpose it was proposed to relate Corpus Linguistics with Gender Studies from the compilation of a corpus, introducing therefore a history of the selected field that justifies the analysis proposed and its characterization as a specialized field. In addition, the theoretical references supporting the work and the analyzed corpus stands out, compiled based on the criteria of Corpus Linguistics and composed by the two main journals in the field of Gender Studies in Brazil, the *Revista Estudos Feministas* and *Cadernos Pagu*. In conclusion, it defends the importance of producing linguistic and terminological researches that converse with contemporary and urgent social demands.

Keywords: Gender Studies; Corpus Linguistics; Terminology; *Revista Estudos Feministas*; *Cadernos Pagu*.

Recebido em 08 de setembro de 2020

Aceito em 28 de outubro de 2020

Introdução

A partir do interesse em demonstrar possibilidades de contribuição dos estudos linguísticos e, mais especificamente, da Linguística de Corpus para a discussão de demandas sociais contemporâneas, a autora do presente artigo desenvolve uma pesquisa em andamento na Universidade de São Paulo (USP) em que propõe analisar a terminologia empregada na veiculação de conhecimento do campo de Estudos de Gênero no Brasil, em português, contribuindo para a precisão da comunicação especializada da área e oferecendo um glossário de termos e contextos definitórios como produto-subsídio para tradutores e revisores de textos produzidos com foco nesse campo. Para isso, apoia-se nos princípios teóricos da Terminologia, particularmente da Teoria Comunicativa da Terminologia, que entende os termos como unidades lexicais que adquirem valor especializado em contextos reais de utilização, para identificação, análise e tratamento das unidades terminológicas empregadas na comunicação dos Estudos de Gênero. Conta, também, com a abordagem metodológica oferecida pela Linguística de Corpus, a partir da qual foi compilado um corpus de artigos acadêmicos da área escritos originalmente em português e publicados nos dois principais periódicos de referência em Estudos de Gênero no Brasil, a *Revista Estudos Feministas* (UFSC) e a *Cadernos*

Pagu (Unicamp). Trata-se, portanto, de uma pesquisa interdisciplinar, na medida em que toma como base a linguagem, os princípios teóricos e os critérios aplicados de diferentes campos do conhecimento e busca devolver às disciplinas que os articulam uma contribuição também teórica, a partir da análise e da descrição terminológica, e aplicada, com a proposta de um glossário de termos para tradutores e demais interessados a ser disponibilizado digital e gratuitamente, em sua conclusão, no site do Projeto CoMET.[1]

O artigo aqui elaborado é um recorte atualizado da realização dessa pesquisa e busca 1) justificar a importância da estruturação terminológica de um campo especializado para a sua sistematização no meio acadêmico e, também, na sociedade de modo geral; 2) apresentar um corpus de estudo, compilado com o fim de análise e extração terminológica a partir de parâmetros da Linguística de Corpus e com base na importância dessas fontes para o campo de Estudos de Gênero no Brasil; e 3) defender que a análise terminológica com uma consequente elaboração de produto terminográfico pode contribuir para a ampliação do acesso ao conhecimento de um campo com reivindicações ainda extremamente atuais.

1 "Por quê?" Estudos de Gênero: breve contexto

Em 2019, o Instituto Brasileiro de Geografia e Estatística (IBGE) publicou o estudo Estatísticas de Gênero,[2] em que apresentou dados sistematizados relativos a indicadores de gênero, realizado com base no Conjunto Mínimo de Indicadores de Gênero (CMIG) (Minimum Set of Gender Indicators – MSGI), organizado pela Comissão de Estatística das Nações Unidas (United Nations Statistical Commission) em 2013.

[1] Disponível em: http://comet.fflch.usp.br/. Acesso em: 6 set. 2020

[2] Disponível em: https://www.ibge.gov.br/estatisticas/multidominio/genero/20163-estatisticas-de-genero-indicadores-sociais-das-mulheres-no-brasil.html?=&t=o-que-e. Acesso em: 6 set. 2020. O estudo é baseado em dados elaborados pelo próprio IBGE, como a Pesquisa Nacional por Amostra de Domicílios Contínua (PNAD Contínua), a Pesquisa Nacional por Amostra de Domicílios (PNAD), as Projeções da População por Sexo e Idade, as Estatísticas do Registro Civil, a Pesquisa Nacional de Saúde (PNS) e a Pesquisa de Informações Básicas Estaduais (Estadic), além de informações de fontes externas, como o Ministério da Saúde, o Tribunal Superior Eleitoral (TSE) e o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP),

Uma análise superficial de seus resultados é suficiente para constatar que o Brasil está longe de ter desmontado a desigualdade de gênero nos mais diversos âmbitos da vida em sociedade. No país, por exemplo, as mulheres trabalham uma média de três horas por semana a mais do que os homens e, mesmo contando com um nível educacional mais alto, ganham, em média, 76,5% do rendimento deles.

O IBGE (2019), com a apresentação desse estudo e de seus diversos dados, objetiva fornecer "[...] valiosos elementos para a reflexão de estudiosos e formuladores de políticas públicas". É no contexto dessa reflexão sobre esse e muitos outros dados que se estrutura uma área já consagrada no Brasil, mas em constante evolução: os Estudos de Gênero.

Embora o termo "gênero" tenha sido introduzido pelo psicanalista estadunidense Robert Stoller, em 1963, no contexto de um debate sobre a distinção entre natureza e cultura, sua elaboração, ao longo da segunda metade do século XX, foi feita por pensadoras feministas com o objetivo de questionar e destituir o "[...] procedimento de naturalização mediante o qual as diferenças que se atribuem a homens e mulheres são consideradas inatas, derivadas de distinções naturais, e as desigualdades entre uns e outras são percebidas como resultado dessas diferenças", como afirma Piscitelli (2009, p. 119). Ainda que se tenha informações sobre a organização do movimento feminista já no século XIX, os dados de 2019 do IBGE aqui mencionados demonstram que o objetivo dessas pensadoras feministas mencionadas por Piscitelli (2009) ainda não foi alcançado: seguimos vivendo em uma sociedade que produz desigualdades e violências (físicas, psíquicas e sexuais) em diversas esferas e com base no gênero.

No mundo todo, esses estudos, ainda que se encontrem, atualmente, já muito expandidos e acolham diversos sujeitos, opressões, questionamentos e desigualdades como objeto de estudo, têm uma inegável origem no movimento feminista, que irrompe, no que algumas correntes denominam como uma "primeira onda", no final do século XIX, reivindicando, para as mulheres, o direito ao voto e o acesso à educação, entre outras demandas. A "segunda onda" viria a partir das décadas de 1950 e 1960, quando as pautas do movimento se modificam e avançam, tendo como impulsos a publicação, a tradução e o consequente alcance de diversas obras feministas, que expõem e reivindicam um olhar atento para as questões de gênero. Nesse contexto, surgem novas categorias na reflexão para discussão, e diferentes ferramentas teóricas são produzidas

para explicar as causas originais da subordinação estabelecida, e mantida, pelos homens em relação às mulheres. Assim, conceitos como opressão, patriarcado e relações de poder se estruturam no movimento feminista, e outras noções, como o que é político, são revisadas e redefinidas. Segundo Piscitelli (2009, p. 35), como consequência, passou-se a reexaminar as formas tradicionais pelas quais se explicavam as diversas disciplinas, de modo a encontrar "[...] conceitos apropriados para dar conta da opressão feminina e da realidade das mulheres".

Em decorrência do desenvolvimento exponencial que vai se realizando no contexto da reflexão e da pesquisa sobre questões de gênero, produz-se uma virada fundamental no pensamento do campo: ao expandir o olhar para a diversidade das experiências, foram se ampliando e, ao mesmo tempo, paradoxalmente, especificando as categorias de análise, de modo que muitos conceitos passaram a ser questionados. O conceito de gênero[3] firmar-se-ia a partir disso, servindo como uma "[...] ferramenta alternativa aos conceitos e categorias considerados problemáticos" (PISCITELLI, 2009, p. 136), e tornar-se-ia central no campo.

Na década de 1980, no entanto, inicia-se uma nova contestação dentro do pensamento feminista em relação ao sujeito político "mulher" criado no movimento: principalmente no contexto da reflexão de feministas negras dos EUA, passa-se a formular críticas à ideia de identidade entre as mulheres e ao apagamento das diferenças entre elas e entre suas experiências, fortemente permeadas por questões raciais, sociais, entre outras. O objetivo é evitar generalizações e situar as opressões em contextos particulares, atravessados por mais questões que apenas a de gênero. Além disso, "[...] as novas leituras sobre gênero se esforçam radicalmente para eliminar qualquer naturalização da noção de diferença sexual" (PISCITELLI, 2009, p. 137) e passam a rejeitar classificações lineares e redutoras, entendendo a necessidade de incluir novos sujeitos não contemplados, ou muito pouco considerados, nas reflexões de até então, conformando uma terceira onda.

Nesse contexto de revisão, destaca-se o trabalho de teóricas como, por exemplo, a filósofa Judith Butler (2004), que traz novas e inovadoras categorias para explicar a importância de evoluir de distinções binomiais,

[3] Atribui-se a Gayle Rubin, antropóloga estadunidense, a elaboração e difusão do termo "gênero" a partir de sua proposição de um sistema sexo/gênero e de sua articulação com uma dimensão política.

como mulher/homem, masculino/feminino. Embora mantendo muito da base teórica iniciada no início do século XX, mudam os paradigmas de análise e ampliam-se, consequentemente, os sujeitos considerados objetos de discriminação.

Os Estudos de Gênero, em um constante processo de reelaboração, com suas diversas pautas e objetos de reflexão, sistematizam-se em instituições acadêmicas, geralmente no contexto do amplo guarda-chuva das ciências sociais, e passam a produzir e difundir o conhecimento elaborado com base no conceito de gênero, que passou, e segue passando, por uma profunda revisão: inicialmente, é utilizado exclusivamente em referência às desigualdades produzidas na relação entre homens e mulheres em processos de dominação e subordinação e, então, posteriormente, passa a incluir novas distinções e a articulá-las com diferentes categorias.

Assim, nesse contexto, com o objetivo de produzir e difundir o conhecimento elaborado nesse campo, criam-se programas de pós-graduação, núcleos e grupos de pesquisa e são escritos artigos, dissertações e teses tendo como foco temas suscitados pelas questões de gênero. Essa reflexão ganha ênfase na sociedade por meio da movimentação política ampla e contestatória, recebendo impulso na era digital – com sua facilidade de comunicação e de troca –, o que fortalece e atribui importância ao campo teórico que se dedica às questões de gênero.

Esta breve e resumida contextualização histórica não pretende dar conta da forte e evidente complexidade conceitual da área, mas apenas apontar seu caráter de produção e constante revisitação de conceitos, categorias e ferramentas de análise para discutir fenômenos complexos e permeados por muitas variantes, mostrando que as discordâncias dentro da reflexão contribuem, e são produtivas, para a área especializada, enriquecendo-a conceitualmente.

1.1 Os Estudos de Gênero no Brasil

No Brasil, assim como na maior parte do mundo, o início da sistematização da área de Estudos de Gênero está fortemente vinculado ao movimento feminista. No início dos anos 1970, em um contexto político efervescente e complexo, começa a mobilização social de mulheres e sua reivindicação por melhores condições e igualdade em relação aos homens. Segundo Heilborn e Sorj (1999, p. 3), no entanto, no Brasil,

esse movimento, desde sua origem, contou "[...] com expressivo grupo de acadêmicas, a tal ponto que algumas versões de sua história consideram que o feminismo apareceu primeiro na academia e, só mais tarde, teria se disseminado entre mulheres com outras inserções sociais". Grossi (2004), nesse sentido, destaca a defesa da tese de livre docência de Heleieth Saffioti, em 1967, na USP, como marco do início dos estudos sobre as mulheres no país.

De acordo com Heilborn e Sorj (1999), alguns anos depois, uma importante mudança terminológica e conceitual vislumbrou-se na consagração do campo em contexto brasileiro: "A partir da década de oitenta, observa-se uma gradativa substituição do termo mulher, uma categoria empírica/descritiva, pelo termo gênero, uma categoria analítica, como identificador de uma determinada área de estudos no país" (HEILBORN; SORJ, 1999, p. 4).

Silva, ao destacar a produtiva parceria entre o meio acadêmico e os movimentos sociais, como o feminismo, a partir de convênios, oferecimento de cursos, seminários etc., sublinha que, assim, "[...] a Universidade valida e valoriza as ações promovidas pelas redes, servindo como suporte teórico e, muitas vezes, também, com sua infraestrutura, promovendo uma maior integração entre a sociedade em geral, os movimentos sociais e os cientistas" (SILVA, 2000).

No Brasil, segundo Heilborn e Sorj (1999, p. 3), a institucionalização dos estudos deu-se muito em função do fato de que, desde o início, houve um claro esforço das pensadoras e acadêmicas feministas em integrar-se à "[...] dinâmica da comunidade científica nacional mediante a obtenção do reconhecimento do valor científico de suas preocupações intelectuais pelos profissionais das ciências sociais".

Assim, a área vai se consagrando no espaço acadêmico, formando grupos de trabalho sobre gênero que se fazem presentes em encontros e congressos como os da ANPOCS (Associação Nacional de Pós-Graduação e Pesquisa em Ciências Sociais), desde a sua origem, e introduzindo em programas de pós-graduação disciplinas focadas nas questões de gênero. A produção da área também passa a contar com o suporte e a divulgação de revistas acadêmicas, produzidas no contexto de programas de pós-graduação de universidades destacadas no país. Além disso, são cada vez mais frequentes os diálogos com pesquisadores e pensadores de diversas partes do mundo, que chegam para contribuir com a formação da reflexão da área no Brasil a partir de traduções, mediadas

para transmitir com precisão e adequação a terminologia e os conceitos produzidos e constantemente revisados.

Costa (2010), nesse sentido, destaca a importância do "tráfico" de teorias feministas através de fronteiras geopolíticas que, se nos separam, são também fundamentais para a consagração do pensamento brasileiro sobre gênero. Para isso, dá o exemplo das feministas latino-americanas e latinas que vivem nos Estados Unidos e que "[...] desenvolvem uma política de tradução que se utiliza de conhecimentos produzidos pelos feminismos latinos, de cor, pós-coloniais no norte das Américas para iluminar análises de teorias, práticas, culturas e políticas no sul e vice-versa" (COSTA, 2010, p. 54). Assim, a tradução é, para este campo, elemento essencial e indispensável:

> [...] em termos políticos e teóricos, para a formação de alianças feministas pós-coloniais/pós-ocidentais, já que a América Latina – entendida mais como uma formação cultural trans-fronteira e não como espaço territorialmente delimitado – deve ser vista como translocal. A noção de translocalidade possibilita, por sua vez, a articulação da colonialidade do poder/gênero em várias escalas (locais, nacionais, regionais, globais) com diferentes posições de sujeito (de gênero, sexual, etno-racial, de classe, etc.) constitutivas da identidade (COSTA, 2010, p. 54).

1.2 Estudos de gênero: portanto, área especializada

Se, como afirma Cabré (2001, p. 20), "[s]em terminologia, não se pode fazer ciência", a consagração dos Estudos de Gênero como disciplina reconhecidamente acadêmica no contexto brasileiro não pode prescindir de seu aspecto terminológico e nem ignorá-lo, já que é com base nele que estrutura seu discurso especializado.

É, inclusive, importante destacar a perspectiva terminológica desse campo, já que "quanto mais estruturada é uma disciplina, maior é o nível de precisão semântica, estabilidade formal e sistematicidade de sua terminologia" (CABRÉ, 2001, p. 30), de modo que as pesquisas sobre o léxico empregado nessa comunicação especializada podem colaborar para a sua estruturação. Nesse sentido, discutindo e defendendo a institucionalização e a sistematização dos Estudos de Gênero no Brasil, Heilborn e Sorj (1999, p. 28) destacam que, "[...] para ganhar posição no campo acadêmico, é necessário demonstrar o valor cognitivo da reflexão

empreendida", com o que a Terminologia, por meio da análise de unidades fortemente marcadas por seu aspecto (e valor) cognitivo, pode contribuir.

Segundo Morel e Rodríguez (2001), a comunicação especializada difere da geral em função 1) do aspecto semântico global, já que deve veicular um conhecimento próprio de um campo a partir da produção de sentidos; 2) da importância atribuída ao léxico, que deve ser preciso em sua representação e transmissão; 3) e do aspecto formal do discurso, que se estrutura em uma interação entre especialistas de um campo e de acordo com a perspectiva, ou nível de especialização, desses especialistas.

Nesse sentido, os Estudos de Gênero se caracterizam como comunicação especializada que veicula um tema específico, a partir de uma perspectiva cognitiva e que se realiza entre interlocutores – especialistas – do campo e por meio de terminologia, isto é, unidades que, no discurso, representam e transmitem um conhecimento particular de uma área. Além da extensa produção acadêmica, em formato de artigos, dissertações e teses, da organização de eventos acadêmicos para a apresentação e discussão de trabalhos,[4] há iniciativas incipientes de análise da terminologia do campo.[5]

A afirmação de Minella (2004), em referência à *Revista Estudos Feministas*, periódico que conforma o corpus desta pesquisa e que será apresentado a seguir, destaca, justamente, a complexa questão terminológica e conceitual da linguagem da área de Estudos de Gênero:

> Do ponto de vista da conceituação, à primeira vista tem-se a impressão de que existe um certo caos teórico, pois se atribui aos conceitos básicos da área (gênero, por exemplo) múltiplos significados. Simultaneamente observa-se que o termo feminismo tampouco funciona como uma categoria monolítica, aparecendo ora como política, ora como movimento social, ora como teoria, filosofia, etc. Dependendo do ângulo de análise, o gênero é interpretado ora como desdobramento do feminismo, ora como categoria que inclui o feminismo. O feminismo por sua vez, ora aparece como algo que inclui o gênero, ora como categoria que ultrapassa o gênero. Compreende-se que talvez esta multiplicidade de interpretações deva-se, primeiro, ao próprio contexto de instabilidade da produção científica da pós-modernidade, dado o

[4] Ver: http://www.fazendogenero.eventos.dype.com.br/. Acesso em: 6 set. 2020.

[5] Ver, por exemplo: http://www.ufpb.br/escolasplurais/contents/noticias/didaticos/genero-e-diversidade-sexual-um-glossario. Acesso em: 6 set. 2020.

> rompimento deste contexto com a pretendida identificação entre o real e o racional defendida pela modernidade. Dado ainda o fato de que este contexto admite e até defende, uma certa (des)ordem dos discursos científicos, ou seja, uma certa autonomia dos conceitos e das metodologias em relação às teorias que os engendraram, compreendendo que eles migrem de um lado para o outro, e que entrem desta maneira, numa cadeia intensamente produtora de novas hipóteses e de novas ideias (MINELLA, 2004, p. 231).

Assim, a pesquisa aqui apresentada foi organizada compreendendo como especializado o discurso no campo de Estudos de Gênero produzido por especialistas em português brasileiro e considerando a importância da sistematização terminológica do conhecimento para a estruturação da área e para a sua difusão na sociedade por meio de traduções. O objetivo central estabelecido para tal foi, então, identificar e analisar a terminologia empregada em um corpus de artigos acadêmicos da área de Estudos de Gênero para compilá-la em formato de glossário on-line – composto pelas unidades terminológicas e por contextos definitórios –, contribuindo, dessa forma, para a sistematização da linguagem especializada e como subsídio para o trabalho de tradutores desses textos especializados. A seguir, apresenta-se "como" definimos alcançar o objetivo da presente pesquisa.

2 "Como": um corpus de Estudos de Gênero

Considerando, como Evers e Finatto (2016, p. 272), com base em autores como Halliday, Sinclair e Gries, que "[...] a única forma segura para se descrever uma língua é através da observação dessa língua em uso, analisando-se registros autênticos em larga escala", para oferecer uma contribuição à sistematização da linguagem do campo de Estudos de Gênero, entende-se que é imprescindível contar com os princípios, pressupostos e recursos instrumentais da Linguística de Corpus. Essa abordagem metodológica estabelece e oferece princípios para a compilação de corpora, assim como ferramentas, programas e recursos que podem colaborar nas diversas etapas que constituem um trabalho terminológico e terminográfico, como a compilação de textos para formar um corpus, a identificação de unidades terminológicas e fraseológicas e até mesmo a elaboração de enunciados definitórios para essas unidades.

O ponto de partida da Linguística de Corpus se estrutura em levantamentos probabilísticos, estatísticos e quantificáveis de dados

linguísticos, feitos com base em concepções sobre a(s) língua(s) envolvida (s) e seu funcionamento. O aspecto quantitativo é uma das propriedades do léxico e da língua, ou seja, a frequência de utilização é uma característica essencial das palavras que dá informações importantes sobre a língua como um todo. Ao mesmo tempo, não se pode entender o enfoque estatístico e seus resultados como um fim em si mesmo, mas, sim, como uma referência, como um recurso para análises linguísticas/terminológicas (EVERS; FINATTO, 2016). A partir de um corpus e de sua análise, com a extração de dados quantitativos que permitam a observação de padrões de uso linguístico, é possível produzir descrições confiáveis sobre sistematicidades das línguas.

Com as línguas em contextos especializados de comunicação não é diferente, e a abordagem da Linguística de Corpus é recorrentemente utilizada nas pesquisas terminológicas, que, com o objetivo de identificar especificidades relacionadas às mais diversas terminologias, vêm se aproximando mais do processamento de grandes conjuntos textuais (FINATTO, 2004).

A partir dos anos 1990, os estudos terminológicos se reestruturam e passam por mudanças de paradigmas que têm profundos efeitos em sua prática: entende-se que é fundamental analisar, descrever e compilar os termos *in vivo*, isto é, em contextos reais de comunicação especializada. Nesse sentido, passa a dialogar intimamente com as pesquisas em Linguística de Corpus para observar, com os princípios dessa abordagem, grandes conjuntos de textos e extrair informação linguística/terminológica. A seguir, apresenta-se um breve referencial que guia nossa perspectiva de trabalho na interface Linguística de Corpus e Terminologia.

2.1 Linguística de Corpus e Terminologia

Da necessidade de nomear e caracterizar as coisas, os processos e as atividades, os quais derivam da produção e da aquisição de novo conhecimento, o léxico das línguas se forma e se renova. É em função justamente dessa necessidade que o léxico não é um inventário de palavras fechado, fixo: o conhecimento humano nas mais diversas áreas está em constante movimento, mutação, evolução, revisão e, portanto, a necessidade de incorporar novos signos, ou de revisar os já existentes, é um processo espontâneo, orgânico e, sobretudo, frequente. Porque configura e permite representar a realidade extralinguística é que o léxico é parte fundamental dos estudos linguísticos. O léxico que caracteriza os

Estudos de Gênero não foge a isso: o próprio termo central, "gênero", é um dos mais complexos da reflexão e, ao longo do tempo, esteve, e segue estando, sujeito a constantes revisitações conceituais.

É nesse contexto, tomando o léxico como objeto de estudo central, que se organiza e estrutura a Terminologia, que também se concentra na comunicação humana por meio de signos linguísticos, mas no contexto de áreas especializadas. Embora sua sistematização teórica tenha se realizado, inicialmente, com objetivos prescritivos, normativos e de univocidade comunicacional, centrada exclusivamente no conceito, a partir dos anos 1990, com profundas mudanças de paradigmas, a teoria terminológica expandiu seus campos de análise e consagrou seu caráter interdisciplinar, incorporando conceitos e subsídios das ciências cognitivas, comunicativas e linguísticas.

Desde os anos 1990, então, principalmente devido ao impulso dado pela Teoria Comunicativa da Terminologia (TCT), as pesquisas terminológicas têm, em sua base, a reflexão lexical e buscam analisar, compilar e descrever unidades que, quando utilizadas em contextos especializados de comunicação, adquirem um valor especializado, representando e transmitindo esse conhecimento. Cai, portanto, a barreira que separava rigidamente as línguas de especialidade das línguas naturais e, consequentemente, a que se interpunha entre termo e palavra. Assim, elementos como uso e contexto passam a ter relevância fundamental na análise terminológica.

A TCT surge para questionar a falta de preocupação da reflexão terminológica com questões como estrutura morfossintática, contexto comunicativo, tradução de textos especializados, entre outros. Com uma perspectiva mais dinâmica e flexível, passa a entender e analisar a terminologia como parte de um sistema comunicativo natural, em que o termo, sua unidade central de análise, é uma unidade lexical cujo valor é ativado no discurso e de acordo com o consenso.

Assim, os termos são entendidos como unidades léxicas, multidimensionais, poliédricas, denominativas e conceituais que desempenham uma dupla função: representação e transmissão do conhecimento (CABRÉ, 1999), compreensão que forma a base desta pesquisa. Trata-se de uma associação de forma e conteúdo, formada por um conjunto de traços em um nó cognitivo de uma estrutura conceitual dada sempre em um contexto especializado. Esse conteúdo nunca é absoluto: varia de acordo com o âmbito e a situação de uso.

Outra contribuição da TCT e da qual não podemos prescindir mais atualmente é a de que os termos, como unidades das línguas, estão sujeitos aos mecanismos léxicos de criação e formação de palavras dessas línguas. Consequentemente, estão associados a características gramaticais e pragmáticas, que devem ser consideradas em qualquer análise terminológica.

Além disso, é importante destacar que os termos se conectam entre si por diferentes tipos de relações e, por isso, devem ser observados em fontes reais, como textos especializados. Assim, os textos têm papel fundamental na análise, na reflexão e no trabalho terminológico que passa a ser realizado a partir dos anos 1990, já que são considerados o hábitat natural em que ocorrem os termos e outras unidades de significação transmissoras de conhecimento, como as fraseologias, por exemplo. É por isso, portanto, que a interface com a Linguística de Corpus começou não apenas a ser frequente, mas se tornou a base da pesquisa terminológica, fornecendo subsídios teóricos e metodológicos para a análise e para a descrição dos objetos de estudo da Terminologia.

Atualmente, a proposta de um olhar sobre a terminologia empregada em um campo do conhecimento raramente pode prescindir da observação de corpus, já que é inegável que todo dado terminológico deve advir de uma fonte real, autêntica.

A mencionada mudança no paradigma teórico da Terminologia, a partir dos anos 1990, trouxe diversas consequências também para a prática terminológica/terminográfica, já que os textos, sendo então considerados o berço das unidades terminológicas, passaram a ser ainda mais o objeto de identificação, análise e compilação dessas unidades. Consequentemente, iniciou-se uma íntima aproximação dos estudos terminológicos com a Linguística de Corpus e suas ferramentas de extração de informação linguística.

Neste trabalho, entende-se corpus como "[...] bancos de textos de linguagem autêntica, criteriosamente construídos, destinados a pesquisa e legíveis por computador" (TAGNIN, 2015a, p. 20) e considera-se que a metodologia com base em corpus, isto é, que obtém "[...] todos os seus dados de um corpus especializado compilado para esse fim específico", é essencial para a produção de "[...] glossários que correspondam às necessidades dos tradutores" (TAGNIN, 2015b, p. 375, tradução nossa).

Assim, Terminologia e Linguística de Corpus compartilham seus objetos de estudo, texto e léxico, e, portanto, servem à pesquisa

terminológica em questão, cujos objetos também são os mesmos. Não pretendendo dar conta de toda a linguagem que representa e veicula o conhecimento produzido no campo de Estudos de Gênero, a pesquisa a partir da qual se estrutura este trabalho busca mostrar uma perspectiva dessa linguagem a partir das unidades terminológicas que essa área apresenta em língua portuguesa brasileira, contribuindo, portanto, para a adequação e para a precisão da comunicação da área, oferecendo um esforço de sistematização terminológica que seja útil para seus pesquisadores e que sirva de referência para a tradução de textos do campo na fase de confirmação e validação de equivalentes terminológicos. O objetivo, com isso, é mostrar especificidades lexicais do discurso da área, com foco nas unidades terminológicas que representam e transmitem seu conhecimento.

Para a compilação dessas unidades, toma-se como base os pressupostos terminográficos que estão de acordo com a TCT e que têm como foco o princípio da adequação (CABRÉ, 1999), a partir do qual a metodologia é aplicada a partir da adoção de uma estratégia que varia em cada trabalho de acordo com a temática, os objetivos, o contexto, os recursos à disposição, os elementos envolvidos, entre outros. É, portanto, uma metodologia que se adapta às circunstâncias de cada pesquisa, mas sem contradizer os princípios da teoria (WAQUIL, 2017). Porém, há princípios-chave que conformam a TCT e que devem ser considerados na realização do trabalho terminográfico alinhado a essa teoria, como, por exemplo (CABRÉ, 1999):

a. as unidades terminológicas devem ser consideradas a partir de seu caráter poliédrico, que afeta tanto a denominação quanto o conceito;
b. discordâncias podem ser identificadas na forma como os (diferentes) grupos especializados conceituam a realidade, que pode ser percebida a partir de perspectivas diferentes;
c. a variação também afeta o conceito e a denominação, que, além disso, está sujeita aos fenômenos da língua geral;
d. as unidades terminológicas podem apresentar polissemia: as denominações podem ser total ou parcialmente coincidentes com unidades de outras áreas especializadas;
e. um termo pertence sempre a uma língua e, por isso, responde a todas as regras de formação e funcionamento da mesma;

f. o valor de um termo é determinado por sua presença em um campo específico do conhecimento;
g. o método é necessariamente descritivo e se baseia na coleta de unidades reais utilizadas pelos especialistas de uma área em diferentes situações de comunicação.

2.2 Um corpus de Estudos de Gênero

Para Berber Sardinha (2000), existem quatro pré-requisitos básicos para a compilação de um corpus:

a) Autenticidade I (do conteúdo): os textos que compõem um corpus devem ser autênticos, tendo sido escritos em linguagem natural e sem que tenham sido criados com o propósito de serem objetos de estudo de pesquisa linguística.
b) Autenticidade II (autores nativos): com exceção de alguns casos, como o exemplo dos corpora de aprendizes, os textos selecionados para a compilação de um corpus devem ter sido escritos por falantes nativos da língua na qual estão escritos.
c) Seleção criteriosa: os textos para a formação de um corpus devem ser escolhidos de forma que este tenha alguma característica (por exemplo, um gênero textual específico ou produtores específicos, como aprendizes de língua), a partir de regras especificadas pelo(s) pesquisador(es).
d) Representatividade: critério relativo em função da dificuldade em identificar elementos objetivos para defini-lo, esse pré-requisito deve servir para que o corpus seja representativo de uma língua, variedade, campo especializado etc. A extensão do corpus também deve ser significativa e, embora não se tenha definido números exatos, costuma-se considerar que, quanto maior o número de palavras e de textos, melhor para a pesquisa em corpus.

Além disso, em termos lexicográficos e terminográficos, não se considera mais a possibilidade de compilar e elaborar produtos como dicionários ou glossários sem o suporte da Linguística de Corpus, já que os corpora tanto subsidiam as decisões de um projeto quanto se constituem como a própria matéria a partir da qual são extraídas informações, relações, propriedades e as próprias unidades que comporão

tal projeto. Dessa forma, para o objetivo de produzir um glossário de termos de Estudos de Gênero, também se considera imprescindível o apoio da Linguística de Corpus.

Se, como princípio, na Linguística de Corpus, entendemos que "[...] é preciso ir dos fatos às teorizações" (EVERS; FINATTO, 2016, p. 272), o primeiro passo, após estabelecida a área objeto de estudo desta pesquisa, foi compilar "fatos", isto é, registros reais da comunicação de especialistas de Estudos de Gênero, para, então, em uma próxima fase, teorizar a partir dos achados nesses "fatos", que, aqui, entendemos como contextos, na forma de artigos, em que o conhecimento é desenvolvido, publicado e divulgado, estruturando-se, para isso, em unidades que representam e veiculam esse conhecimento, os termos.

Assim, com base em pesquisa e em referencial teórico (COSTA, 2004; DINIZ; FOLTRAN, 2004; FACCHINI, 2017; MALUF, 2008; PISCITELLI; BELELI; LOPES, 2003) que discute a produção do campo no Brasil, para compilar um corpus segundo os critérios da Linguística de Corpus, buscou-se o subsídio da produção científica da área de Estudos de Gênero no Brasil, em que se destacam duas publicações fundamentais na área e que contribuem para a divulgação do conhecimento produzido no campo e no contexto acadêmico: a *Revista Estudos Feministas* (REF) e a *Cadernos Pagu*.[6]

Considerando que, "[...] para corpora especializados, a coleta de dados contextuais sobre o contexto a partir do qual os textos ou discursos foram coletados pode ser essencial, já que muitas vezes não é possível dar sentido a esse discurso especializado sem algum conhecimento prévio" (KOESTER, 2010, p. 67, tradução nossa) e que "[...] tais informações contextuais são extremamente valiosas: muitas vezes são essenciais para a interpretação dos dados e podem ser utilizadas na análise qualitativa dos resultados do corpus" (KOESTER, 2010, p. 72, tradução nossa), a seguir, apresenta-se o histórico e a descrição desses dois periódicos selecionados para a realização da pesquisa aqui em questão.

2.3 Corpus REF/PAGU

No Brasil, a produção no contexto acadêmico sobre gênero se desenvolveu com intensidade, a partir do aumento do ritmo do surgimento de

[6] Ambas as publicações contam com comitê e conselho editorial e tem Qualis A1 no sistema de avaliação de periódicos da CAPES.

cursos e de núcleos de pesquisa voltados para essa temática, o que, segundo Costa (2004), indicava a possibilidade de um significativo potencial de demanda por espaços de publicação. Nesse contexto, no começo dos anos 1990, surgem, com um ano de diferença, o que viriam a ser os dois mais reconhecidos periódicos da área de Estudos de Gênero, a *Revista Estudos Feministas* (doravante REF) (1992) e a *Cadernos Pagu* (1993).

A REF, reconhecidamente uma das mais importantes publicações do campo de Estudos de Gênero no Brasil e na América Latina, foi criada em 1992, na Universidade Federal do Rio de Janeiro (UFRJ), onde foi editada até 1998, quando foi transferida para a Universidade Federal de Santa Catarina (UFSC). Segundo Diniz e Foltran (2004, p. 245), a REF surge "[...] como parte de uma estratégia deliberada para fortalecer os estudos feministas e de gênero no Brasil" e é idealizada em 1990, em um "[...] seminário histórico que reuniu feministas e acadêmicas em uma cidade do interior do estado de São Paulo", em que se "[...] decidiu pela oferta de cursos itinerantes para fortalecer os estudos de gênero nas universidades". A REF, nesse contexto, é proposta como um instrumento de difusão das pesquisas produzidas na área e como veículo de formação e educação política. Minella (2004, p. 224), nesse sentido, destaca que "[...] o surgimento da Revista resulta de um longo processo de maturação epistemológica e política, intensificado nos anos oitenta no Brasil em torno da questão feminina, como reflexo do debate internacional e da expansão das lutas e movimentos feministas".

Um grande e importante impulso à publicação foi o apoio financeiro da Fundação Ford, que permitiu, entre outros, a internacionalização da REF: "[...] por um lado, traduziram-se artigos chave para o debate feminista e de gênero internacional, e, por outro, artigos nacionais foram também traduzidos para a língua inglesa, tendo havido inclusive um número especial neste idioma como forma de promover e divulgar a REF" (DINIZ; FOLTRAN, 2004). A seção de artigos avulsos da REF, considerada "a mais criteriosa e de maior notoriedade da revista" (DINIZ; FOLTRAN, 2004), inclui os textos traduzidos de idiomas como inglês e francês, enquanto os em espanhol são mantidos nessa língua.

A descontinuidade do apoio financeiro da Fundação Ford, com a consequente indisponibilidade da UFRJ, que dependia desses recursos para seguir mantendo a revista, acabaria levando a REF para a UFSC, mas sem interferência na qualidade da produção e da publicação, já que sempre foi um preceito da revista ser uma publicação nacional, não se

restringindo a um núcleo de pesquisadoras ou pesquisadoras de uma única universidade (GROSSI, 2004). Assim:

> É graças a este lugar que a REF já tinha ao chegar em Florianópolis que mantivemos os critérios de excelência como a periodicidade, o rigor d@s pareceristas, a rapidez na avaliação dos artigos submetidos, a busca de traduções de textos fundamentais para a formação d@s estudantes de graduação, o empenho em entrevistar expoentes da teoria feminista internacional (GROSSI, 2004, p. 215).

Segundo Grossi (2004), a REF, além disso, sempre se preocupou com a questão formal, aliada ao conteúdo, com o objetivo de destacar-se e tornar-se uma revista de ponta na área, o que é também evidente na diagramação diferenciada da publicação. Além disso, baseia-se em uma "política de democratização do acesso à produção científica e acadêmica" (MALUF, 2008, p. 123), com o que, por exemplo, é editada eletronicamente na "SciELO Social Science, um portal ligado à SciELO, mas que, buscando implementar uma política de tradução e visibilidade internacional da produção científica e acadêmica brasileira, disponibiliza os artigos em inglês" (MALUF, 2008, p. 125).

A *Cadernos Pagu*, por sua vez, surge na Unicamp, segundo Piscitelli, Beleli e Lopes (2003, p. 243), em um contexto em que os Estudos de Gênero já tinham se estabelecido no Brasil e já contavam com certa legitimidade acadêmica, mas sua intenção era "[...] ampliar o espaço já existente, difundindo e estimulando a produção de conhecimento na área", incluindo, nisso, a diversificação das temáticas apresentadas. Segundo as autoras, a criação da *Cadernos Pagu* se estabeleceu após "[...] mais de dois anos de leituras, pesquisas e debates, nos quais integrantes do Núcleo de Estudos de Gênero – Pagu mapeavam os avanços na produção sobre gênero e seus impasses". Com o tempo, o periódico foi se estruturando e contando com a colaboração de pesquisadores estrangeiros. No quinto número, a publicação também começou a receber financiamento externo, o que contribuiu para o seu crescimento tanto em termos de qualidade gráfica e quantidade de textos em cada edição, mas também no que se refere a questões como conselho e normas editoriais e ao registro em diversos indexadores nacionais e internacionais (PISCITELLI; BELELI; LOPES, 2003).

Assim, foi-se organizando com a preocupação de incentivar, em contexto brasileiro, a reflexão e a produção no campo de gênero, mas sempre

a partir da perspectiva de diversidade temática, em que se entende que é possível o diálogo com essa questão nos mais variados recortes temáticos. Nesse sentido, assim como a REF, a *Cadernos Pagu* sempre buscou diálogo com a comunidade internacional por meio da publicação de traduções de textos sobre gênero que são referência no mundo, com o objetivo de, com isso, "[...] promover a leitura crítica da produção internacional" (PISCITELLI; BELELI; LOPES, 2003, p. 243), sempre acompanhada de produção nacional (que constitui 85% da publicação), e colaborar para o desenvolvimento de mais e novos conhecimentos no campo.

Apesar do enfoque e do seu contexto acadêmico, segundo as autoras, a publicação *Cadernos Pagu* também atinge e contribui com outros leitores, situados em organizações governamentais, não governamentais e em movimentos sociais variados.

Regina Facchini (2017) destaca que a *Cadernos Pagu* se renova constantemente, acompanhando o desenvolvimento do campo no país e no mundo e desempenhando fundamental papel. Segundo a autora, a publicação é um importante legado e patrimônio da comunidade científica engajada nos estudos feministas e de gênero no país, insistindo na produção com diversidade regional e disciplinar, o que também caracteriza o corpo de revisores, editores e tradutores que colaboram com a publicação.

Com periodicidade quadrimestral desde 2016, a *Cadernos Pagu* reforça, como mencionado, desde seu surgimento, a importância do diálogo com a comunidade estrangeira com uma política de tradução e publicação de textos-chave do campo, de modo a possibilitar "[...] não só a difusão de conhecimento, mas uma leitura crítica da produção internacional, contribuindo para a formação de jovens pesquisadores" (FACCHINI, 2017).

Desde 2014, a *Cadernos Pagu* vem sendo publicada exclusivamente em formato digital, on-line, seguindo uma tendência cada vez mais comum e que faz "[...] com que os artigos se consolidem como unidades na rápida circulação do conhecimento por meio da internet" (FACCHINI, 2017), alcançando novos e diferentes leitores.

Essas duas publicações, como vemos, não apenas foram firmadas por um contexto propício para seu surgimento e sucesso, mas, também, da mesma forma, ajudaram enormemente a firmar o campo de Estudos de Gênero no Brasil. Referindo-se à REF, Minella (2004, p. 224), por exemplo, destaca que a publicação atuou mantendo "[...] uma

relação peculiar com o campo dos estudos feministas e das relações de gênero, atuando no sentido de modificá-lo e sendo simultaneamente metamorfoseada por ele".

QUADRO 1 – Comparativo REF e *Cadernos Pagu*

	REF	*Cadernos Pagu*
Ano de criação	1992	1993
Projeto gráfico	Inovador – impulso do apoio financeiro da Fundação Ford	Ascetismo formal
Filiação	Pretende ser um periódico não diretamente institucional (institui-se na UFRJ e se transfere para a UFSC)	Revista de um núcleo universitário (Unicamp)
Objetivo/ descrição	"A *Revista Estudos Feministas* (REF) tem como foco as questões de gênero e feminismos, que podem ser tanto relativos a uma determinada disciplina quanto interdisciplinares em sua metodologia, teorização e bibliografia. A cobertura temática contribui para o estudo das questões de gênero, sendo provenientes de diversas disciplinas: sociologia, antropologia, história, literatura, estudos culturais, ciência política, medicina, psicologia, teoria feminista, semiótica, demografia, comunicação, psicanálise, entre outras". (REVISTA ESTUDOS FEMINISTAS, s.d.).	"*Cadernos Pagu*, publicação quadrimestral interdisciplinar, tem como objetivo contribuir para a ampliação e o fortalecimento do campo interdisciplinar de estudos de gênero, dando visibilidade à produção realizada no Brasil e promovendo o intercâmbio de conhecimento internacional sobre a problemática. Publica artigos inéditos com contribuições científicas originais, que colaborem para a inovação teórica, metodológica e/ou agreguem conhecimento empírico inovador, e debates em torno de textos teóricos relevantes no campo dos estudos de gênero, viabilizando, assim, a difusão de conhecimentos na área e a leitura crítica da produção internacional" (CADERNOS PAGU, s.d.).
Número de artigos	Total de textos: 1050 Textos escritos originalmente em português: 829 Artigos em língua estrangeira ou tradução: 231 Resenhas: 286	Total de textos: 641 Textos escritos originalmente em português: 541 Artigos em língua estrangeira ou tradução: 140 Resenhas: 138
Ambas contam com comitê e conselho editorial e tem Qualis A1 no sistema de avaliação de periódicos da CAPES.		
Ambas estão disponíveis na Scientific Electronic Library Online (SciELO)		

Fonte: Elaborado pela autora com base em Costa (2004).

A compilação desse corpus (nomeado REF/PAGU) se deu com base nos mencionados critérios apresentados por Berber Sardinha (2000) e sua análise inicial foi realizada com os princípios discutidos em Tagnin (2013), com o desenvolvimento dos seguintes passos:

1. Os textos[7] foram extraídos de suas páginas web, salvos em pastas correspondentes e nomeados para sua identificação posterior em outras etapas da pesquisa (PAGU_1; REF_1; etc.).
2. Em formato .pdf, os textos foram convertidos por um leitor de OCR (optical character recognition), a ferramenta on-line Convertio, para o formato .txt. Nesta etapa, em função da impossibilidade de leitura de OCR de material com formatação com baixa resolução (caso dos artigos mais antigos), foi necessário excluir textos inicialmente previstos como componentes do corpus. Chegou-se, assim, a um total de 688 artigos extraídos, convertidos e limpos – 380 da REF e 288 da *Cadernos Pagu*. Segundo a tipologia de Berber Sardinha (2000), podemos caracterizar o cropus compilado da seguinte forma:
   - modo: escrito. Todos os textos que compõem o corpus são escritos;
   - tempo: diacrônico. Os textos selecionados foram produzidos nas últimas três décadas;
   - seleção: de amostragem e estático. O corpus foi compilado e planejado para ser entendido como uma amostra finita da linguagem do campo de Estudos de Gênero. É, também na tipologia de Berber Sardinha, estático, em oposição a dinâmico;
   - conteúdo: especializado. Todos os textos se inserem no contexto de publicações do campo de Estudos de Gênero, estando compostos pela linguagem que o caracteriza. É também monolíngue, já que foram todos escritos em língua portuguesa do Brasil;
   - autoria: de língua nativa. Os textos foram escritos por falantes nativos do português brasileiro;
   - finalidade: de estudo. O corpus em si é o objeto de estudo, análise e extração – terminológica.

[7] Foram compilados apenas os textos das seções de artigos científicos para que, na próxima etapa da pesquisa, seja possível analisar a terminologia em empregada nesse gênero textual especificamente.

3. Após a conversão para .txt, os textos passaram por um processo de limpeza manual, em que foram apagados trechos que emitiriam ruído na análise e que não eram de interesse para o alcance dos objetivos propostos.
4. Com o corpus organizado, foi feita sua inserção na ferramenta de análise e extração linguística AntConc.[8]

Embora a análise terminológica ainda esteja em fase de realização, alguns importantes e significativos dados já foram observados. De acordo com o Antconc, há 2.848.938 tokens e 105.233 types. Entendendo que se trata de um número alto e suficiente para uma profunda análise terminológica, considera-se que, "[...] mais importante que tamanho do corpus em si é o quão bem ele foi planejado e se é representativo" (KOESTER, 2010, p. 67), ou seja, o corpus deve ser projetado de forma que seja adequado para a proposta da pesquisa. Nesse sentido, se a compilação de um corpus depende do que se quer investigar, acredita-se que o corpus PAGU/REF,[9] pelo histórico e pela descrição aqui apresentados é um corpus representativo da área de Estudos de Gênero, que, no Brasil, como vimos, estruturou-se, em grande medida, justamente em função dessas duas publicações.

---

[8] Programa gratuito de ferramentas de análise de corpus para análise de texto criado pelo Prof. Lawrence Anthony, da Universidade Waseda, no Japão.

[9] Cabe ressaltar que, no Brasil, outras publicações especializadas na área de Estudos de Gênero vêm se estruturando em distintos centros universitários, produzindo e divulgando conhecimento, como a REF e a Cadernos Pagu. Nesse sentido, destacam-se periódicos como "Caderno Espaço Feminino" (Qualis B2), publicação do Núcleo de Estudos de Gênero e Pesquisa sobre a Mulher do Centro de Documentação e Pesquisa em História (CDHIS), da Universidade Federal de Uberlândia; "Gênero" (Qualis B3), periódico de circulação nacional, iniciativa do Núcleo Transdisciplinar de Estudos de Gênero e que, atualmente, está vinculado ao Programa de Estudos Pós-Graduados em Política Social da Universidade Federal Fluminense; e a Revista Ártemis (Qualis B2), periódico semestral, interdisciplinar, vinculada aos Programas de Pós Graduação em Sociologia e Letras da Universidade Federal da Paraíba. Considerando como critérios Qualis segundo o sistema da CAPES, tempo de publicação e reconhecimento na área de Estudos de Gênero, para a presente pesquisa, optou-se pela extração apenas de textos da REF e da Pagu, embora não se descarte, como perspectiva futura, a ampliação do corpus considerando os periódicos aqui mencionados.

Na extração terminológica inicial feita como teste do corpus, dados bastante interessantes confirmam uma expectativa inicial da pesquisadora a respeito dos termos cuja frequências se destacaria. Nas primeiras posições, destacam-se types como mulheres, homens, gênero, corpo, sexual, poder, feminino, termos expressivos da comunicação do campo de Estudos de Gênero e cujo caráter especializado será confirmado com análise contextual.

QUADRO 2 – Corpus REF/PAGU: 25 primeiras posições na Wordlist do Antconc

Posição	Frequência	Palavra
1	25749	mulheres
2	12973	mulher
3	8823	trabalho
4	8522	homens
5	7762	ela
6	7575	social
7	7387	forma
8	7075	gênero
9	6725	vida
10	6377	anos
11	6028	sociais
12	5776	corpo
13	5394	ele
14	5162	sexual
15	4840	parte
16	4824	outro
17	4728	brasil
18	4422	poder
19	4309	feminino
20	4294	tempo
21	4288	homem
22	4279	sexo
23	4156	pessoas
24	3965	elas
25	3868	sociedade

Fonte: Elaborado pela autora.

O termo “mulher”, que ocupa a primeira e a segunda posição da lista de palavras mais frequentes do corpus REF/PAGU aponta uma questão prevista e que deve ser observada com cuidado nas pesquisas

que observam a comunicação especializada da área de Estudos de Gênero no Brasil: como indicam as pesquisas historiográficas sobre o seu desenvolvimento no país, esse campo teórico teve sua origem fortemente vinculada ao movimento feminista e, portanto, às questões das "mulheres": nos anos 1980, "[...] a temática da **mulher** foi incluída em prestigiosos congressos e encontros de ciências sociais e aumentou o número de pesquisas, dissertações e teses com orientação feminista" (PISCITELLI, 2013, p. 381, grifo nosso). Costa (2004, p. 206, grifo nosso), também revisando as origens dos estudos no Brasil, destaca que:

> A vitalidade da produção acadêmica sobre **mulher** era invejável; como atestavam inúmeros seminários, grupos de trabalho nas principais associações de pós-graduação (ANPOCS, Associação Brasileira de Antropologia, Associação Brasileira de Estudos Populacionais, Associação Nacional de Pós-Graduação e Pesquisa em Letras e Linguística, Associação Nacional de História) e o crescimento progressivo de núcleos de pesquisa nas universidades.

A categoria "mulher", no entanto, foi constantemente discutida, criticada, questionada e reelaborada e, por isso, Grossi (2004, p. 218) já apontava, 16 anos atrás, a variedade temática no campo, que deixou de centrar-se apenas nessa categoria, expandindo-se e produzindo, além de "pesquisas sobre mulheres, pesquisas sobre homens, pesquisas que analisam as relações de gênero, pesquisas preocupadas com questões teóricas, pesquisas sobre o movimento feminista e de mulheres, etc.". Apesar disso, "mulher" e "mulheres", de acordo com essa primeira extração-teste, são as duas formas mais frequentes em um corpus de mais duas milhões de palavras, o que parece indicar uma ainda muito forte ligação das pesquisas da área com as questões que atravessam essa categoria em toda a sua, atualmente observada, diversidade.

Nesta análise preliminar, também se destaca o fato de que os itens "homem" e "homens" apareçam entre os mais frequentes do corpus, o que pode estar relacionado ao fato de que "mulher"/"mulheres" também o estejam: os Estudos de Gênero, como vimos, especialmente no Brasil, estão intimamente vinculados com o questionamento do movimento feminista em referência à opressão sistematicamente perpetrada por "homens", em um sistema patriarcal, contra as "mulheres".

Nesse sentido, também é importante apontar que "Brasil" é um termo presente nessa lista, o que parece indicar um forte enfoque da

produção dos artigos nas questões de gênero em contexto nacional, em seus sujeitos e em suas experiências. Nos últimos anos, por exemplo, vem tomando força o desenvolvimento dos feminismos transnacionais, cuja base reside em situar-se, marcando sempre de onde se fala, o que se analisa, para que não sejam produzidas generalizações que pretensamente deem conta da totalidade das experiências das mulheres. Há, nesse sentido, a possibilidade de que a perspectiva transnacional esteja produzindo efeitos nos discursos compilados no corpus e, portanto, na escolha terminológica de seus autores.

Dessa forma, embora o foco do presente artigo não seja a aprofundada extração/análise dos termos do corpus REF/PAGU, entende-se que esses resultados preliminares destacados no Quadro 2, e que ainda serão tratados também com o auxílio de outras ferramentas do AntConc, como Keywords, Clusters e Collocates, indicam importantes e interessantes perspectivas para a pesquisa.

3 Para quê?

O ano de 2020, provavelmente, ficará marcado na história da humanidade por uma série de motivos diferentes. A pandemia do novo coronavírus, por exemplo, vem produzindo efeitos catastróficos nas vidas e em diversos setores das sociedades ao redor do mundo inteiro. Embora, evidentemente, chame a atenção o número de vidas perdidas e a crise econômica que vem se alastrando aos poucos, outro problema, também de escala global e que é preciso destacar, refere-se ao aumento, nos mais diversos países, de violências sendo produzidas em função de questões de gênero e no contexto do isolamento como consequência da pandemia.

Exemplos disso são facilmente encontrados: de acordo com dados do Ministério da Mulher, da Família e dos Direitos Humanos (MMDH), com apenas um mês de isolamento no Brasil, em abril de 2020, as denúncias de violência contra a mulher feitas pelo canal 180 já tinham subido 40% em relação ao ano anterior (ESTADÃO CONTEÚDO, 2020). Não à toa, temos visto, como resposta a esse aumento, surgirem diversas campanhas, produzidas tanto na esfera pública quanto na privada, para orientar a população e divulgar canais de denúncia para essas violências e de combate a esses abusos, físicos e emocionais.

Nesse sentido, a Organização das Nações Unidas (ONU) também tem se mobilizado, estruturando parcerias para enfrentar as consequências

do que, segundo Phumzile Mlambo-Ngcuka, diretora executiva da ONU Mulheres e vice-secretária geral das Nações Unidas, é uma "pandemia invisível": a violência contra mulheres e meninas, que, vem emergindo como efeito/consequência do coronavírus e é "um espelho e um desafio aos nossos valores, nossa resiliência e humanidade compartilhada" (MLAMBO-NGCUKA, 2020).

Ao mesmo tempo, outros sujeitos têm se visto afetados ainda mais no contexto pandêmico: segundo notícia da Agência Brasil (2020), "[...] no primeiro semestre deste ano, 89 pessoas transgênero foram assassinadas no Brasil, quantidade que supera em 39% a registrada no mesmo período de 2019"; segundo a Associação Nacional de Travestis e Transexuais (Antra), esses números apontam o quão vulneráveis estão esses sujeitos, mas:

> [...] não refletem exatamente a realidade da violência transfóbica em nosso país, uma vez que [sua] metodologia de trabalho possui limitações de capturar apenas aquilo que de alguma maneira se torna visível. É provável que os números reais sejam bem superiores. Mesmo com essas limitações, os dados já demonstram que o Brasil vem passando por um processo de recrudescimento em relação à forma com que trata travestis, mulheres transexuais, homens trans, pessoas transmasculines e demais pessoas trans (BOND, 2020, documento).

Os Estudos de Gênero, portanto, veem-se diante, continuamente, de novas questões, novas demandas e diferentes sujeitos. A reflexão do campo é, assim, cada vez mais fundamental e desafiada a constantemente renovar-se e atualizar-se. Nesse sentido, são inevitáveis, também, impactos nas formas linguísticas/terminológicas por meio das quais essas reflexões se estruturam.

Mas de que forma os estudos linguísticos/terminológicos podem contribuir efetivamente para o debate dessas demandas? Recentemente, junto ao contexto da pandemia, vimos tomar força, nos Estados Unidos, o movimento Black Lives Matter (em português, literalmente, "vidas negras importam"), impulsionado por casos de violência despropositada perpetrados contra sujeitos negros. Além de inúmeras manifestações, protestos, discussões e ações políticas, o movimento produziu efeitos em outras esferas, e a da língua é uma delas. Em reportagem publicada em junho deste ano pelo jornal *The New York Times*, Hauser (2020)

apresentou uma importante mudança em andamento em um dos mais conhecidos e renomados dicionários da língua inglesa, o Merriam-Webster, incentivada pela reivindicação da uma usuária, Kennedy Mitchum, mulher de 22 anos recém-formada na Universidade Drake, no estado do Missouri. Mitchum escreveu aos editores da obra lexicográfica contestando as três acepções da entrada "racismo", que, em sua opinião, contemplam apenas a ideia de que se trata de preconceito contra uma certa raça devido à cor da pele; sua demanda é por uma definição do item lexical que inclua a concepção de que o racismo é uma prática que combina esse preconceito com poder social e institucional, um sistema de vantagens baseado em cor de pele; portanto, que reflita o racismo sistêmico e estrutural que permeia a sociedade.

Os editores do dicionário admitiram que a entrada "racismo" não era revisada há décadas e que as revisões são produzidas, justamente, quando veem "[...] mudanças em larga escala ocorrendo na linguagem" (HAUSER, 2020, s.p., tradução nossa). Na troca de correspondência entre Mitchum e os editores do Merriam-Webster, a mulher questionou, inclusive, as fontes utilizadas por eles em termos de representatividade – elemento fundamental de qualquer corpus segundo os princípios da Linguística de Corpus, independentemente do objetivo da pesquisa.

Embora não se discuta esse tema com a devida frequência, os estudos linguísticos têm, efetivamente, impactos diretos no mundo, e vice-versa. Mitchum, em um de seus argumentos pela reivindicação da revisão da entrada, destaca ter observado muitas pessoas brancas defenderem seu ponto de vista "copiando e colando" as definições apresentadas nos dicionários, o que mostra a importância das reflexões e produções lexicográficas na sociedade. Os próprios editores admitem que palavras com conceitos mais abstratos, como racismo, são constantemente consultadas pelas pessoas em dicionários.

Ainda que esse seja o caso de um impacto de uma demanda social em um contexto não especializado, de língua geral, e de uma obra lexicográfica – o Merriam-Webster –, entendendo que as unidades terminológicas passam pelas mesmas regras de funcionamento que as unidades lexicais que compõem obras lexicográficas, a correlação desse caso com esta pesquisa é justificada e coerente: considera-se que áreas que se baseiam em demandas sociais, como os Estudos de Gênero, sejam objeto de estudo linguístico/terminológico, que podem contribuir para a sistematização de unidades que, como "racismo", têm impacto importante

nas reflexões sobre questões da área, seja em contextos especializados ou não. Assim, seguindo os princípios da TCT, não é possível ignorar que os termos, assim como as unidades lexicais, estão inerentemente permeados por características gramaticais e pragmáticas, que devem ser consideradas, observadas e descritas em qualquer análise terminológica.

Nesse sentido, entendem-se como essenciais as pesquisas que se debrucem sobre a comunicação de teóricos, pensadores, pesquisadores e especialistas da área para a observação da representação e da transmissão de um conhecimento que constantemente se atualiza em função das demandas sociais. Esta pesquisa, portanto, tem como objetivo colaborar para este esforço: chamar a atenção de campos como o da Linguística de Corpus – com sua abordagem funcionalista da linguagem, fornecendo subsídios teóricos e aplicados para análise – e o da Terminologia – em uma perspectiva comunicativa, que não entende os termos como unidades fixas e preestabelecidas, mas como elementos das línguas que, de acordo com o contexto em que são utilizadas, adquirem caráter especializado – para a área de Estudos de Gênero, apresentando-a a partir da observação da produção e transmissão do conhecimento que produz em forma de corpus.

Para isso, neste artigo, foi apresentado um recorte que constitui uma etapa fundamental: a compilação, a descrição e a justificativa de um corpus representativo da comunicação da área e que cumpre um papel fundamental e simbiótico com a mesma no Brasil, as publicações REF e *Cadernos Pagu*. Com isso, mostrou-se um corpus que tem origem em fontes bem estruturadas e confiáveis para extração terminológica e que, além de servir à presente pesquisa, pode configurar-se como objeto de estudo para futuras investigações, essenciais para a estruturação, para a revisão e para a atualização da língua portuguesa e de sua utilização no campo especializado de Estudos de Gênero. Embora esta pesquisa ainda esteja em vias de realização e espere-se que muitos outros resultados importantes terminológicos sejam alcançados, entende-se que a fase apresentada aqui, de compilação de corpus, é a base para quaisquer futuros achados.

O "[...] ativismo não muda o dicionário, [...] o ativismo muda a língua" (HAUSER, 2020, tradução nossa); portanto, é necessário, para as mais diferentes sociedades, que os estudos que observam as línguas, em contextos gerais ou especializados, estejam atentos aos efeitos que movimentos sociais, como os vinculados às questões de gênero e suas interfaces, vêm produzindo constantemente.

Referências

BERBER SARDINHA, T. Lingüística de corpus: histórico e problemática. *DELTA*, São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005

BOND, L. Pesquisa mostra aumento da violência contra pessoas trans no Brasil. *Agência Brasil*, 28 jun. 2020. Disponível em: https://agenciabrasil.ebc.com.br/direitos-humanos/noticia/2020-06/pesquisa-mostra-aumento-da-violencia-contra-pessoas-trans-no-brasil. Acesso em: 5 set. 2020.

BUTLER, J. *Undoing Gender*. New York: Routledge, 2004. DOI: https://doi.org/10.4324/9780203499627

CABRÉ, M. T. *La terminología*. Representación y comunicación. Una teoría de base comunicativa y otros artículos. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra, 1999.

CABRÉ, M. T. Sumario de principios que configuran la nueva propuesta teórica y consecuencias metodológicas. *In*: CABRÉ, M. T.; FELIU, J. (org.). *La terminología científicotécnica*: reconocimiento, análisis y extracción de información formal y semântica: Informe DGES PB-96-0293. Barcelona: Universitat Pompeu Fabra; Institut Universitari de Lingüística Aplicada, 2001. p. 17-25.

CADERNOS PAGU. São Paulo, [s.d.]. Disponível em: https://www.pagu.unicamp.br/pt-br/cadernos-pagu. Acesso em: 5 set. 2020.

COSTA, A. de O. *Revista Estudos Feministas*: primeira fase, locação Rio de Janeiro. *Revista Estudos Feministas*, Florianópolis, v. 12 n. especial, p. 211-221, 2004. DOI: https://doi.org/10.1590/S0104-026X2004000300022

COSTA, C. de L. Feminismo, tradução cultural e a descolonização do saber. *Fragmentos*, Florianópolis, v. 21, n. 2, p. 45-59, 2010.

DINIZ, D.; FOLTRAN, P. Gênero e feminismo no Brasil: uma análise da *Revista Estudos Feministas*. *Revista Estudos Feministas*, Florianópolis, v. 12, n. especial, p. 245-253, 2004. DOI: https://doi.org/10.1590/S0104-026X2004000300026

ESTADÃO CONTEÚDO. Violência contra a mulher aumenta em meio à pandemia; denúncias ao 180 sobem 40%. *IstoÉ/Dinheiro*, 1 jun. 2020. Disponível em: https://www.istoedinheiro.com.br/violencia-contra-a-mulher-aumenta-em-meio-a-pandemia-denuncias-ao-180-sobem-40/. Acesso em: 8 ago. 2020.

EVERS, A.; FINATTO, M. J. B. Linguística de Corpus, Léxico-Estatística Textual e Processamento de Linguagem Natural: perspectiva para estudos de vocabulário em produções textuais. *Revista GTLex*, Uberlândia, v. 1, n. 2, p. 271-295, 2016. DOI: https://doi.org/10.14393/Lex2-v1n2a2016-3

FACCHINI, R. *Cadernos Pagu*: desafios, nossas respostas e novidades. *Blog SciELO em Perspectiva*: *Humanas*, [*S.l.*], 30 jun. 2017. Disponível em: https://humanas.blog.scielo.org/blog/2017/06/30/cadernos-pagu-desafios-nossas-respostas-e-novidades/. Acesso em: 8 ago. 2020.

FINATTO, M. J. B. Terminologia e Lingüística de Corpus: da perspectiva enunciativa aos novos enfoques do texto técnico-científico. *Letras de Hoje*, Porto Alegre, v. 39, n. 4, dezembro, p. 97-106, 2004.

GROSSI, M. P. A *Revista Estudos Feministas* faz 10 anos – uma breve história do feminismo no Brasil. *Revista Estudos Feministas*, Florianópolis, v. 12, Número Especial, p. 211-221, 2004. DOI: https://doi.org/10.1590/S0104-026X2004000300023

HAUSER, C. Merriam-Webster Revises 'Racism' Entry After Missouri Woman Asks for Changes. *The New York Times*, New York, s. p., 10 jun. 2020. Disponível em: https://www.nytimes.com/2020/06/10/us/merriam-webster-racism-definition.html. Acesso em: 5 set. 2020.

HEILBORN, M. L.; SORJ, B. Estudos de Gênero no Brasil. *In*: MICELI, S. (org.). *O que ler na Ciência Social brasileira* (1970-1995). São Paulo: Editora Sumaré; ANPOCS; Brasília: CAPES, 1999. p. 183-222.

IBGE. Estatísticas de Gênero - Indicadores sociais das mulheres no Brasil. 2019. Disponível em: https://www.ibge.gov.br/estatisticas/multidominio/genero/20163-estatisticas-de-genero-indicadores-sociais-das-mulheres-no-brasil.html?=&t=o-que-e. Acesso em: 6 set. 2020.

KOESTER, A. Building Small Specialised Corpora. *In*: O'KEEFFE, A.; MCCARTHY, M. (ed.). *The Routledge Handbook of Corpus Linguistics*. New York: Routledge, 2010. p. 66-79.

MALUF, S. W. As edições eletrônicas da REF (e a democratização do acesso à produção acadêmica e científica). *Revista Estudos Feministas*, Florianópolis, v. 16, n. 1, p. 123-127, 2008. DOI: https://doi.org/10.1590/S0104-026X2008000100012

MINELLA, L. S. A contribuição da *Revista Estudos Feministas* para o debate sobre gênero e feminismo. *Revista Estudos Feministas*, Florianópolis, v. 12, número especial, p. 223-234, 2004. DOI: https://doi.org/10.1590/S0104-026X2004000300024

MLAMBO-NGCUKA P. Violência contra as mulheres e meninas é pandemia invisível, afirma diretora executiva da ONU Mulheres. *ONU Mulheres*, 7 abr. 2020. Disponível em: http://www.onumulheres.org.br/noticias/violencia-contra-as-mulheres-e-meninas-e-pandemia-invisivel-afirma-diretora-executiva-da-onu-mulheres/. Acesso em: 8 ago. 2020.

MOREL, J.; RODRÍGUEZ, C. Consecuencias metodológicas de la propuesta teórica. *In*: CABRÉ, M. T.; FELIU, J. (org.). *La terminología científico-técnica*: reconocimiento, análisis y extracción de información formal y semántica. Barcelona: Informe DGES PB-96-0293. Barcelona: Universitat Pompeu Fabra; Institut Universitari de Lingüística Aplicada, 2001. p. 37-53.

PISCITELLI, A. Gênero: a história de um conceito. *In*: BUARQUE DE ALMEIDA, H.; SZWAKO, J. (org.). *Diferenças, igualdade*. São Paulo: Berlendis & Vertecchia, 2009. p. 116-148.

PISCITELLI, A. Atravessando fronteiras: teorias pós-coloniais e leituras antropológicas sobre feminismos, gênero e mercados do sexo no Brasil. *Contemporânea - Revista de Sociologia da UFSCar*, São Carlos, v. 3, n. 2, p. 377-405, 2013.

PISCITELLI, A., BELELI, I., LOPES, M. M. Cadernos Pagu: contribuindo para a consolidação de um campo de estudos. *Revista Estudos Feministas*, Florianópolis, v. 11, n. 1, p. 242-246, 2003. DOI: https://doi.org/10.1590/S0104-026X2003000100015

REVISTA ESTUDOS FEMINISTAS. Políticas Editoriais. [s.d.]. Disponível em: https://periodicos.ufsc.br/index.php/ref/about/editorialPolicies#focusAndScope. Acesso em: 5 set. 2020.

SILVA, S. V. da. Os estudos de Gênero no Brasil: Algumas Considerações. Biblio 3W. *Revista Bibliográfica de Geografía y Ciencias Sociales*, Barcelona, n. 262, p. 1-13, 2000.

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. Barueri: Disal, 2013.

TAGNIN, S. E. O. A Linguística de Corpus na e para a Tradução. *In*: VIANA, V.; TAGNIN, S. (org.). *Corpora na Tradução*. São Paulo: HUB, 2015a. p. 19-56.

TAGNIN, S. E. O. Corpus-Driven Glossaries in Translator Training Courses. *Oslo Studies in Language*, Oslo, v. 7, n. 1, p. 359-377, 2015b. DOI: https://doi.org/10.5617/osla.1447

WAQUIL, M. L. *Traduzindo "Traducción y Traductología"*: problemas terminológicos de tradução. 2017. 288 f. Tese (Doutorado em Letras) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, 2017.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 771-803, 2021

CC BY

Construindo corpora bilíngues quimbundo-português-quimbundo

Building Kimbundu-Portuguese-Kimbundu bilingual corpora

Paulo Jeferson Pilar Araújo
Universidade Federal de Roraima (UFRR), Boa Vista, Roraima / Brasil
paulo.pilar@ufrr.br
https://orcid.org/0000-0002-9965-3444

Resumo: Discute-se neste artigo as possibilidades para a construção de corpora bilíngues quimbundo-português-quimbundo para o estudo de fenômenos de contato linguístico entre essas duas línguas. Faz-se uma aproximação entre as áreas da Linguística de Corpus e da Linguística Africana, enfatizando-se o caso dos contatos linguísticos presentes nos corpora em vista. Defende-se que o quimbundo e o português têm uma relação histórica que permite a elaboração de um corpus escrito a partir de sua tradição descritiva iniciada no século XVII e de um corpus de fala decorrente de projetos de pesquisa recentes que se ocupam de variedades vernaculares do português e sua relação com a língua africana deste estudo. Para tanto, buscou-se fazer um estudo do estado da arte da descrição do quimbundo e do português angolano com o objetivo de demonstrar a necessidade e viabilidade da produção de corpora bilíngues escrito e de fala quimbundo-português-quimbundo. Espera-se que a produção de corpora bilíngues quimbundo-português-quimbundo possa contribuir para o conhecimento da situação de contato visualizado entre as duas línguas, pautando-se em material empírico necessário para o entendimento da real situação de contato entre essas línguas de Angola além de embasar com dados empíricos hipóteses como a de um continuum afro-brasileiro de português.

Palavras-chave: corpora bilíngues; quimbundo; português; contato linguístico; Angola.

Abstract: This article discusses the possibilities for building Kimbundu-Portuguese-Kimbundu bilingual corpora in order to studying language contact phenomena between these two languages. An approximation is made between the areas of Corpus Linguistics and African Linguistics, emphasizing the case of linguistic contact in the corpora in

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.771-803

sight. It is argued that Kimbundu and Portuguese have a historical relation that allows the elaboration of written corpus based on the documents of its descriptive tradition that started in the 17th century and a spoken corpus resulting from recent research projects that deal with vernacular varieties of Portuguese and its relation to the African language of this study. To this end, we sought to study the state of the art of the description of Kimbundu and Portuguese in Angola in order to demonstrate the need and feasibility of building bilingual written and spoken Kimbundu-Portuguese-Kimbundu corpora. It is hoped that the production of these bilingual corpora may contribute to the knowledge of the situation of contact between both Angolan languages under study, based on empirical material necessary to understand the real situation of contact between these languages of Angola besides to support hypothesis about an Afro-Brazilian continuum of Portuguese.

Keywords: bilingual corpora; Kimbundu; Portuguese; Language contact; Angola.

Recebido em 11 de outubro de 2020

Aceito em 25 de novembro de 2020

1 Introdução

Nos meses de julho e agosto de 2013 foi realizado trabalho de campo exploratório no município do Libolo, província do Kwanza Sul, Angola. Foram realizadas várias entrevistas e gravações diversas com os moradores da região e coletados basicamente dados do quimbundo e do português local. Com o início das transcrições dessas entrevistas, em 2014, verificou-se que uma parte considerável dos dados era bilíngue quimbundo-português, apresentando ocorrências de empréstimos e *codeswitching*, trazendo para a produção do corpus específico das línguas do Libolo a problemática de como trabalhar na produção de corpora bilíngues quimbundo-português e português-quimbundo (a partir daqui quimbundo-português-quimbundo) com fins de analisar adequadamente a situação de contato linguístico entre as duas línguas explicitada nos casos de *codeswitching* e de empréstimos encontrados na variedade da língua africana ali denominada *ngoya*, do grupo quimbundo (H20).[1]

[1] A tradição bantuista, desde a classificação das línguas bantas por Guthrie (PETTER, 2015, p. 60) utiliza letras e números para identificação das línguas. Uma letra indica uma zona, por exemplo H, uma letra e número um grupo: H20 é o grupo quimbundo, já H21 é o grupo dialetal mbundo, quimbundo. Assim, ao se fazer referência ao quimbundo

Pensando nisso, buscaram-se exemplos de casos de corpora bilíngues de outras línguas bantas ocupados com a questão dos empréstimos e *codeswitching* que pudessem servir como diretriz para esta investigação. No entanto, mesmo nas duas edições da obra de referência da linguística bantuista (NURSE; PHILIPPSON, 2003; VELDE *et al*., 2019), apesar de um capítulo dedicado para o contato de línguas, quase nada é dito sobre a construção de corpora para línguas africanas. Este artigo discute essa problemática, a da inter-relação entre a Linguística de Corpus (SARDINHA, 2004) com a Linguística Africana (PETTER, 2015) via as questões do contato linguístico de línguas pouco ou não descritas (ADAMOU, 2016), tomando como exemplo o caso da produção de corpora bilíngues quimbundo-português-quimbundo.

As seções que compõem este artigo são: seção teórica e de problematização em 2 onde se busca relacionar a Linguística de Corpus com a Linguística Africana com uma breve apresentação dos corpora de línguas africanas existentes e a particularidade dos corpora bilíngues para essas línguas. A seção 3 apresenta a língua quimbundo (H20) e o português de Angola e os projetos atuais ocupados em descrevê-los. A seção 4 detalha a metodologia na constituição do corpus do Libolo. Em 5 são discutidos mais diretamente a produção de corpora bilíngues quimbundo-português-quimbundo e os desafios de tal empreitada, tomando os fenômenos de contato linguístico como especificidade de corpora bilíngues como os pretendidos neste trabalho. Ao final da seção 5, são exemplificados os casos de empréstimos do português no quimbundo e a forma como a inter-relação entre corpus escrito e de fala pode contribuir para um melhor conhecimento dos contatos históricos entre o quimbundo e o português (angolano e brasileiro) em teorias e hipóteses que tentam explicar a formação do português brasileiro (PB) e um continuum afro-brasileiro de português (LÓPEZ; GONÇALVES; AVELAR, 2018; PETTER, 2008).[2]

---

sem a especificação dialetal será utilizada a indicação do grupo, H20, mesmo sabendo da classificação para o dialeto *ngoya* como H23 (HAMMARSTRÖM, 2019, p. 39), levando em conta que uma investigação sobre o continuum dialetal do quimbundo ainda aguarda uma descrição mais aprofundada.

[2] Na impossibilidade de dedicar uma subseção para conceituar a hipótese de um continuum afro-brasileiro de português citado neste último parágrafo, por fugir do escopo do trabalho, remeto o leitor aos trabalhos supracitados, reservando as próximas seções à temática central do artigo. Da mesma forma, considerando que a literatura sobre

2 A Linguística de Corpus e a Linguística Africana

Diferentemente da Linguística de Corpus e outras disciplinas como a Linguística do Contato para as quais se costumam indicar datas de nascimento (o Corpus Brown 1964 e a obra de Weinreich, 1953, respectivamente), a Linguística Africana tem sua história ligada ao passado colonial, remontando ao século XVI com as primeiras obras de catequese ocupadas com o aprendizado de línguas africanas. Será apenas em meados do século XX que o estudo das línguas africanas toma um caráter disciplinar, conforme periodização apresentada por Petter e Araújo (2015, p. 29-41). No entanto, algumas obras são apontadas como referência na constituição da Linguística Africana como a de Greenberg (1963) *The Languages of Africa*. Mesmo com um histórico extenso na sua constituição, a Linguística Africana ainda se ocupa basicamente na descrição das mais de 2 mil línguas do continente africano, o que de certo modo interfere em uma maior proximidade entre ela e a Linguística de Corpus como esta é conhecida para línguas majoritárias.

Conforme pode ser verificado em obras de referência da Linguística Africana (GÜLDEMANN, 2018; VOSSEN; DIMMENDAAL, 2020; WOLFF, 2019), ao menos um capítulo é dedicado ao contato linguístico, mesmo que transversalmente sobre áreas linguísticas, enquanto para a Linguística de Corpus ocorrem apenas menções a corpus específicos em determinados capítulos. Os estudos baseados em corpus para o caso das línguas africanas são ainda escassos, dentre alguns motivos citam-se: (i) o estatuto de línguas pouco ou não descritas da maioria das línguas africanas; (ii) as políticas linguísticas; e (iii) a representatividade das línguas africanas na glotopolítica internacional. Em breves palavras, o primeiro item desvela a situação das línguas da África e a necessidade de descrição, documentação e publicização dos corpora utilizados no processo de gramatização dessas línguas. O segundo item está relacionado às políticas linguísticas de cada país em relação à língua oficial, geralmente de origem colonial, e as línguas nacionais ou nativas. Cada país tem diferentes políticas, a exemplo da África do Sul que possui 9 línguas africanas com estatuto de língua oficial enquanto Angola possui apenas o português como língua oficial e oferece atenção para 6 línguas

empréstimos e *codeswitching* é bastante ampla e relativamente acessível em artigos, teses e dissertações de acesso aberto, sugiro a consulta a obras mais específicas sobre esses fenômenos do contato linguístico.

nacionais, das mais de 40 línguas do país. Eleger línguas africanas como línguas oficiais é uma medida favorável para a produção de corpora para essas línguas, a exemplo das línguas da África do Sul (ALLWOOD; HENDRIKSE, 2003). O terceiro item tem a ver com a visibilidade e importância política dada às línguas e a preocupação com seu estudo e promoção. O melhor exemplo é o suaíli, considerada língua franca em grande parte do leste africano e que goza de estatuto internacional dentro do continente africano.

Outras questões podem ser elencadas transversalmente àquelas acima, como a da escassa existência de documentos históricos escritos de grande parte das línguas africanas, dentre outras. Apesar da relação parcialmente distante entre a Linguística Africana com a Linguística de Corpus, podem-se citar alguns exemplos de corpora de línguas africanas e um promissor desenvolvimento futuro das duas áreas.

2.1 Corpora de línguas africanas

Roux e Ndinga-Koumba-Binza (2019, p. 632-633) elencam alguns recursos on-line com indicação de corpora de línguas africanas disponíveis. Dentre eles, o *Open Language Archives* (OLAC), que apresenta informações sobre diferentes línguas africanas.[3] Os autores mencionam ainda um blog com informações úteis sobre corpora existentes de línguas africanas.[4] Citam rapidamente os websites da *Columbia University Libraries*,[5] os arquivos da *African Language Technology* (AFlaT),[6] além, claro, do Ethnologue.[7] E por fim, uma iniciativa da *Oxford University Press* com o Projeto *Oxford Global Languages* (OGL).[8] As referências virtuais oferecidas pelos autores não são exaustivas, mas dão uma ideia das empreitadas na construção de corpora de línguas africanas. Podem ser citadas outras iniciativas

[3] Disponível em: http://www.language-archives.org/area/africa Acesso em: 10 out. 2020.
[4] Disponível em: https://corplinguistics.wordpress.com/2012/02/08/african-language-corpora/ Acesso em: 10 out. 2020.
[5] Disponível em: https://library.columbia.edu/ Acesso em: 10 out. 2020.
[6] Disponível em: https://www.aflat.org/ Acesso em: 10 out. 2020.
[7] Disponível em: https://www.ethnologue.com/ Acesso em: 10 out. 2020.
[8] Disponível em: https://languages.oup.com/research/community/ Acesso em: 10 out. 2020.

não mencionadas pelos autores, como o *Langage, Langues et Cultures d'Afrique* (LLACAN),[9] as propostas da *African Academy of Languages* (ACALAN),[10] o CorpAfroAs,[11] dentre outros, principalmente página de centros de Estudos Africanos de diversas universidades. No entanto, esses corpora são de tamanhos modestos comparados com outros, até por se tratar de línguas sub-representadas e de minorias étnicas.

O quimbundo é um exemplo de língua sub-representada que, apesar de ser objeto de estudos e ter uma tradição de publicações no passado, é dificilmente encontrado nesses repositórios de corpora. Alguns fatos históricos explicam em parte a pouca atenção voltada para essa língua banta, considerando sua importância nos estudos sobre a influência africana no português brasileiro, por exemplo. A Guerra Civil Angolana que durou de 1975, ano da independência de Angola, até 2002, contribuiu sobremaneira para o atraso de estudo mais atuais sobre a língua.

2.2 Corpora bilíngues

Em se tratando de corpora monolíngues é possível encontrar alguns recursos on-lines como aqueles apresentados na subseção anterior, no entanto, o cenário se torna menos animador para corpora bilíngues de línguas africanas. Com algumas exceções, por exemplo, em países anglófonos, encontram-se corpora bilíngues para diferentes línguas africanas em relação com o inglês africano (ESIMAJE; GUT; ANTIA, 2019). Talvez o exemplo mais conhecido seja o *Helsinki Corpus of Swahili* (HCS 2.0) em suas versões anotadas e não anotadas.[12] Para a construção de corpora similares ao HCS é necessário um estágio de descrição da língua razoável, o que não é o caso da maioria das línguas bantas (VELDE *et al.*, 2019). As etapas na construção de corpora bilíngues requerem, além do aparato tecnológico bem detalhado na literatura (DEUCHAR, *et al*, 2014; BARRIÈRE, 2016), os recursos

[9] Disponível em: http://llacan.vjf.cnrs.fr/ressources_en.php Acesso em: 10 out. 2020.
[10] Disponível em: https://acalan-au.org/aboutus.php Acesso em: 10 out. 2020.
[11] Disponível em: http://corpafroas.tge-adonis.fr/ Acesso em: 10 out. 2020.
12 Disponível em: https://metashare.csc.fi/repository/browse/helsinki-corpus-of-swahili-20-hcs-20-annotated-version/232c1910b9eb11e5915e005056be118e59fb2e920f1f4c0cafc94915fc6f5cac/ Acesso em: 10 out. 2020. Outro website de interesse está disponível em: https://www.goswahili.org/ Acesso em: 10 out. 2020.

imprescindíveis da gramatização das línguas, ou seja, a produção de gramáticas de referência e dicionários, sendo que o caminho inverso é também válido, o da produção de gramáticas e dicionários baseados em corpora. Mas este último caso para línguas majoritárias com uma tradição de descrição consistente.

Verifica-se ainda que boa parte dos corpora bilíngues como aqueles coligidos em Esimaje, Gut e Antia (2019) voltam-se para aspectos descritivos ou corpus de aprendizagem de língua, aspectos quantitativos são mais modestamente explorados.

3 O quimbundo e o português em Angola

O quimbundo e o português têm um histórico de contato de séculos. É interessante ressaltar essa característica como forma de fundamentar a necessidade da construção de corpora bilíngues para essas duas línguas. Em primeiro lugar pela inegável participação histórica do quimbundo na formação do português brasileiro (BONVINI, 2009), segundo por essas duas línguas ainda estarem em contato contínuo em Angola, ensejando hipóteses de contato linguístico que enfatizem as semelhanças dos processos de mudança nas variedades do português no Brasil e em países africanos lusófonos da área banta, o já mencionado continuum afro-brasileiro de português (PETTER, 2008).

Faz-se, então, nas próximas subseções, uma breve apresentação das línguas angolanas e uma contextualização dos seus estudos.

3.1 O quimbundo e sua tradição descritiva

O quimbundo tem uma tradição descritiva considerável desde o seu passado colonial e missionário, como se pode observar no Quadro 1, no qual são elencados gramáticas, dicionários e coletâneas sobre a língua banta:

QUADRO 1 – Obras sobre o quimbundo (Século XVII ao XX)

Tipo	Ano	Autor	Obra
Gramáticas	1697	Dias	*Arte da língua de Angola*
	1805	Cannecattim	*Collecção de observações grammaticaes sobre a lingua bunda, ou angolense*
	1888/1889	Chatelain	*Gramática elementar do Kimbundu ou* língua de Angola
	1891	Batalha	*A Lingua de Angola*
	1934	Quintão	*Gramática de Kimbundu*
	1946	Baião	*Quimbundo sem mestre*: gramática popular da língua Kimbundu conforme é falada nos distritos de Luanda e Malange O Kimbundu prático ou guia de conversação em Português-Kimbundu (2 v.)
	1951 1957	Maia	*Guia prático para a aprendizagem das línguas Portuguesa e Omumbuim;* *Lições de gramática de quimbundo*
Dicionários	s./d.	Assis Jr.	*Dicionário Kimbundu-Português*
	1804	Cannecattim	*Diccionario da lingua bunda, ou angolense explicada na portugueza, e latina*
	1893	Matta	*Ensaio de diccionario Kimbúndu-Portuguez*
	1961	Maia	*Dicionário complementar português-kimbundo-kikongo*
Diversos	1894/1964	Chatelain	*Folk tales of Angola/Contos populares de Angola*
	1642	Pacconio	*Gentio de Angola sufficientemente instruido nos mysterios de nossa sancta Fé*
	1922	Wendling	*Catecismo da Doutrina Cristã em Portuguez com uma versão em Kimbundo, Dialeto do Libolo*[13]

Fonte: Elaboração do autor

Praticamente todas as obras anteriores ao século XX estão disponíveis em bibliotecas digitais do Brasil, Portugal, Alemanha e

[13] Além dos primeiros catecismos produzidos no século XVII como o de Pacconio para o quimbundo, vale mencionar este catecismo produzido em 1922 no Libolo, de autoria do Padre Victor Wendling e que se encontra em Lisboa, no Centro de Documentação da Província Portuguesa da Congregação do Espírito Santo. Tal documento foi localizado pelo professor Carlos Figueiredo, da Universidade de Macau e está sendo analisado pelo seu descobridor e colegas de pesquisa. Além desse catecismo, o professor Carlos Figueiredo (c. p.) informa que já localizou mais três dicionários, uma gramática e traduções bíblicas para a língua africana, documentos esses ainda em organização para análise.

Estados Unidos. É possível localizá-las em rápidas buscas na Internet. No entanto, versões digitalizadas das gramáticas do quimbundo de Quintão, Baião e Maia são de difícil acesso. A apreciação dessas obras tem sido realizada por diversos autores conforme indicado no trabalho de Fernandes (2015). Vale mencionar aqui a carência de estudos primorosos sobre esses documentos como o estudo de Rosa (2013) para a *Arte da língua de Angola* de Dias.

A disponibilidade de algumas dessas obras em meio digital possibilita o cotejo com descrições mais recentes ou em progresso. É importante frisar que apesar da distância temporal, a gramática do Chatelain por exemplo é considerada ainda como referência nos estudos do quimbundo, sendo utilizada em cursos (principalmente voltado para o público de adeptos de religiões afro-brasileiras como o candomblé angola) e em publicações que carecem de acesso a dados primários da língua.

Uma análise documental dessas obras teria, além do valor histórico, uma importante iniciativa para estudos diacrônicos do quimbundo em cotejo com dados atuais da língua.

3.2 Pesquisas recentes sobre o grupo quimbundo H20 e sua relação com o português

Nos últimos quase 40 anos o quimbundo tem sido objeto de estudo de trabalhos acadêmicos, notadamente teses, como os de Huth (1984), Pedro (1993) e Xavier (2010). Fora esses trabalhos, outros estudos interdisciplinares mais atuais sobre a língua, em seus aspectos históricos (VANSINA, 2001; VIEIRA-MARTINEZ, 2006) figuram entre as publicações que tomam a língua como foco. Vem ressurgindo também o interesse por trabalhos que discutem a situação e classificação do grupo H20 e seus dialetos (ANGENOT, MFUWA, RIBEIRO, 2011; ANGENOT; ANGENOT; HUTA-MUKANA, 2013; SOUSA; KUKANDA; SANTIAGO, 2011) além de análises de documentos históricos sobre o quimbundo (ANGENOT; KEMPF; KUKANDA, 2011; BONVINI, 2009; ROSA, 2013) e sua influência no português tanto no Brasil como na África lusófona (LÓPEZ; GONÇALVES; AVELAR, 2018; OLIVEIRA; ARAÚJO, 2018).

Logo após essa retomada de interesse pelo quimbundo, em 2012 e 2013, tiveram início dois projetos que deram impulso ao estudo dessa

língua angolana e sua relação com o português: o Projeto Temático "A Língua Portuguesa no Tempo e no Espaço"[14] e o Projeto Libolo,[15] respectivamente. Esses dois projetos tiveram subprojetos de doutorado e pós-doutorado relacionados tendo como foco principal a descrição de variedades do português angolano e o quimbundo como língua de substrato. Enquanto o Projeto Temático visava a produzir análises históricas do português a partir do Corpus Tycho Brahe (CTB),[16] o Projeto Libolo, além de outros enfoques, buscava enriquecer o CTB com documentos do eixo África-Brasil. A obra de Oliveira e Araújo (2018) foi produzida no bojo de trabalhos relacionados aos dois projetos, tendo dois capítulos voltados para o quimbundo e o português do Libolo (ARAÚJO; PETTER; JOSÉ, 2018; FIGUEIREDO, 2018).

Verifica-se que, dos primeiros catecismos do quimbundo no século XVII (Cf. QUADRO 1) à sua gramática pedagógica (ARSÉNIO; SEBASTIÃO; ADÃO, 2012), essa língua conta com uma bibliografia considerável em relação às demais línguas bantas de Angola. Falta, no entanto, uma linha de investigação descritiva que contemple todo esse corpo de trabalhos existentes sobre a língua, de modo que a relação histórica entre a língua banta e o português (e por que não outras línguas bantas vizinhas?) seja devidamente analisada. Um dos últimos trabalhos que se ocuparam de fenômenos de contato entre o português e

[14] Disponível em: https://bv.fapesp.br/pt/auxilios/55149/a-lingua-portuguesa-no-tempo-e-no-espaco-contato-linguistico-gramaticas-em-competicao-e-mudanca-pa/ Acesso em: 10 out. 2020.

[15] O projeto *Município do Libolo, Kwanza Sul, Angola: aspectos linguístico-educacionais, histórico culturais, antropológicos e sócio-identitários*, também conhecido como *Projeto Libolo*, é parcialmente financiado pela Universidade de Macau e por entidades privadas filantrópicas de Angola. Trata-se de um projeto internacional e multidisciplinar cujos pesquisadores intervêm, de forma articulada, em pesquisas nas áreas de Linguística, História, Antropologia, Etnografia, Filologia e Ações Pedagógicas. O *Projeto Libolo* é também membro da Cátedra UNESCO em Políticas Públicas para o Multillinguismo e está devidamente patenteado pelo Centro de Investigação e Desenvolvimento (R&DAO) da Universidade de Macau, sob o número de referência SRG011-FSH13-CGF, encontrando-se, desta forma, ao abrigo da vigente proteção de direitos autorais de propriedade intelectual designada por Copyright © 2016, R&DAO *University of Macau*. O Projeto Libolo está com site em construção a ser disponível em: https://www.projetolibolo.com/ Acesso em: 10 out. 2020.

[16] Disponível em: https://www.tycho.iel.unicamp.br/home Acesso em: 10 out. 2020.

o quimbundo é o de Miguel (2019), com foco na variedade de Luanda. Anteriormente, Mingas (2000) se ocupou da mesma temática sob um viés substratista.[17]

4 Metodologia

As propostas e os dados que serão apresentados nas próximas seções são decorrentes da minha participação em estágio de pós-doutorado nos diferentes projetos de pesquisa elencados em 3.2, o Projeto Libolo e o Projeto Temático "A Língua Portuguesa no Tempo e no Espaço" já apresentados. De início a preocupação sobre os estudos do quimbundo e do português da região do Libolo visava a entender os fenômenos de contato linguístico entre essas duas variedades linguísticas considerando que já havia uma equipe ocupada com a descrição da variedade do quimbundo *ngoya* da região e outra equipe ocupada com a variedade do português do Libolo. Sendo assim, seria necessário um trabalho de interseção entre as duas equipes de modo que se contemplasse a questão dos contatos entre a língua banta e a língua oficial angolana.

Em um workshop promovido pela FAPESP e pelo *British Council*, o *Researcher Links*,[18] propus inicialmente a ideia de construção de um corpus anotado do português e do quimbundo como línguas em contato. Desde essa primeira proposta, iniciei a compilação e digitalização das diversas obras sobre o quimbundo (Cf. QUADRO 1). Paralelamente a isso, eram realizadas reuniões de grupos de pesquisa coordenadas pela professora Margarida Petter com o estudo da gramática escrita por Chatelain (1888/1889). Os trabalhos de transcrição de entrevistas e contos

[17] Vale mencionar a ausência ainda de uma gramática de referência do quimbundo. Mark van de Velde (c.p.) informou que foi submetida uma proposta de capítulo sobre o quimbundo para a segunda edição do *The Bantu Languages* (VELDE *et al.*, 2019), mas o proponente não enviou o texto a tempo. Olga Kharytonava conduz nos últimos anos um projeto de pesquisa sobre o quimbundo. Disponível em: http://kimbundu.ca/ Acesso em: 23 nov. 2020.

[18] O workshop *The New Historical Linguiticss and the World of Annotated Corpora* financiado pela FAPESP (Processo 14/50501-9) em convênio com o *British Council* agregou por cinco dias pesquisadores brasileiros e britânicos para discussões sobre a construção de corpora anotados e humanidades digitais. Informações sobre o evento estão disponíveis em: https://www.york.ac.uk/language/research/centres/clhd/nhlwac/ Acesso em: 23 nov. 2020.

em quimbundo foram realizados com o auxílio de um colaborador de pesquisa nativo da variedade *ngoya* do quimbundo (H23) durante um mês em 2014. Com isso, foi possível realizar as análises iniciais sobre o contato do português e do quimbundo do Libolo com dados preliminares apresentados neste trabalho.

As pesquisas que têm sido desenvolvidas no âmbito do Projeto Libolo enfocam notadamente a descrição da variedade de português falado naquela região do Kwanza Sul, conforme os diversos trabalhos produzidos por seus investigadores (FIGUEIREDO; OLIVEIRA, 2016, 2013; FIGUEIREDO, 2018, 2016). Em uma nova fase do Projeto Libolo, iniciada em 2018, em parceria com o Laboratório de Estudo Empíricos e Experimentais da Linguagem (LEEL) da Faculdade de Letras da Universidade Federal de Minas Gerais (UFMG), os dados do português do Libolo têm sido coletados e transcritos seguindo a metodologia do Projeto C-Oral-Angola, como mostram duas publicações (OLIVEIRA; ZANOLI; ANDRADE, 2018; ROCHA; MELLO; RASO, 2018). Antes da escolha metodológica pela família C-ORAL[19] para o tratamento dos dados do português do Libolo, o processo de transcrição dos dados do português do Libolo foi iniciado com uma proximidade com a chave de transcrição do Projeto Vertentes, utilizada nos trabalhos em Lucchesi, Baxter e Ribeiro (2009).[20]

Os materiais orais coletados em 2013 no Libolo eram então transcritos de forma contínua, contando com o trabalho de alunos de Iniciação Científica vinculados ao Projeto Libolo, passando por uma revisão feita pelos seus coordenadores. Para as transcrições das entrevistas em quimbundo ou com trechos em quimbundo e português, os pesquisadores recorriam a o auxílio de um falante da língua africana que soubesse ler e escrever em quimbundo. Para que se tenha uma ideia da constância da presença do quimbundo e do português nos dados, as Tabelas 1 e 2 apresentam as informações das entrevistas utilizadas nas primeiras análises. A Tabela 1 apresenta cerca de 4 horas de entrevistas da equipe do português, cada uma atentando para a caracterização da L1 e L2 dos entrevistados. Basicamente, metade dos entrevistados tinham

[19] Para uma visão sobre o referido projeto, conferir o site do C-ORAL Brasil, disponível em: http://www.c-oral-brasil.org/ Acesso em: 23 out. 2020.

[20] Disponível em: http://www.vertentes.ufba.br/ Acesso em: 10 out. 2020.

uma língua africana como L1, sendo que um deles tinha o umbundo como L1:[21]

TABELA 1 – Transcrições da Equipe do Português do Libolo[22]

Identificação	Sexo	Idade	Local/ Comuna	Duração	L1	L2
[TEMALM3]	feminino	38	Mbanza da Cabuta/ Cabuta	00:25:18	Quimbundo	Português
[HALDOM2]	feminino	20	Mbanza do Kitondo/ Cabuta	00:10:26	Português	Quimbundo
[JOMAJH2]	masculino	15	Calulo	00:19:38	Português	Quimbundo
[HALDOM2]	masculino	68	Mbanza do Kitondo/ Cabuta	00:09:36	Português	Quimbundo
[DOKITHX]	masculino	?	Mbanza do Kitondo/ Cabuta	00:06:51	Quimbundo	Português
[MIJOMH2]	masculino	20	Mbanza do Kitondo/ Cabuta	00:08:09	Quimbundo	Português
[LUSAMH1]	masculino	8	Calulo	00:08:59	Português	?
[ALBAGH4]	masculino	43	Calulo	01:00:20	Quimbundo	Português
[ANPAVM4]	feminino	53	Calulo	00:48:35	Português	Quimbundo
[VACHIH5]	masculino	67	Fazenda da Quitila/ Calulo	00:51:15	Umbundo	Português
Total do tempo de gravações:				***04:12:00***		

Fonte: Elaboração do autor

Quanto à equipe do quimbundo, a base de dados inicial constituía-se de cerca de 29h de áudio compreendendo entrevistas, relatos, contos, provérbios e sessões de reunião sobre o ensino do quimbundo nas escolas. A temática das entrevistas era, na sua maioria, sobre o quimbundo ou do cotidiano dos entrevistados. Nas entrevistas, os documentadores fazem perguntas sobre os entrevistados e sobre o uso do quimbundo na comunidade e na família. Em quase todas as entrevistas houve a participação de um intérprete; em outras vezes o próprio entrevistado

[21] O município do Libolo é composto por quatro comunas (distritos): Calulo, sede do município, Munenga, Cabuta e Quissongo.

[22] A Tabela em questão foi organizada a partir das transcrições disponibilizadas por um dos pesquisadores da equipe do português responsável pela organização e sistematização do corpus do Libolo, que será acessível em uma futura *webpage* do Projeto.

falava em quimbundo, dando uma tradução de sua fala em seguida. Somam-se aos áudios da língua africana o acervo de vídeos produzidos nas quatro comunas do município do Libolo.[23]

As primeiras entrevistas que passaram por transcrição, com cerca de 2 horas e meia no total, não tiveram a categorização por L1 ou L2 do falante como na Tabela 1, já que as entrevistas ocorreram predominantemente em quimbundo, com exceção das duas entrevistas mais longas. Praticamente todos os entrevistados pela equipe do quimbundo são bilíngues em português e quimbundo e em resposta ao formulário sociolinguístico não souberam informar qual das línguas consideraram primeira ou segunda. As entrevistas foram classificadas então como tendo o quimbundo ou o português como predominante:

TABELA 2 – Transcrições da Equipe do Quimbundo

Tipo de entrevista	**Identificação**	**Local/ Comuna**	**Duração**	**Língua predominante na entrevista**	**Língua secundária na entrevista**
Entrevista mediada por um intérprete	[LUAMAR]	Jongo/Quissongo	00:18:46	Quimbundo	Português
Interlocutores em espaço público	[VARINT1]	Mercado Kamama/ Calulo	00:03:55	Quimbundo	Português
Entrevista em espaço público	[VARINT2]	Mercado Kamama/ Calulo	00:04:27	Quimbundo	Português
Contos em uma reunião familiar	[GILHH1]	Jongo/Quissongo	00:42:36	Português	Quimbundo
Reunião Pedagógica	[VARPROF]	Missão Católica/ Calulo	01:06:38	Português	Quimbundo
Total do tempo de gravações:			***02:25:03***		

Fonte: Elaboração do autor

Para a transcrição dos dados do quimbundo, antes da nova fase em 2018, os pesquisadores da equipe do quimbundo utilizavam o ELAN

[23] Santos (2015, p. 66), em nota de rodapé, informa que o espólio do Projeto Libolo contava com cerca de 150 horas de material para análise, entre áudios e vídeos, depois da pesquisa exploratória ao Libolo em 2013, até aquela data. Só de áudio, cada equipe contribuiu com as seguintes quantidades de horas: 40 horas de gravações realizadas pelas equipes de Linguística; 50h de entrevistas realizadas pela equipe de História e cerca de 21 horas de entrevistas realizadas pela equipe de Antropologia.

por esse software permitir a criação de trilhas que indicassem os casos de empréstimos e *codeswitching* recorrentes nas entrevistas.

Neste ponto, vale oferecer algumas informações quanto às questões éticas na coleta dos dados em Angola. Os membros do Projeto Libolo têm seguido as normas vigentes de Angola considerando a realidade política do país, conforme procedimentos descritos por Rocha, Melo e Tommaso (2018, p. 143) e por Figueiredo *et al*. (2016, p. 21-22). Os autores relatam que foram realizadas reuniões com os administradores do Município do Libolo, além de solicitada a permissão dos sobas, as autoridades comunitárias. As visitas às comunas e as entrevistas eram sempre acompanhadas por um representante do governo angolano e com a participação do soba da localidade, corroborando as palavras de Rocha, Melo e Tommaso (2018, p. 143) que afirmam: "O *Soba*, assistido pelos *sobetos*, é de fato a maior autoridade civil da comunidade, desde os tempos pré-coloniais."

5 Para a produção de corpora bilíngues quimbundo-português-quimbundo

Diante do material considerável sobre o quimbundo e a retomada de interesse por essa língua em diversos projetos de investigação, é proposta nas próximas subseções a construção de diferentes corpora, de modo a se pensar também na produção de corpora específicos quimbundo-português-quimbundo, em formato de corpora paralelos e/ou bilíngues.

5.1 Corpus escrito

O Corpus Tycho Brahe (CTB) já mencionado constitui-se de diversos corpora preocupados com a história do português, tanto em Portugal como no eixo África-Brasil (GALVES, 2018, 2019). Conforme Galves (2018, p. 49): atualmente o corpus contém 76 textos (= 3.302.696 palavras) e pretende ser alargado para 1.500.000 palavras de textos portugueses, 600.000 palavras de textos brasileiros e 150.000 de documentos africanos (GALVES, 2018, p. 50).

Além do corpus do Português Histórico e o Corpus Sintático, o Tycho Brahe abarca ainda o corpus Cafundó e Kadiwéu.[24] É nessa nova plataforma multilíngue do CTB que as gramáticas e documentos

[24] Disponível em: https://www.tycho.iel.unicamp.br/browser. Acesso em: 10 ago. 2020.

históricos do quimbundo podem figurar, sendo processados pelas ferramentas como eDictor[25] (PAIXÃO DE SOUZA; KEPLER; FARIA, 2012), utilizado na produção do corpus Kadiwéu (GALVES *et al.,* 2017). Em princípio, pelo valor histórico e pelo empenho de estudos historiográficos (ROSA, 2013), as gramáticas do século XVII e XIX, por já estarem digitalizadas e disponibilizadas em diferentes plataformas (Cf. QUADRO 1) podem ser as primeiras a serem inseridas em um corpus quimbundo-português-quimbundo escrito no CTB. De qualquer modo, tanto esses documentos digitalizados quanto as gramáticas do século XX enriquecerão sobremaneira os estudos não apenas historiográficos como também de descrição linguística.

Deve-se considerar que esses documentos são bilíngues, em português e quimbundo, e que a inserção em plataformas como o CTB deve levar em conta as particularidades de corpora bilíngues (BARRIÈRE, 2016; DEUCHAR *et al*., 2014;). Por exemplo, a gramática de Dias (1697) representa também o português seiscentista (ROSA, 2013) e o quimbundo daquele período pode ser considerado um quimbundo clássico. Os mecanismos de etiquetagem e de busca para a produção de um corpus anotado poderá auxiliar pesquisadores do contato a procurar por empréstimos e construções que indiquem também a influência do português sobre o quimbundo (Cf. 5.4). A inserção de documentos históricos do quimbundo no CTB foi pensada em 2015, mas esbarrava naquele momento no manejo de textos bilíngues como é o caso das gramáticas e documentos diversos do quimbundo. Tais questionamentos deverão ser úteis para as demais plataformas com caráter bilíngue no referido Corpus.

5.2 Corpus de fala

Ciente de que boa parte dos trabalhos do Projeto Libolo se concentra sobre a variedade de português desse município, os pesquisadores do referido projeto logo perceberam o contato entre o português e a língua banta local, como mostra o exemplo (1) abaixo em que desde as transcrições preliminares da pesquisa de campo exploratória ao Libolo em 2013 a língua africana já se fazia presente:

[25] Disponível em: https://humanidadesdigitais.org/edictor/ Acesso em: 23 nov. 2020.

(1) Português do Libolo [TEMALM3]

DOC1: E assim vocês falavam com o avô, não é? Você lembra esse tempo ainda?

INF: É// quan... quando falava com o avô?

DOC1: Hum.

INF: Lembro. Ele te manda. Fala assim [*fala em quimbundo*] em quimbundo já quando falava [*fala em quimbundo*]. O avô já é assim porque ele *num* sabe falar português. É. Português fala vai buscar panela, vai buscar cesto.

DOC1: Ham.

INF: Agora ele que *num* sabe te fala embora [*fala em quimbundo*].

Exemplos como em (1) confirmam que situações de bilinguismo são pervasivas nas entrevistas coletadas no Libolo, conforme constatado nas Tabelas 1 e 2 da seção 4. O recurso utilizado no início das transcrições era de indicar a alternância de língua do português para o quimbundo, ou vice-versa, entre colchetes: [*fala em quimbundo*]. Com o trabalho da equipe do quimbundo e o auxílio de colaboradores bilíngues quimbundo-português, os casos de entrevistas contendo ocorrências de *codeswitching* ou empréstimos puderam ser melhor detectados, transcritos e analisados. No entanto, considerando duas das publicações mais recentes seguindo a metodologia do C-ORAL-Angola (OLIVEIRA; ZANOLI; ANDRADE, 2018; ROCHA; MELLO; RASO, 2018), verifica-se ainda a proeminência dos estudos da variedade de português do Libolo em relação com a língua africana. Ressalta-se, portanto, que o ideal é a consolidação de corpora do quimbundo em paralelo com corpora do português em Angola por motivos que serão melhor detalhados nas subseções que seguem.

Por ora vale enfatizar que uma das línguas bantas de Angola tem agora a possibilidade de entrar no rol de línguas africanas que podem contar com corpora bilíngues, a exemplo dos corpora suaíli-inglês (Cf. 2.1 e 2.2). Para isso, é preciso que o quimbundo receba a atenção devida por parte dos pesquisadores do Projeto Libolo e que as questões de contato linguístico sejam consideradas e debatidas na produção dos corpora dos projetos que se ocupam de uma ou outra língua ou mesmo de ambas as línguas do Libolo. A depender da metodologia usada na constituição de corpora de fala do quimbundo, seja no bojo da família C-ORAL ou de outros corpora de línguas bantas como o HCS 2.0 (Cf. 2.2), o suporte empírico de corpora de fala contribuirão inestimavelmente nos trabalhos de descrição da língua quimbundo

propriamente dita e nos fenômenos decorrentes do contato com o português. Outra possibilidade é a construção de corpora paralelos quimbundo-português para o caso de dados como os de [GILHH1] da Tabela 2 em que um mesmo conto de animais foi contado por um idoso em quimbundo e em seguida recontado em português por um jovem da comuna do Quissongo.

Outro detalhe a ser considerado é a necessidade de descrição linguística do quimbundo (H20) e seu contínuo dialetal, no caso o *ngoya* (H23) frente a outros grupos dialetais. Desse modo, a constituição de corpora do quimbundo deve andar junto com o melhor entendimento da gramática da língua banta ensejando a produção de gramáticas descritivas que devem auxiliar os pesquisadores bantuistas e do português com um conhecimento mais seguro da língua e seus contatos com as variedades de português angolano (ARAÚJO; PETTER; JOSÉ, 2018).

5.3 Questões de contato linguístico entre o quimbundo e o português

Nesta subseção são exemplificados os possíveis casos que a produção dos corpora deverá encarar no manejo dos dados das variedades linguísticas do Libolo. Foram selecionados alguns exemplos que ilustram a realidade bilíngue do Libolo. Com isso, a alternância de língua nas entrevistas é bem documentada, ocorrendo exemplos como abaixo:

(2) *Codeswitching* português-quimbundo [VARINT1]

Samo daqui mesmo, ***kumbala iami ku Kibuma***, tava lá na Kibuma. É agora que vieu (viemos?) aqui *na* Kapemba.

'Somos daqui mesmo, ***o meu bairro é o Kibuma***, (mas) estávamos lá na Kibuma. É agora que viemos para Kapemba.

Na entrevista em (2), a língua predominante era o português, mas a falante alterna com o quimbundo além de produzir formas que indicam ser o português sua L2, o caso de "vieu", a tentativa de utilizar o passado do verbo "vir" sem a concordância com a primeira pessoa do plural iniciada na sentença. Esse exemplo é retirado de transcrições prévias realizadas ainda em Angola. Observa-se que o trabalho de segmentação e de glosa ainda não havia sido realizado com auxílio do colaborador bilíngue.

Como as entrevistas da equipe do quimbundo se concentravam sobre a língua banta, ocorrem principalmente exemplos de *codeswitching* quimbundo-português, como no exemplo (3):

(3) *Codeswitching* quimbundo-português[26] [VARINT1]

Aí na Bula.	Iji	wo	ku	mu-igile	***esse nu lhe conhece***.		
Aí, na Bula	Disse	lhe	NEG	MO-conhecer	***esse***	***não lhe conhece***.	
Mu-kage	a	***só***		***Fikeletu***	we-kexi	ku	Longa.
CL-Mulher	dele,	***senhor***		***Figueiredo***	2sg-COP	LOC	Longa.

'Aí no (bairro) Bula. Disse se você lhe conhece, esse não lhe conheço. A mulher dele, do senhor Figueiredo que estava no Longa.'

Observa-se que tanto nas entrevistas em português ou em quimbundo há a alternância entre as línguas, caracterizando alguns casos de *codeswitching* e em outros a utilização de empréstimos, talvez o caso de *só Fikeletu* para "senhor Figueiredo". Por se tratar de dados de fala espontânea, alguns exemplos são de produção de sentença em uma língua seguidos da repetição em outra língua, como em (4):

(4) *Codeswitching* quimbundo-português [GILHH1]

êh	Ni-lombol-a	***não***	***tem***	***medo***	eme	ke	ni	kala
MD	MS-pedir-VF	***não***	***tenho***	***medo***	1ps	NEG	COM	COP
uoma	uaia							
medo	PRON							

'Oh... peço com pena... não tenho medo... eu não tenho medo deles...'

---

[26] Para os exemplos em línguas africanas, utiliza-se uma transcrição ortográfica, sem marcação tonal, apresentando primeiramente os segmentos e as glosas, em seguida uma tradução livre entre aspas simples. Para fins de simplificação, não é indicada a numeração das classes nominais, comum na literatura bantuista. As abreviaturas das glosas são: 1, 2, 3 sg = primeira, segunda, terceira pessoa singular; 1, 2, 3 pl = primeira, segunda, terceira pessoa plural; CL = classe nominal; CONJ = conjunção; COM = comitativo; COP = cópula; IDEO = ideofone; INF = infinitivo; LOC = locativo; MD = Marcador discursivo; MS = marca do sujeito; MO = marca do objeto; NEG = negativa; RFL = reflexivo; PPF = pré-prefixo; PRON = pronome; TAM = marca de tempo, modo e aspecto; VF = vogal final. Para uma rápida apresentação da estrutura das línguas bantas, sugiro o capítulo 2 de Araújo (2013).

Esse tipo de dados em (4) é bastante comum nas entrevistas quando os interlocutores sabiam do interesse dos pesquisadores no quimbundo. Os entrevistados falavam em quimbundo seguindo de uma tradução. Esses dados demonstram a necessidade de uma categorização dos exemplos, a depender dos objetivos de pesquisa.

Além desses casos, verificam-se outros de adaptação morfofonológica de palavras e expressões do português no quimbundo, o que pode ser denominado como uma quimbundização (equivalente ao aportuguesamento) do português do Libolo. No exemplo abaixo vê-se a repetição que seria em português do quimbundo, mas com a realização dos itens em conformidade com o sistema fonológico da língua banta:

(5) *Codeswitching* quimbundo-português[27] [GILHH1]

êh	***sô***	mu-ki-fut-a	ohi?	***Paka***	***kiê,***	***paka***	***kwantu***?
MD	***senhor***	CL-?-pagar?-VF	quanto?	***Paga***	***quê***	***paga***	***quanto***

'É... então senhores pagam quanto? Paga o quê? Paga quanto?'

Ao invés de uma alternância simples para o português que levaria a forma "paga quê, paga quanto", o falante utiliza formas adaptadas à fonologia do quimbundo: *paka* e *kiê*, e mesmo *sô* para "senhor".

Os exemplos de (2) a (5) confirmam que os fenômenos de contato entre o quimbundo e o português são bastante arraigados na comunidade de fala do Libolo. Conhecer melhor a situação de contato contribui para uma caracterização do estatuto de cada língua e as consequências sociolinguísticas desses contatos. Por exemplo, qual a melhor forma de descrever a situação atual de contato entre o quimbundo e o português no Libolo? Convergência de línguas, atrito ou substituição de língua (*language shift*)? Araújo e Petter (Manuscrito) levantam alguns desses questionamentos que apenas uma análise dos contatos quimbundo-português podem esclarecer melhor. Por outro lado, entender os contatos linguísticos entre as línguas do Libolo pode colaborar no trabalho de etiquetagem na produção de corpora anotados quimbundo-português-

[27] É indicado com o sinal de interrogação "?" no lugar da glosa sempre que houver alguma dúvida quanto à categoria de um morfema ou a melhor glosa para uma categoria. O exemplo (5) foi transcrito com auxílio de colaborador falante nativo do quimbundo e a sugestão da grafia *paka* e *kiê* indicam a acomodação do português no quimbundo L1 dos falantes.

quimbundo. Em outras palavras, frente a exemplos como os enumerados acima, como definir adequadamente se são situações de empréstimo ou *codeswitching*? A próxima subseção busca exemplificar este último questionamento.

5.4 Os corpora bilíngues escrito e de fala e análises do quimbundo e do português como línguas em contato

Com a produção do corpus bilíngue específico do português e do quimbundo será possível verificar a real produtividade e integração de alguns empréstimos do português no quimbundo e vice-versa. Enquanto alguns empréstimos parecem já figurar em gramáticas da língua de fins do século XIX e meados do século XX, outras ocorrências caem na dubiedade entre *codeswitchiing* e empréstimo, a exemplo de *xitalata* (estrada) a seguir:

(6) Quimbundo [VARINT1]

Jina	die	Losita.	W-aiula	**o**	**xitalata**	xa	Longa
Nome	dele	Rochita	2ps-arranjar	PPF	***estradas***	do	Longa
ne	xa	Kabuta	ne	ia	Kisongo.		
CONJ	da	Cabuta	CONJ	do	Quissongo.		

'O nome dele é Rochita, que arranjou as estradas do Longa, da Cabuta e do Quissongo.'

A aparente dúvida se manifesta por não se ter informação certa se *xitalata* é realmente usado em todos os contextos referentes a estrada ou apenas em contextos específicos, já que existe palavra para estrada em quimbundo. Outro ponto é o uso do pré-prefixo *o* do quimbundo comumente confundido com o artigo masculino singular do português. Estudos sobre a integração de empréstimos do português em línguas angolanas é praticamente inexistente, por outro lado, o da situação inversa é mais conhecido. É o caso do trabalho de Miguel (2019) que partindo de 36 entrevistas realizadas em Luanda em 2012 e 2013 selecionou 255 empréstimos lexicais, em sua maioria do quimbundo. No entanto, como as entrevistas foram conduzidas em português e o foco do trabalho era a integração de empréstimos bantos no português de Luanda, o caso inverso não é mencionado, ou seja, o da integração de empréstimos do português no quimbundo. A produção de um corpus escrito e anotado do

quimbundo poderá contribuir na identificação dos possíveis empréstimos já registrados em gramáticas do quimbundo e que indiquem seu real estatuto de empréstimos integrados.

Nos próximos exemplos, de (7) a (14), são apresentados casos de empréstimos de preposições "para" e "até" e conjunções, "mas" e "se" do português no quimbundo registrados tanto nas entrevistas orais quanto nas gramáticas do século XIX e XX. Esses exemplos são interessantes porque logo no início das transcrições esses casos eram encarados como *codeswitching* ou *nonce borrowing* (empréstimos de uma única vez, em tradução livre) até que foram encontrados exemplos similares nas gramáticas do quimbundo ainda no século XIX, o que indica que podem ser na verdade empréstimos já integrados na língua banta.

Para fins de padronização foram acrescidas glosas nos exemplos retirados das gramáticas do quimbundo, seguindo as mesmas abreviaturas já propostas.

– Preposição "para" do português como *pala* no quimbundo:

(7) Corpus do quimbundo do Libolo [LUAMAR]

tutulukutu	***pala***	ku-zol-a
IDEO (bater os pés)	***para***	INF-rir-VF

'Batam os pés para rir'

(8) (MAIA, 1964, p. 94)

Eme	ng-el-e	ko	Sikola	***pala***	ku-li-longes-a
1sg	MS.1sg-vir-VF	LOC	escola	***para***	INF-RFL-ensinar-VF

'Eu vim à escola para aprender'

– Preposição "até" do português como *katé* no quimbundo:

(9) Corpus do quimbundo do Libolo [GILHH1]

Otxó	we-riendé	***katé***	obó,	mumamé	iji:	eie
quando	MS-andar	***até***	lá	mulher	dizer	2sg
wa-landuka	eie					
MS-malandrar	2sg					

'Quando ele andou até lá, a mulher dele disse: você malandreia'

(10) (CHATELAIN, 1888/89, p. 116)

Tunde mu Luanda ***katé*** mu Ndongo **ji-lekua**.
Desde LOC Luanda ***até*** LOC Ndongo CL-léguas
'De Luanda até o Ndongo são léguas'

– Conjunção "mas" do português como *maji* no quimbundo:

(11) Corpus do quimbundo do Libolo [GILHH1]

bit-el-a kuno tu-ban-e ***maji*** ku-be kibalo
chegar-?-VF aqui 3pl-dar-VF ***mas*** NEG-dar contribuição
Chega aqui, damos, **mas** não dás contributo.

(12) (MAIA, 1951, p. 114)

É, iene muê, o-kuendje; ***maji*** ke ***mbalão*** mokonda
é isso mesmo PPF-rapaz ***mas*** NEG ***balão*** porque
íí lo ***motor*** ***pala*** u-ana lo mapapa ***pala***
COP COM *motor* ***para*** MS-puxar COM asas ***para***
ukatuka ku elu.
ficar LOC céu
'É isso mesmo, rapaz; mas não é balão porque tem motor para o puxar e asas para se conservar no ar'

– Conjunção "se" do português como *se* no quimbundo:

(13) Corpus do quimbundo do Libolo [LUAMAR]

eie u-a-il-e ni o mu-ana mu ***xicola*** ***se***
2ps MS-TAM-ir-VF COM PPF cl-criança LOC ***escola*** ***se***
ue-kala o mu-an-a mu ***xicola***
2ps-COP PPF cl-criança LOC ***escola***
'Você foi com o filho à escola, se estava o filho na escola.'

(14) (CHATELAIN, 1988/89, p. 49)

Se ua-le-valel-e, uo-jofuta

se MS-TAM-dever-VF MS-pagar

'Se ele devesse, havia de pagar'

Além dos empréstimos de preposições e conjunções, já se verificam empréstimos de nomes como *ji-lékua* (léguas) em (10), *mbalão* (balão) e *motor* em (12) e *xicola* (escola) em (13).

Todos esses exemplos que caracterizam o quimbundo e o português do Libolo como línguas em contato devem ser considerados na construção dos corpora bilíngues das línguas. Deve ter ficado patente nos exemplos discutidos que a perspectiva enfatizada neste trabalho recai mais sobre a língua africana, considerando-se que em grande parte dos debates teóricos sobre a relação entre o quimbundo e o português enfatiza-se as influências da(s) língua(s) banta(s) no português (MINGAS, 2000; MIGUEL, 2019) subjacente até mesmo na própria hipótese de um continuum afro-brasileiro de português (LÓPEZ; GONÇALVES; AVELAR, 2018; PETTER, 2008).

Com a organização dos corpora bilíngues pretendidos e propostos neste trabalho, espera-se que se tenha um entendimento maior sobre a influência do português sobre o quimbundo ao se analisar a produtividade dos empréstimos *pala*, *katé*, *maji* e *se*, dentre outros. Casos como esses indicam que o histórico de contato entre o quimbundo e o português favorecem uma situação de convergência, nos termos de Myers-Scotton (2002). No entanto, tais elucubrações só serão de fato constatadas com o exame em corpora escritos e de fala quimbundo-português-quimbundo, como demonstrado rapidamente nos exemplos (7) a (14).

Retomo, assim, algumas considerações das subseções 5.1 e 5.2 no que diz respeito à complementaridade entre esses diferentes corpora. O quimbundo, como já apresentado em 3.1, goza de uma relativa tradição descritiva que pode contribuir com a constituição de trabalhos históricos, filológicos ou diacrônicos (BONVINI, 2009; ROSA, 2013) sobre a língua e sua relação com o português, no Brasil e em Angola. Reafirmo, portanto, a importância de se levar a cabo a inserção mais que oportuna dos documentos escritos do quimbundo no CTB, como também a necessidade de que os pesquisadores do Projeto Libolo levem a descrição do quimbundo mais seriamente, não apenas para fins comparativos ou contrastivos com variedades do português, mas para uma descrição da

realidade sociolinguística do Libolo condizente com um contexto de multilinguismo, conforme observada na área de transição etnolinguística encontrada na região do Libolo.

Um outro detalhe que deve ser mencionado na construção dos corpora bilíngues é justamente a realidade sociolinguística do Libolo, que poderá trazer mais surpresas. Por ser área de transição etnolinguística entre as zonas H e R (ANGENOT; MFUWA; RIBEIRO, 2011), pode decorrer disso a possibilidade de fenômenos de contato não apenas entre o português e o quimbundo, mas do grupo quimbundo (H20) com o grupo umbundo (R10), o que poderá acarretar dados bilíngues quimbundo-umbundo ou mesmo trilíngues quimbundo-umbundo-português. Tal constatação é exemplificada no caso de *codeswitching* quimbundo-umbundo em (15):

(15) *Codeswitching* quimbundo-umbundo [LUAMAR]

D - A senhora fala quimbundo? *o kimbundu zuela*?
quimbundo... fala...?

A - ***wa*** eye eye eye **o** **limi** **lia** o kimbu.ndu mwene?
MS 2ps 2ps 2ps PPF língua de PPF quimbundo mesmo
'Você, você, você, (fala) a língua do quimbundo mesmo?'

No trecho da entrevista acima, o documentador D pergunta em português, logo depois tenta repetir a pergunta em quimbundo. Já o intérprete A pergunta para a informante, inicia a pergunta em quimbundo e produz um *codeswitching* com o umbundo. No processo de transcrições das entrevistas, principalmente das comunas mais distantes do centro do município, Calulo, como as do Quissongo, esse tipo de *codeswitching* entre quimbundo e umbundo é esperado por ser aquela região uma zona de transição etnolinguística. Veja-se ainda que um dos entrevistados na Tabela 01 tinha como língua materna o umbundo.

6 Conclusão

Trazendo para a discussão a relação mais do que necessária entre a Linguística Africana (PETTER, 2015) e a Linguística de Corpus (SARDINHA, 2004) como áreas disciplinares independentes e que podem ser mais relacionadas, este artigo tomou como foco o caso da

produção de corpora bilíngues quimbundo-português-quimbundo em suas modalidades escrita e de fala para apresentar e exemplificar as possibilidades de explorar a situação de contato existente no caso da língua banta em sua história com o português.

Com uma breve apresentação dos corpora existentes de línguas africanas e a carência de mais corpora bilíngues dessas línguas (subseções em 2), o quimbundo e o português de Angola foram discutidos em 3, desde os documentos históricos sobre a língua africana deste estudo (em 3.1) aos estudos atuais dedicados às duas línguas (em 3.2). As pesquisas sobre as relações de contato entre o quimbundo e o português tiveram um impulso nas últimas 3 a 4 décadas com pesquisas acadêmicas sobre a língua banta, mas principalmente pela atenção de projetos de cunho internacional, como o Projeto Temático "O Português no Tempo e no Espaço" e o Projeto Libolo". A metodologia foi apresentada da forma como era seguida inicialmente pelos pesquisadores do Projeto Libolo no tratamento dos dados das variedades linguísticas do quimbundo e do português presentes no Libolo, interior de Angola (Kwanza Sul). Através desses dois projetos, foram apresentadas propostas de construção de corpus escrito (em 5.1) para o quimbundo e o português, sugerindo-se a plataforma multilíngue do Corpus Tycho Brahe (CTB) com o uso de suas ferramentas de tratamento de documentos históricos. Já para a produção de um corpus de fala (em 5.2) foram discutidos algumas problemáticos sobre o tratamento de dados bilíngues em corpora. Por fim, em 5.3 foram discutidas questões relativas ao tratamento de ocorrências de *codeswitching* e empréstimos nesses a partir de exemplos retirados do corpus do Projeto Libolo. Em 5.4 foram apresentados dados de como a correlação entre os corpora escrito e de fala podem corroborar a integração de empréstimos do português no quimbundo, já que as pesquisas sobre a integração de empréstimos do quimbundo no português recebem uma maior atenção dos pesquisadores (MIGUEL, 2019).

Espera-se que este artigo descortine iniciativas antevistas nas seções anteriores, assim como o estímulo de pesquisas sobre o contato linguístico entre o português e o quimbundo para tornar mais empiricamente informados os estudos sobre as chamadas influências das línguas bantas nas variedades de português do Brasil e de Angola, assim como a hipótese de um continuum afro-brasileiro de português (LÓPEZ; GONÇALVES; AVELAR, 2018; PETTER, 2008).

Agradecimentos

Agradeço a Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) pela bolsa de pós-doutorado (Processo 13/20567-5), que financiou parte dos resultados de pesquisa deste artigo decorrentes do projeto "O português e o quimbundo (H20) do Libolo, Kwanza Sul, Angola – avaliando modelos teórico de línguas em contato" vinculado ao Projeto Temático "A Língua Portuguesa no Tempo e no Espaço: contato linguístico, gramáticas em competição e mudança paramétrica" (Processo 12/06078-9). Agradeço ainda a um parecerista anônimo por suas sugestões e observações que muito contribuíram para uma apresentação da temática deste artigo mais acessível. Os problemas remanescentes são de minha total responsabilidade.

Referências

ADAMOU, E. *A Corpus-Driven Approach to Language Contact*: Endangered Languages in a Comparative Perspective. Berlim: de Gruyter, 2016. DOI: https://doi.org/10.1515/9781614516576

ALLWOOD, J.; HENDRIKSE, A. Spoken Language Corpora for the Nine Official African Languages of South Africa. *Southern African Linguistics and Applied Language Studies*, [*S.l.*], v. 21, n. 4, p. 189-201, 2003. DOI: https://doi.org/10.2989/16073610309486343

ANGENOT, J.-P.; ANGENOT, G. de L.; HUTA-MUKANA, D. M. Comparision between the Ipala-Ngoya, Kimbundu and Umbundu Tone-Cases Systems. *Revista Língua Viva*, Porto Velho, RO, v. 3, n. 1, p. 1-28, 2013.

ANGENOT, J.-P.; KEMPF, C. B.; KUKANDA, V. Arte da Lingua de Angola de Pedro Dias (1697) sob o prisma da dialetologia Kimbundu. *Papia*, São Paulo, v. 21, n. 2, p. 231-252, 2011.

ANGENOT, J.-P.; MFUWA, N.; RIBEIRO, M. A. As classes nominais do kibala-ngoya, um falar bantu de Angola não documentado, na intersecção dos grupos kimbundu [H20] e umbundo [R10]. *Papia,* São Paulo, v. 21, n. 2, p. 253-266, 2011.

ARAÚJO, P. J. P.; PETTER, M. *O português e o quimbundo do Libolo (Angola)*: línguas em contato (Manuscrito).

ARAÚJO, P. J. P.; PETTER, M.; JOSÉ, J. A. Variedades de português angolano e línguas bantas em contato. *In*: OLIVEIRA, M. S. D. de; ARAÚJO, G. A. de (org.). *O português na África Atlântica*: Angola, Cabo Verde, Guiné-Bissau, São Tomé e Príncipe. São Paulo: Humanitas, 2018. p. 17-46.

ARSÉRNIO, M. J.; SEBASTIÃO, J. J. C.; ADÃO, A. *Manual de Alfabetização em Kimbundu*. Luanda: África Internacional, 2012.

ASSIS JR., A. *Dicionário Kimbundu-Português*. Luanda: Argente, Santos e Comp., [s./d.].

BAIÃO, D. V. *Quimbundo sem mestre*: gramática popular da língua Kimbundu conforme é falada nos distritos de Luanda e Malange/O Kimbundu prático ou guia de conversação em Português-Kimbundu. Porto: Imprensa Moderna, 1946.

BARRIÈRE, C. *Natural Language Understanding in a Semantic Web Context*. Cham, Switzerland: Springer, 2016. DOI: 10.1007/978-3-319-41337-2

BATALHA, L. *A Lingua de Angola*. Lisboa: Companhia Nacional Editora, 1891.

BONVINI, E. Revisiter, trois siècles après, Arte da língua de Angola de Pedro Dias S. I. (1697), première grammaire du kimbundu. *In*: PETTER, M.; BELINE, R. (org.). *Proceedings of the Special World Congress of African Linguistics, São Paulo, 2008:* Exploring the African Language Conection in the Americas. São Paulo: Humanitas, 2009. p. 15-45.

CANNECATTIM, B. M. de. *Diccionario da língua bunda, ou angolense explicada na portuguesa, e latina*. Lisboa: Impressão Régia, 1804.

CANNECATTIM, B. M. de. *Collecção de observações grammaticaes sobre a língua bunda, ou angolense*. Lisboa: Impressão Régia, 1805.

CHATELAIN, H.. *Folk tales of Angola/Contos populares de Angola.* Nova York/Lisboa: The American Folk-Lore Society/Agênia Geral do Ultramar, 1894/1964.

CHATELAIN, H. *Grammatica elementar do kimbundu ou língua de Angola/Kimbundu grammar*. Genève: Type de Charles Schuchardt, 1888/1889.

DEUCHAR, M.; DAVIES, P.; HERRING, J. R.; PARAFITA COUTO, C.; CARTER, D. Building Bilingual Corpora. *In*: THOMAS, E. M.; MENNEN, I. (org.). *Advances in the Study of Bilingualism*. Bristol: Multilingual Matters, 2014. p. 93-110. DOI: https://doi.org/10.21832/9781783091713-008

DIAS, P. *Arte da língua de Angola, oferecida a Virgem Senhora N. do Rosario, Mãy, e Senhora dos mesmos Pretos, pelo P. Dias da Companhia de Jesu.* Lisboa*:* Officina de Miguel Deslandes, 1697. Edição fac-similar da Fundação Biblioteca Nacional. Rio de Janeiro: Fundação Biblioteca Nacional, 2006.

ESIJAME, A.; GUT, U.; ANTIA, B. (org.). *Corpus Linguistics and African Englishes*. Amsterdam: John Benjamins, 2019. DOI: https://doi.org/10.1075/scl.88

FERNANDES, G. Primeiras descrições das línguas africanas em língua portuguesa. *Confluência*: *Revista do Instituto de Língua Portuguesa*, Rio de Janeiro, n. 49, p. 43-67, 2015. DOI: https://doi.org/10.18364/rc.v1i49.88

FIGUEIREDO, C. F. G. Aspectos histórico-culturais e sociolinguísticos do Libolo: aproximações com o Brasil. *In*: OLIVEIRA, M. S. D. de ARAÚJO, G. A. de. (org.). *O português na África Atlântica*: Angola, Cabo Verde, Guiné-Bissau, São Tomé e Príncipe. São Paulo: Humanitas, 2018. p. 47-97.

FIGUEIREDO, C. F. G. *Retratos do Libolo*. Lisboa: Chiado Editora, 2016. v. 2.

FIGUEIREDO, C. F. G.; NEGRÃO, E. V.; OLIVEIRA, M. S. D.; PETTER, M. Autorização de recolha e de apresentação de dados e imagens. *In*: FIGUEIREDO, C. F. G.; OLIVEIRA, M. S. D. de. (org.). *Linguística, História, Antropologia e Ensino no Kwanza Sul, Angola*. Lisboa: Chiado Editora, 2016. p. 21-22.

FIGUEIREDO, C. F. G.; OLIVEIRA, M. S. D. de. (org.). *Linguística, História, Antropologia e Ensino no Kwanza Sul, Angola*. Lisboa: Chiado Editora, 2016. v. 1.

FIGUEIREDO, C.; OLIVEIRA, M. S. D. de. Português do Libolo, Angola, e português afro-indígena de Jurussaca, Brasil: cotejando os sistemas de pronominalização. *Papia*, São Paulo, v. 23, n. 2, p. 105-185, 2013.

GALVES, C. O *corpus* Tycho Brahe: um *corpus* sintaticamente anotado do português histórico. *Revista Binacional Brasil Argentina: Diálogo entre as Ciências*, Vitória da Conquista, BA, v. 8, p. 181-204, 2019. DOI: https://doi.org/10.22481/rbba.v8i1.5585

GALVES, C. The Tycho Brahe Corpus of Historical Portuguese: Methodology and Results. *Linguistic Variation*, Amsterdam; Philadelphia, v. 18, n. 1, p. 49-73, 2018. DOI: https://doi.org/10.1075/lv.00004.gal

GALVES, C.; SANDALO, F.; SENA, T.; VERONESI, L. Annotating a Polysynthetic Language: From Portuguese do Kadiwéu. *Cadernos de Estudos Linguísticos*, Campinas, v. 59, n. 3, p. 361-648, 2017. DOI: https://doi.org/10.20396/cel.v59i3.8651003

GREENBERG, J. *The Languages of Africa*. Bloomington: Mouton de Gruyter, 1963.

GÜLDEMANN, T. (org.). *The Languages and Linguistics of Africa*. Berlim: de Gruyter, 2018. DOI: https://doi.org/10.1515/9783110421668

HAMMARSTRÖM, H. An Inventory of Bantu Languages. *In*: VELDE, M. van de; BOSTOEN, K.; NURSE, D.; PHILIPSON, G. (org.). *The Bantu Languages*. Londres: Routledge, 2019. p. 17-78. DOI: https://doi.org/10.4324/9781315755946-2

HUTH, K. *Untersuchungen zum nominalklassensystem des kimbundu (vr Angola) unter Berücksishtigung der entwicklungstendenzen siener urbanen varianten*. 1984. 169 f. Tese (Doutorado em Linguística) – Karl-Marx-Universität, Leipzig, 1984.

LÓPEZ, L. Á.; GONÇALVES, P.; AVELAR, J. O. de. (org.). *The Portuguese Continuum in Africa and Brazil*. Amsterdam: John Benjamins, 2018.

LUCCHESI, D.; BAXTER, A.; RIBEIRO, I. (org.). *O português afro-brasileiro*. Salvador: Edufba, 2009. DOI: https://doi.org/10.7476/9788523208752

MAIA, A. da S. *Dicionário Complementar Português-Kimbundu-Kikongo* (Linguas Nativas do Centro e Norte de Angola). Luanda: Cooperação Portuguesa, 1961.

MAIA, A. da S. *Lições de gramática de quimbundo:* português e banto (Dialecto Omumbuim). Cucujães: Escola Tipográfica das Missões, 1957.

MAIA, A. da S. *Guia prático para a aprendizagem das línguas Portuguesa e Omumbuim* (Língua indígena de Gabela-Amboim-Quanza-Sul-Angola) - Dialecto do Kimbundo. Cucujães: Escola Tipográfica das Missões, 1951.

MATTA, J. D. C. da. *Ensaio de Diccionario Kimbúndu-Portuguez*. Lisboa: Casa Editora António Maria Pereira, 1893.

MIGUEL, A. J. *Integração morfológica e fonológica de empréstimos lexicais bantos no Português Oral de Luanda*. 2019. 401f. Tese (Doutorado em Linguística) - Universidade de Lisboa, 2019.

MINGAS, A. A. *Interferência do Kimbundu no português falado em Luanda*. Lisboa: Campo das Letras, 2000.

MYERS-SCOTTON, C. *Language Contact*: Bilingual Encounters and Grammatical Outcomes. Oxford: Oxford University Press, 2002.

NURSE, D.; PHILIPSON, G. (org). *The Bantu Languages*. Londres: Routledge, 2003.

OLIVEIRA, M. S. D. de; ARAÚJO, G. A. de (org.). *O português na África Atlântica*: Angola, Cabo Verde, Guiné-Bissau, São Tomé e Príncipe. São Paulo: Humanitas, 2018.

OLIVEIRA, M. S. D. de; ZANOLI, M. de L.; ANDRADE, G. M. Marcadores discursivos no português falado em Angola, subvariedade Libolo: um estudo inicial de base prosódico-pragmática. *Filologia e Linguística Portuguesa*, São Paulo, v. 20, n. Especial, p. 159-186, 2018. DOI: https://doi.org/10.11606/issn.2176-9419.v20iEspecialp159-186

PACCONIO, F. *Gentio de Angola sufficientemente instruido nos mysterios de nossa sancta Fé.* Lisboa: Lopes Rosa, 1642.

PAIXÃO DE SOUZA, M. C.; KEPLER, F. N.; FARIA, P. P. F. de. E-Dictor: novas perspectivas na codificação e edição de corpora de textos históricos. *In*: PINTO, M. V.; SHEPERD, T.; SARDINHA, T. B. (org.). *Caminhos da Linguística de Corpus*. Campinas: Mercado de Letras, 2012. p. 191-224.

PEDRO, J. D. Étude grammaticale du kimbundu (Angola). 1993. 380 f. Tese (Doutorado em Linguística) – Universidade René Descartes, Paris, 1993.

PETTER, M. A classificação das línguas africanas. *In*: PETTER, M. (org.). *Introdução à Linguística Africana*. São Paulo: Contexto, 2015. p. 49-86.

PETTER, M. *Variedades lingüísticas em contato*: português angolano, português brasileiro e português moçambicano. 2008. 212 f. Tese (Livre-Docência) - Universidade de São Paulo, 2008.

PETTER, M.; ARAÚJO, P. J. Linguística Africana: passado e presente. *In*: PETTER, M. (org.). *Introdução à Linguística Africana*. São Paulo: Contexto, 2015. p. 27-48.

QUINTÃO, J. L. *Gramática de Kimbundu*. Luanda: Edições Descobrimentos, 1934.

ROCHA, B.; MELLO, H.; RASO, T. Para a compilação do C-ORAL-ANGOLA: um corpus de fala espontânea informal do português angolano. *Filologia e Linguística Portuguesa*, São Paulo, v. 20, n. Especial, p. 139-157, 2018. DOI: https://doi.org/10.11606/issn.2176-9419.v20iEspecialp139-157

ROSA, M. C. *Uma Língua Africana no Brasil Colônia de Seiscentos:* o quimbundo ou língua de Angola na Arte de Pedro Dias. S. J. Rio de Janeiro: Faperj; 7 Letras, 2013.

ROUX, J. C.; NDINGA-KOUMBA-BINZA, S. African Languages and Human Language Technologies. *In*: WOLFF, E. (org.). *The Cambridge Handbook of African Linguistics*. Cambridge: Cambridge University Press, 2019. p. 623-649. DOI: https://doi.org/10.1017/9781108283991.022

SANTOS, E. F. dos. *Sentenças marcadas para o foco no português do Libolo*: uma proposta de análise derivacional. 2015. 157f. Tese (Doutorado em Filologia e Língua Portuguesa) - Universidade de São Paulo, São Paulo, 2015.

SARDINHA, T. B. *Lingüística de Corpus*. Barueri: Manole, 2004.

SOUSA, M. de F. L. de; KUKANDA, V.; SANTIAGO, J. L. A posição lexical do Songo dentro do grupo H20 (Kimbundu *strictu sensu*, Sama, Bolo, Songo). *Papia*, São Paulo, v. 21, n. 2, p. 303-314, 2011.

VANSINA, J. Portuguese vs Kimbundu: Language Use in the Colony of Angola (1575-c. 1845). *Bulletin des Seances de l'Academie des Sciences d'Outre-Mer*, Bruxela, Bélgica, v. 47, n. 3, p. 267-281, 2001.

VELDE, M. van de; BOSTOEN, K.; NURSE, D.; PHILIPSON, G. (org). *The Bantu Languages*. Londres: Routledge, 2019.

VIEIRA-MARTINEZ, C. E. *Building Kimbundu*: language community reconsidered in West Africa, c. 1500-1750. 2006. 264 f. Dissertação (Mestrado em História) - University of California, Los Angeles, 2006.

VOSSEN, R.; DIMMENDAAL, G. (org.). *The Oxford Handbook of African Languages*. Oxford: Oxford University Press, 2020. DOI: https://doi.org/10.1093/oxfordhb/9780199609895.001.0001

WEINREICH, U. *Languages in Contact*. Findings and Problems. Nova York: Linguistic Circle of New York; De Gruyter, 1953.

WENDLING, V. *Catecismo da Doutrina Cristã em Portuguez com uma versão em Kimbundo, Dialeto do Libolo*. Lisboa: Província Portuguesa da Congregação do Espírito Santo, 1922. (Manuscrito.)

WOLFF, E. (org.). *The Cambridge Handbook of African Linguistics*. Cambridge: Cambridge University Press, 2019. DOI: https://doi.org/10.1017/9781108283991

XAVIER, F. da S. *Fonologia Segmental e Supra-Segmental do Quimbundo* - Variedades de Luanda, Bengo, Quanza Norte e Malange. 2010. 158f. Tese (Doutorado em Linguística) - Universidade de São Paulo, São Paulo, 2010.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 805-828, 2021

Brazilian Sign Language *corpus*: Acre Libras Inventory

Corpus da Língua Brasileira de Sinais: inventário de Libras do Acre

Ronice Müller de Quadros
Universidade Federal de Santa Catarina (UFSC), Florianópolis, Santa Catarina / Brasil
ronice.quadros@ufsc.br
http://orcid.org/0000-0002-5152-8716

Alexandre Melo de Sousa
Universidade Federal do Acre (UFAC), Rio Branco, Acre / Brasil
alexlinguista@gmail.com
http://orcid.org/0000-0002-2510-1786

Abstract: This paper draws on the theoretical methodological proposal of a Brazilian Sign Language (Libras) *corpus* to be developed under the scope of the Brazilian Sign Language (Libras) Inventory in the region of Rio Branco municipality, in the State of Acre project. First, we address some issues regarding *corpus* definitions and characteristics, some aspects of Libras, and documentation of sign languages. Second, we address the methodology used in gathering, transcription and analysis of data from Brazilian Sign Language Inventory focusing on the Region of Rio Branco – Acre, shedding light on the contributions of the gathered data to identification, recognition, valuing, and documentation of the Brazilian Sign Language in use in the State of Acre.

Keywords: Inventory; Brazilian Sign Language (Libras); Rio Branco; Acre.

Resumo: Este artigo se baseia na proposta teórico-metodológica de um *corpus* de Língua Brasileira de Sinais (Libras) a ser desenvolvido no âmbito do projeto Inventário de Língua Brasileira de Sinais na Região do município de Rio Branco, no estado do Acre. Em primeiro lugar, abordamos algumas questões relativas às definições e às características do *corpus*, alguns aspectos da Libras e documentação das línguas de sinais. Em segundo lugar, abordamos a metodologia utilizada na coleta, transcrição e análise de dados do Inventário Brasileiro de Língua de Sinais com foco na Região de Rio Branco – Acre, destacando as contribuições dos dados coletados para identificação, reconhecimento, valorização e documentação da Língua Brasileira de Sinais em uso no Estado do Acre.

Palavras-chave: Inventário; Língua Brasileira de Sinais (Libras); Rio Branco; Acre.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.805-828

Submitted on September 8th, 2020
Accepted on December 21th, 2020

1 Introduction

The proposal of building an inventory of Brazilian Sign Language in the region of Rio Branco in the State of Acre is integrated to the National Brazilian Sign Language Inventory (INDLibras), established by Universidade Federal de Santa Catarina, as part of the National Inventory of Linguistic Diversity (INDL), implemented by the decree 7387/10, as a tool for identification, recognition, valuing and promotion of the languages spoken in Brazil. In this sense, INDL stands as an instrument of the National Program of Immaterial Patrimony (IPHAN), which aims at embracing the semiotic, sociocultural, political specificities of the languages spoken in Brazil, in contrast to the cultural references encompassed by IPHAN, namely the Registration and the National Inventory of Cultural References (INRC) (IPHAN, 2016, p. 1). The present paper follows the proposal of methodological description in the compilation of the Brazilian Sign Language (Libras) Inventory as observed in the works of Quadros (2016a) with regard to the inventory of Florianópolis region (headquarters of the original project), and Ludwig *et al* (2019), regarding the inventory of Palmas region – in the State of Tocantins.

INDL, as a whole, might be defined as follows: a) a set of information about the languages spoken in Brazil; b) a way to support language knowledge and heritage; c) a policy catalyzing resources as well as governmental and non-governmental actions in order to protect those languages (IPHAN, 2016).

Once Libras is a national language, legally recognized by means of Law 10.436/2002 and regulated by Decree 5.626/2005, the development of a Libras Inventory leaves room for compilation of a *corpus* with information about the language and mapping of its linguistic aspects. Furthermore, once a consistent and broad inventory is created, one is likely to provide a Libras dataset for linguistic investigation, cultural valuing, educational feeding, and recognition of deaf identity.

2 The concept of *corpus*

Corpus compilation has been a reliable resource in linguistic research. One may define *corpus* from two main perspectives as follows: a Linguistic perspective and a *Corpus* Linguistics perspective. In this section, we will address such perspectives alongside the gathering of a sign language *corpus* in particular.

2.1 Corpus in Linguistics

Galisson and Coste (1983), for example, define *corpus* as a finite set of utterances comprising a type of language and taken as object of description, analysis, and, sometimes, creation of an explanatory model of such language. It might be comprised of oral documents (either recorded or transcribed), written documents or in both formats (depending on the nature of the research carried out) whose dimension is determined by the objectives and/or the phenomena under investigation.. If all the utterances are used, the *corpus* may be classified as exhaustive; however, if they are partially used, the *corpus* might be regarded as selective.

In their turn, Dubois *et al.* (1993) define *corpus* as a set of utterances that are the foundation for a descriptive grammar of a given language, despite being a sample of it and, therefore, being a representative of the structural characteristics of the language. The scholars claim that:

> One might think that difficulties may arise if a *corpus* is exhaustive […]. Indeed, once the number of possible utterances might not be defined, there seems to be no true exhaustivity and, besides this, a significant amount of useless data may only make the research complicated, making it heavy. Thus, the linguist should aim at obtaining a truly significant *corpus*. The linguist must take everything that may turn their *corpus* into a non-representative one (i.e., research method chosen, anomaly constituting linguist intrusion, prejudice against language) with a pinch of salt (DUBOIS *et al.*, 1993, p. 158-159).[1]

[1] "Poder-se-ia pensar que as dificuldades serão levantadas se um corpus for exaustivo [...]. Na realidade, sendo indefinido o número de enunciados possíveis, não há exaustividade verdadeira e, além disso, grandes quantidades de dados inúteis só podem complicar a pesquisa, tornando-a pesada. O linguista deve, pois, procurar obter um corpus realmente significativo. Enfim, o linguista deve desconfiar de tudo o que pode tornar o seu corpus não-representativo (método de pesquisa escolhido, anomalia que constitui a intrusão de linguista, preconceito sobre a língua)" (DUBOIS *et al*., 1993, p. 158-159).

Fromm (2003) concludes that a *corpus*, from a Linguistic perspective, constitutes a set of texts, either from similar or different areas, which has a specific investigative objective. Nevertheless, this group of texts differs from "a collection (of excerpts from literature works) or from an anthology (a collection of texts of renowned authors), which puts together works or scattered parts of works with didactic or purely commercial purposes" (FROMM, 2013, p. 1).

2.2 *Corpus* in *corpus* linguistics

Corpus Linguistics is the area in Linguistics focused on *corpora* studies and compilation. In Berber Sardinha's (2004) words, this field of research is responsible for gathering and exploring *corpora* (or sets of textual linguistic data accurately collected) for research in language variation That is to say, "it focuses on exploration of language by means of empirical evidence extracted by computer" (BERBER SARDINHA, 2004, p. 3).

In a previous work, Baker (1995, p. 229) addresses some criteria related to the inner workings of a *corpus*, as follows:

> *Corpora* are generally designed on the basis of a number of selection criteria, the most important of which are: (i) general language vs. restricted domain (ii) written vs. spoken language (iii) synchronic vs. diachronic (iv) typicality in terms of range of sources (writers/speakers) and genres (e.g., newspaper editorials, radio interviews, fiction, journal articles, court hearings) (v) geographical limits, e.g., British vs. American English (vi) monolingual vs. bilingual or multilingual.

From a *Corpus* Linguistics perspective, Sinclair (2005) defines *corpus* as follows:

> [*corpus*] is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variation as a source of data for linguistic research (SINCLAIR, 2005, p. 16).

Therefore, a *corpus* is a set of authentic computer-readable linguistic data that are representative of a given language (or language variation), accurately gathered (BERBER SARDINHA, 2004; TAGNIN; TEIXEIRA, 2004).

McEnery and Wilson (1996) explain that the notion of *corpus* is comprised of four main pillars as shown in Chart 1:

CHART 1 – Characteristics of a *corpus*

Characteristic	Description
Sampling and representativeness	A *corpus* must be comprised of sufficient sampling of a language or language variation to be analyzed in order to obtain maximum representativity of such language or language variation .
Finite size	A *corpus* must have a finite length, e.g., 500,000 words, 1 million words, 10 million words – except for corpus-monitor 1.
Machine-readable form	A *corpus* must be comprised of digital texts, which offer the following benefits: i) the *corpora* could be researched and manipulated quickly; ii) the *corpora* could be easily fueled with additional information.
Standard reference	A *corpus* constitutes a standard reference for the variation of language that it represents and it must be available for other researchers'(re)use.

Source: The authors – based on McEnery and Wilson (1996).

With regard to the first characteristic pointed out by McEnery and Wilson (1996) – s*ampling and representativeness* – Sinclair (2005) highlights the importance of making choices that lead to the *corpus* mirroring the linguistic behavior of the community whose language is analyzed.

When it comes to the second characteristic – *finite size* – Kennedy (1998) considers that the *corpus* length might take into account not only tokens amount, but also the quantity and diversity of categories to be analyzed, according to the type of research carried out.

Regarding *machine-readable form* (MCENERY; WILSON; 1996), the analysis of digital *corpus* leaves room for accurate remarks, providing reliable and objective information about linguistic facts.

With regard to the last characteristic, it is worth noting that the fact that the *corpus* compiled should be available for future studies (and the fact that the *corpus* may be a standard reference in the language or in its variation) is one of the main characteristics of a corpus from the *Corpus* Linguistic perspective – which relies on storage and exploration of data through computer tools.

2.3 A *corpus* of Sign Language

According to Berber Sardinha (2004, p. 20), when it comes to typologies, oral (transcribed) and written *corpora*, one may ask whether it is possible to compile a *corpus* when there is a visual-spatial language (such as Libras, American Sign Language (ASL), Portuguese Sign Language (LGP) and so forth) at stake.

In a study on the procedures underlying the compilation of a linguistic Libras *corpus*, Veras (2014) points out that the *corpus* investigation network in a given language is still comprised of written language data – which might pose a constraint to an analysis of a visual-spatial language the way it really is: a complex production of gestures, prosody, intonation, eye gaze, expressions, reference, body movements, among other elements that might not arise in written texts. The sign language material available is mostly shown in video. It is this format that leaves room for a more accurate analysis of linguistic phenomena in Libras, for instance.

Quadros (2019) points out the importance of compilation and availability of sign language *corpora*, taking into account a number of factors such as documentation, linguistic valuing, mapping of different language varieties, preservation of records, as well as availability of data for investigation of various linguistic phenomena. The author also sheds light on sign language *corpora* from a number of countries:

a) Libras *Corpus* (www.corpuslibras.ufsc.br);
b) Australian Sign Language (AUSLAN) *Corpus* (http://www.auslan.org.au/about/corpus);
c) British Sign Language (BSL) *Corpus* (http://www.bslcorpusproject.org/);
d) German Sign Language (DGS) *Corpus* (http://www.sign-lang.uni.hamgurg.de/dgs-korpus/index.php/welcome.html);
e) Dutch Sign Language (DSL) *Corpus* (http://www.ru.nl/corpusngtuk/);
f) Polish Sign Language (PSL) *Corpus* (http://www.plm.uw.edu.pl/en/node/241);
g) Japanese Sign Language (JSL) *Corpus* (http://research.nii.ac.jp/jsl-corpus/public/en/index.html).

By means of a *corpus*, be it in oral or sign languages, one may investigate phonetic-phonological, morphological, syntactic, semantic, textual-discursive aspects that could be relevant to linguistic research. When it comes to the constitution of a sign language *corpus* in particular, video platforms have turned out to be a reliable environment for the creation of a linguistic *corpus*. In particular, such platforms provide linguistic variation within the geographic scope of the data.

The proposal of a Libras National Inventory, as a way of compiling linguistic data from deaf individuals' (male and female) productions, from different age gaps, under a strict methodological process, leaves room for reference language sampling as proposed by the *corpora* compilation guidelines (LEITE; QUADROS, 2014; QUADROS, 2016a; QUADROS *et al.,* 2018, 2020, in press). It is worth noting that systematization of sign language documentation procedures, i.e., gathering, registration, storage, and recovery of data and metadata of sign languages worldwide has gained much attention in the literature over the last years, as discussed by Crasborn, van der Kooij and Mesch (2004); Efthimiou and Fotinea (2007); Hanke (2000); Leeson, Saeed and Byrne-Dunne (2006); Schembri (2008); and Chen Pichler *et al.* (2010), among others.

3 Libras and the Libras National Inventory

In the academic area, one may note that research on sign language has been developed only recently in comparison to studies on oral languages. The study of sign languages as natural languages itself had been questioned up to the 1960s so that constraints were posed to the development of Linguistics as a science and the Brazilian deaf community would meet up with challenges in terms of social and educational development.

In light of William Stokoe's (2005) seminal work, the 1960s stand as a reference for sign languages studies. In such work, the American linguist sheds light on the possibility of describing and analyzing sign languages, such as the ASL, based on the same theoretical methodological procedures adopted in the description and analysis of oral languages. In such a unique way, Stokoe showed that sign languages, similarly to oral languages, would also have the particularity of *articulation*, once the signs are formed by a limited number of minimal components that rearticulate to produce a limited number of signals, configuring a highly productive and saving set of contrasts.

From Stokoe's theory to current date, sign language studies has developed significantly, encompassing research that has heavily contributed to linguistic science in two ways: on the one hand, by demonstrating that specificities underlying natural languages are similarly present in sign languages – which have been progressively analyzed at their different levels of study (phonetic and prosodic, phonological, morphological and lexical, syntactic, semantic and pragmatic); on the other hand, by emphasizing similarities and differences in the way sign languages and oral languages are structured at various levels of analysis in order to contribute to a deeper debate over the linguistic theory and its applicability in society.

Unarguably, the academic and social demands for expertise in Brazilian Sign Language is high, notwithstanding the timid steps taken in such field of research. In a broad context, one may note, in Brazil, the same difficulty pervades the area of sign language studies in a global scale: there seems to be ample variation and uncertainty in the criteria regarding registration, documentation, analysis and display of linguistic data of sign languages for the academy (MILLER, 2001). In light of such circumstances, little room seems to be left for a rich empirical debate over the various linguistic aspects of sign languages as well as over the use of such knowledge in a number of applied domains, namely education of deaf community, Libras teaching as L1 and L2, Libras interpreter training, translation of literary works into Libras and so forth (see some developments in *Portal de Libras*).

However, regulamentation of gathering systems, documentation and recovery of data and metadata of sign languages have achieved higher importance worldwide over the last decade (cf. CHEN PICHLER *et al*., 2010; CRASBORN; VAN DER KOOIJ; MESCH, 2004 EFTHIMIOU; FOTINEA, 2007; HANKE, 2000; LEESON; SAEED; BYRNE-DUNNE, 2006;; SCHEMBRI, 2008). In that sense, Libras could not be in a different position. Thus, the development of Libras *corpora* and systematization of its creation process may contribute, in various forms, to consolidation of theory and practice with regard to sign language in the country.

Research on Libras started in late 1980s, with the seminal works of Ferreira-Brito (1984, 1990, 1995), Quadros (1995, 1999), and Karnopp (1994, 1999) among others (QUADROS, 2013, 2018). Other studies have been developed based on the possibility of observing Libras from different perspectives, ranging from its phonetic-phonological basis to the multiple perspectives related to discourse and, therefore, its relationship

with culture. With respect to the relationship between language and culture, Chacon *et al.* (2014, p. 2) claim that:

> [...] both languages and cultures are means and raw material for symbolic and identity frameworks of a given social group and its relationships with other groups; both are transmitted through learning and they are recognized as structured systems of symbols and norms. Language is the vehicle for culture dissemination, and it is also one of constituting elements of several aspects of culture; and vice-versa.[2]

Such intrinsic connections between language and culture, that subside actions to enhance and value linguistic policies, might leave room for recognition of linguistic diversity, linguistic variation (observed in all levels of the system as well as outside it), and recognition of every Brazilian language (including minority groups) as cultural reference, and, therefore, as immaterial patrimony – as stated in the National inventory of Linguistic Diversity (INDL) (QUADROS *et al.*, 2018, p. 11).

According to Chacon *et al.* (2014), Decree 7.387/2010 implemented INDL, which aimed at "establishing an instrument for identification, documentation, recognition and valuing of languages standing as referent for identity, action and memory of the different groups comprising the Brazilian society." Thus, INDL provides mapping and linguistic diversity patrimony policy, protecting the languages of specific linguistic communities in Brazil. In such case, the languages of Brazilian deaf communities are included, based on the viewpoint presented by Chacon *et al.* (2014, p. 4):

> [...] language serves to mark positions and social identities of collectivities and individuals, creating a symbolic and communicative fabric of a community; on the one hand, social practices create the various contexts of language use, marking both its symbolic and structural evolution and social norms and values.[3]

---

[2] "[...] tanto as línguas como as culturas são meios e a matéria para os referenciais simbólicos e identitários de um grupo social e suas relações com outros grupos; ambas são transmitidas através da aprendizagem, e são reconhecidas como sistemas estruturados de símbolos e normas. Língua é veículo para a transmissão da cultura e é também um dos elementos constituintes de vários aspectos da cultura; e vice-versa."

[3] "[...] a língua serve para demarcar posições e identidades sociais de coletividades e indivíduos, criando o tecido simbólico e comunicativo de uma comunidade; por um

This way, although linguistic studies focusing on Libras have expanded over recent years, they still lack greater empirical foundation, mainly when it comes to recording and manipulation of data (QUADROS *et al.*, 2018, p. 12). The initiative to compile a Libras *corpus* – comprised of the National Inventory of Brazilian Sign Language at *Universidade Federal de Santa Catarina* – has contributed significantly to fostering research on sign language and deafness in Brazil as well as to providing theoretical and empirical framework to Libras didactic material production and recording of life experiences from the Brazilian deaf community (LEITE; QUADROS, 2014, QUADROS, 2016a, QUADROS *et al.,* 2018, 2020, in press).

Such National Inventory has been applied to other states, entitled *Inventário de Libras de Alagoas*, *Inventário de Libras do Ceará* e *Inventário de Libras do Tocantins* (Ludwig et al, 2019) as the main reference for collection, recording and analyses (QUADROS *et al.,* in press).

4 Brazilian Sign Language Inventory in Rio Branco - Acre

The Brazilian Sign Language Inventory in the Region of Rio Branco, Acre aims at constituting a Libras *corpus* that is representative of the State of Acre as well as at enhancing social, intellectual and cultural reflection by deaf audience in the State of Acre by means of individuals' engagement alongside valuing of deaf language and culture. In addition to such main goal, the Inventory also aims at providing both a large empirical *corpus* of Libras – relying on theoretical and methodological bases as well as at representing a Libras *Corpus* from the region of Rio Branco, in the State of Acre – and open access to researchers and professionals who are involved with deaf community and would use it for theoretical and applied linguistics.

Furthermore, the inventory aims at providing guidelines for the constitution of a *corpus* of Libras for future research, mainly when it comes to recording, documentation and recovery of data for linguistic analysis purposes. It is important that the current technological

---

lado, as práticas sociais criam os contextos diversos de usos de uma língua, marcando a sua evolução tanto estrutural e simbólica, quanto com relação a normas e valores da sociedade."

possibilities be spread in the academic area in order to provide a consistent empirical basis for studies on Libras as well as develop linguistic, historical and cultural records of the lifestyle of the deaf community, leading to their inclusion in Brazilian society. Also, by following a consistent methodological approach over the whole country, the National Libras Inventory brings comparable data to allow identification of Libras variation. The methodological proposal described in the Inventory of the Region of Rio Branco – in the State of Acre is in accordance with the original Project (QUADROS, 2016a), which has also been adopted in the inventory of the region of Palmas – in the State of Tocantins (LUDWIG *et al*, 2019).

4.1 The informants

One of the main criteria for composing sign language *corpora* is the deaf participation in different stages of the process. A decisive methodological issue for such participation is inviting deaf individuals actively engaged in local deaf communities in their respective cities, once it is known that signers might not use their vernacular Language in the presence of unfamiliar interlocutors.

Nowadays, mainly due to deliberations from Decree 5,626, which determines Libras instruction for teaching, special education, and speech therapy[4] undergraduate students, there is a considerable number of deaf professors working as researchers at various state and federal universities in Brazil. In light of such fact, those deaf researchers stand as ideal contributors for the project, given the national dimension it might reach in the future alongside the possibility of using university infrastructure for data gathering. Indeed, one of the most significant contributions of *corpora* of natural languages is their applicability in procedures in language teaching and learning. In such matter, the professors at those universities may either enhance their research skills or use the *corpus* in their Libras classes for teaching purposes. In cities where there are no deaf professors in their academic institutions, associations, federations, or other institutions engaged in deaf community might stand as alternatives with the help of local deaf representatives.

[4] According to Decree 5,626, the course in Brazilian Sign Language – Libras is not mandatory for other undergraduate courses.

This Libras Inventory, based in the city of Rio Branco, in the *Letras Libras* undergraduate program (a Libras program) at the Federal University of Acre (UFAC), employs the same methodological procedures adopted in the National Inventory of Libras, taking into account the fact that the original project may encompass the 27 capital cities of Brazil (QUADROS *et al.,* 2018, 2020, in press).

The reason for choosing Rio Branco for gathering, recording and transcribing data is due to the fact that the UFAC *Letras Libras* course is based in the capital of the state, Rio Branco, and it is an education center with the highest number of deaf individuals fluent in sign languages.

The group of informants is comprised of 36 deaf individuals from the city of Rio Branco, who participate, in pairs, in 18 interviews, totaling around 40 hours of video (multiplied by 4 video perspectives taken by 4 different cameras). Not only the selection of participants, but also data collection itself are to be carried out by a local deaf researcher – who is supposed to meet the following requirements: i) be from Rio Branco or have had contact with the local deaf community for at least 10 years; ii) be an extrovert, preferably with academic experience in undergraduate or graduate courses; iii) have experience in technologies that are fundamental for the project objectives as well as have daily access to a computer and the Internet.

The informants must meet the following requirements: i) be from Acre or have lived in Acre for at least 10 years; ii) have learned Brazilian Sign Language up to seven years old or with evident proficiency in the community; iii) both individuals in the pair interviewed must be close to each other (i.e., being friends or relatives) and, preferably, having the same gender and being at the same age. It is worth noting that the local researcher should choose pairs from all walks of life. In order to do so, the following criteria may be taken into account: iv) the deaf individuals selected might be from three different age groups, including individuals aged up to 29 years, middle-aged individuals from 30 to 49 years, and individuals over 50 years; v) the deaf individuals selected must be either male or female; v) the deaf individuals selected may also have different education backgrounds. Only those informants consenting with all the use purposes of their image, without any restrictions, are to be selected as stated in the *Consent Form for Research Participation*.

Informants are to be selected for data collection from June 2021 on, in accordance with Ethics Committee's approval (approved by CAAE 35002620.9.0000.5010).

4.2 Ethical issues

The establishment of a Libras *Corpus*, also in the State of Acre, is a project that might only be carried out through active participation of the deaf community. The study starts from conversation with associations and other institutions from deaf community, aiming at clarifying the objectives and relevance of the present work to deaf education in Brazil. Furthermore, the coordination of the project shows clear interest in understanding the preference of deaf community for certain text types or issues to be documented as well as for the forms of methods for data collection and for their expectations when it comes to the contribution offered to the Brazilian deaf community.

As mentioned above, the participation of informants in this project relies on their full consent alongside filling out the Consent Form for Research Participation. The research goals are clearly stated in the form and special focus is placed on the social and academic relevance of research on Brazilian Sign Language and, consequently, enhancing social inclusion of the deaf as well as making the informant understand the implications of allowing the use of their images for research purposes, teaching material and its availability on the Internet.

The general information will be collected through a sheet with questions on their language, family and educational background. This form is bilingual, presented in Portuguese and Libras, so that deaf informants are fully informed about the importance and the implications of their participation in the project in accordance with Resolution 510/2016 of the *National Health Council*.

4.3 Data collection

This item is based on Quadros (2016a, p. 162-167). The video recordings occur in a studio prepared at the University of Acre, in the *Letras Libras* department. The team in charge of data collection is comprised of a volunteer researcher from the coordination team and a technician. The volunteer researcher is responsible for conducting the entire interview. In turn, the technician is in charge of assembling the

itinerant studio and of offering technical guidance over the recording process and storage there.

The studio has four cameras in order to ensure that informants are captured in different perspectives – which is important for an accurate diagnosis of manual and non-manual articulators in conversational circumstances (LEITE, 2008). Each informant has access to a laptop providing the visual stimuli that are the basis for their production and the researcher has a third laptop to manipulate the stimuli as well as to record useful information in recording sessions.

Furthermore, there are lampposts and walls painted in different shades of blue serving as background for the recordings in order to provide optimal conditions for perspective visualization.

The four cameras are placed according to space settings that are previously adjusted and tested. It is worth noting that such settings may vary, as their disposal depends on the activity that is being recorded. For instance, individual eliciting and free conversation require different camera placements. In order to do so, a close-up on the informants' faces is necessary, as well as a take on the signaling of both informants, and a take above them, which is done by means of a camera placed on the ceiling of the studio, as shown in Figure 1.

FIGURE 1 – Camera takes

Sources: Quadros (2016a, p. 167); Quadros *et al.* (2018, p. 27).

Each interview with a pair of informants lasts, on average, two hours and it is carried out by means of the following tasks:

i) ice-breaking task and interview about the informants' personal information to be carried out in 30 minutes: in a semi-structured and partially-open interview, the researcher aims at eliciting, from informants, personal information on a wide array of topics, as follows: the story of their signal, their experience in Brazilian Sign Language acquisition and participation in the local deaf community; their contact with Portuguese and Libras with regard to use and attitudes; remarkable experiences; personal and professional goals;[5]

ii) the task involving eliciting of storytelling, to be performed in 20 to 30 minutes: the informant is supposed to tell three stories (*Pear Story,* a *Frog: where are you?* and *Canary Row*, by Tweetye Sylvester) previously used in the literature; for such reason, they might be used in studies comparing oral languages and sign languages.

iii) 20-minute rest intervals;

iv) eliciting tasks of both grammatical and lexical nature to be performed within 30 minutes: informants receive stimuli adapted from the German Sign Language *corpus* project (NISHIO *et al*., 2010) that are intended to elicit grammar constructions alongside lexical items in Brazilian Sign Language;

v) conversation within 20-30 minutes: each pair of informants is left on their own in the studio and is encouraged do talk about any random topic or about a current issue suggested by the researcher.

Finally, the interviews are carried out in a way that fosters the recording of verbal expressions underlying informants' culture based on demonstration of words, of linguistic borrowing, as well as of utterances illustrating elements concerning grammar, vernacular dialectal varieties pervading the cultural background of each region and in a universal way.

[5] This is the only free-access interview session made available on the corpus website. The other interviews have limited access and are made available in case of previous request and registration of external researchers.

In the case of the region of Rio Branco, in accordance with the Ethics Committee's approval (CAAE nº 35002620.9.0000.5010), the data are to be collected from the second semester of 2021 on. It is worth noting that the data collection procedures are similarly adopted by the *Palmas – TO Inventory* compilation project (LUDWIG *et al*, 2019) as well as the *Maceió – AL Inventory* compilation Project.

4.4 Data annotation

This item is based on Quadros (2016a, p. 168-169). The annotation process is time-consuming and requires commitment – in sign language studies in particular, as there is not a standard writing system fully adapted to a computer. In light of such issue, as sign language research projects have pointed out, it is estimated that one hour of annotation might correspond to 1 minute of a recording.[6] Indeed, the project, for its three first years, expects 40-45 hours (2,400-2,700 minutes) of recordings, so that 2,400-2,700 working hours might be necessary to basic data annotation, not to mention another 2,400-2,700 hours for review of annotation and gloss translation into Portuguese. All of the above things considered alongside time constraints posed to the project, the annotation phase may encompass, over these three initial years, considerable display of part of the data collected for 10-12 hours. In order to achieve such aim, two scholarship holders, under supervision, will be in charge of transcription of data and design of a data transcription manual over 36 months.

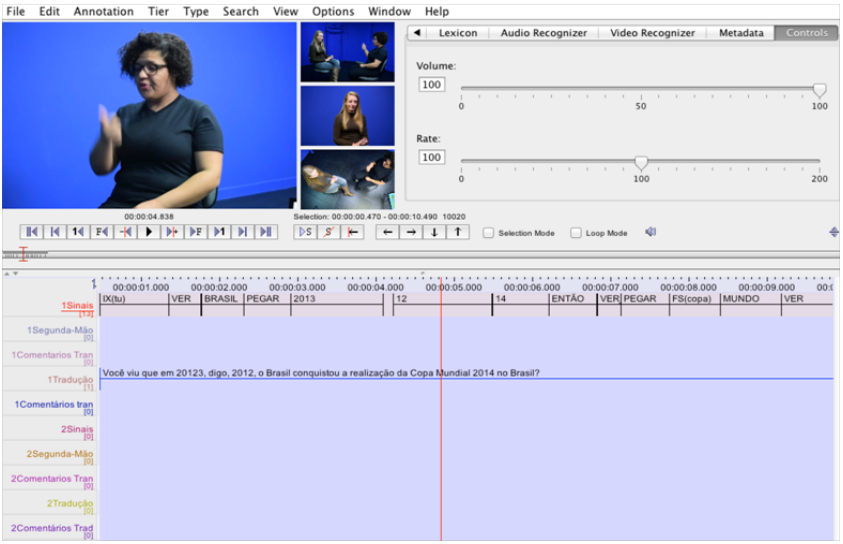
In this first phase, a special focus will be placed on the development of conventions and criteria for transcription based on the data samples that may define elements of the sign language inventory. Thus, transcription of the entire data of the *corpus* may walk hand in hand with necessary financial support for having undergraduate scholarship holders assuming these specific attributions. Although the details of annotation procedures might provide guidance to the study, they are not supposed to be addressed in detail in the first phase of the project.

Given the complexity underlying the process of annotating Libras (LEITE, 2008), the annotation work may be carried out in two basic ways: a) thorough glossing of manual signs, with a Sign Identification for right hand and for left hand; b) translation of utterances into Portuguese. The

[6] Available at: http://www.sign-lang.uni-hamburg.de/intersign/workshop4/baker/baker.html. Access on: Jun. 30, 2012.

program used for the Libras *corpus* data transcription is ELAN – software developed for audio and video purposes. It available for free download at *http://www.lat-mpi.eu/tools/*.

FIGURE 2 – Interview takes in ELAN



Source: Quadros (2016a, p. 168).

The annotation may follow a transcription template file for ELAN designed by the project coordination team (see more details in QUADROS, 2016b). The template will be shown to volunteer researchers during training. Even though the annotation template may account for all manual and non-manual articulators that are key to the description of Brazilian Sign Language (LEITE, 2008; CHEN PICHLER *et al*., 2010), the volunteer researchers are supposed to work solely on tracks regarding the two main issues mentioned in the previous paragraph. Therefore, the annotation of other articulators may be addressed in future studies. As ELAN only enables visualization of tracks of immediate interest, saving the other tracks, opting for transcribing the other articulators in the future seems to be a viable choice.

The validation process is vital to all transcriptions, which is assigned to the project members skilled at transcription, occurs based on statistically viable display of the data collected in other states and aims at drawing a comparison with original transcriptions. Such process occurs

frequently with the aims of evaluating and adjusting the transcription process when necessary. Hence, a researcher heads up the review of the original transcription in order identify potential inconsistencies in annotation conventions used in the project (QUADROS *et al.,* in press).

4.5 Data organization and availability online

The *corpus* data collected is to be stored in three ways: (i) on a specific server for the Libras *corpus*, based in the Data Processing Center of the Federal University of Santa Catarina (UFSC); (ii) on an external HD stored by the project coordinator; (iii) on a backup hard disk in the Multimedia Lab based in the Languages-Libras department of UFAC.

The data will be organized in accordance with a hierarchical structure, namely the capital where collection occurred, and denomination of pair participating in the interview. In such last type of folder, two subfolders could be created as follows: "raw data" – in which data gathered directly from collection are stored; and "edited data" – in which files edited and configured to be used in ELAN are stored. Both subfolders would be subdivided into folders entitled "informant_1" and "informant_2", which would encompass the following folders: "data type" – specifying whether the file encompasses interview, storytelling, eliciting, or conversation; "specific text" (whenever necessary), i.e., *Pear Story*, when it is storytelling, or classifiers, when there is an eliciting session. Storage of the transcribed data may occur in the same folders where edited videos are stored – which may serve as basis for transcription (in accordance with NALS database in QUADROS *et. al.*, 2014). Thus, this might stand as a template framework for implementation of the project to be developed by the Libras National Inventory.

In order to provide the storage of both data and metadata of the project in a reliable database for free online access to the *corpus*, the database to be developed must be developed in accordance with the online version of the *corpus*. In order to avoid any possible constraints or adversities, conversation between website programmers and the executive board of the project stands as a key factor.

At this first stage, the infrastructure provided in the metropolitan region of Florianópolis, in the State of Santa Catarina, stands as a benchmark for the Libras *corpus* inventory in other capital cities in Brazil. Therefore, in the same vein, the "Brazilian Sign Language Inventory in the Region of Rio Branco – Acre" uses similar studio

configuration, collection procedures, as well as way of recovering, storing and transcribing data.

5 Final remarks

The creation of an Inventory of Brazilian Sign Language in the Region of Rio Branco – Acre holds significant importance as it encompasses not only linguistic components, but also sociocultural as well as political aspects of Libras in the deaf community from Acre, aligned to the National Libras Inventory. Then, Acre state becomes part of Libras *Corpus* together with Santa Catarina (Florianópolis area), described by Quadros (2016a); Alagoas (Maceió area); Ceará (Fortaleza area); and Tocantins (Palmas area) – regarding the latter, check description in Ludwig *et al.* (2019).

The *corpus* represents Libras in the metropolitan region of Rio Branco as it is comprised of video recordings of both elicited and spontaneous language use situations for research and other applied purposes, not to mention the fact that the *corpus* involves the creation of a set of guidelines for registration and storage of data and metadata regarding Libras use to be also used in other states in Brazil. The *corpus* also encompasses the creation of a form with gaps and standardized items for systematization of the final results of the study carried out with the Libras *Corpus* of the State of Acre.

All in all, the development of a Libras *corpus* in the scope of the Inventory of the Libras in Rio Branco – Acre alongside the systematization of its creation process might play a significant role in the consolidation of both theory and practice of sign language research in Brazil, once the linguistic data set, accurately gathered, are representative of the language and may be available to other researchers for future studies. The data are to be gathered from 2021 on.

Acknowledgements

This work was made possible partially by the resources of the National Council for Scientific and Technological Development - CNPQ (# 440337 / 2017-8), as well as partially by resources from the National Historical and Artistic Heritage Institute (IPHAN), of the Ministry of Culture, in partnership with the Institute of Linguistic Policies (IPOL). We are very grateful to translator Raquel Rossini Martins Cardoso.

Contribution of each author to the manuscript

The paper "Brazilian Sign Language corpus: Acre Libras Inventory" stems from the original project (Brazilian Sign Language (Libras) National Inventory) developed by Ronice Müller de Quadros and adapted by Alexandre Melo de Sousa for the region of Rio Branco – Acre State. The first part of the theoretical framework regarding the concept of corpus was written by the second author. The first author was responsible for the methodological outline of the study alongside the core description of the research. The text was both written and revised by both authors.

References

BAKER, M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target,* Amsterdam, v. 7, n. 2, p. 223-243, 1995. DOI: https://doi.org/10.1075/target.7.2.03bak

BERBER SARDINHA, T. *Linguística de corpus*. São Paulo: Manole, 2004.

CHACON, T. C. *et al*. *Guia de pesquisa e documentação para o INDL:* patrimônio cultural e diversidade linguística/pesquisa. Brasília: IPHAN, 2014.

CHEN PICHLER, D. *et al.* Conventions for sign and speech transcription of child bimodal bilingual *corpora* in ELAN. *Language, Interaction and Acquisition,* [*S.l.*], v. 1, n. 1, p. 11-40, 2010. DOI: https://doi.org.10.1075/lia.1.1.03che

CRASBORN, O.; VAN DER KOOIJ, E.; MESCH, J. European Cultural Heritage Online (ECHO): Publishing Sign Language Data on the Internet. *In:* CONFERENCE ON THEORETICAL ISSUES IN SIGN LANGUAGE RESEARCH, 8th., 2004, Barcelona. *Proceedings* […] Barcelona: ECHO, 2004. p. 33-37.

DUBOIS, J. *et al. Dicionário de Linguística.* São Paulo: Cultrix, 1993.

EFTHIMIOU, E.; FOTINEA, S. E. Creation and Annotation of a Greek Sign Language *corpus* for HCI. Universal Access in Human Computer Interaction: Coping with Diversity. *In:* INTERNATIONAL CONFERENCE ON UNIVERSAL ACCESS IN HUMAN-COMPUTER INTERACTIONS, 4th., 2007, Beijing. *Proceedings* […]. Beijing: ILSP, 2007. p. 657-666.

FERREIRA-BRITO, L. Epistemic, Alethic, and Deontic Modalities in a Brazilian Sign Language. *In:* FISHER, S. D.; SIPLE, P. (ed.). *Theoretical Issues in Sign Language Research.* Chicago: University of Chicago Press. 1990. p. 224-260.

FERREIRA-BRITO, L. *Por uma gramática de línguas de sinais*. Rio de Janeiro: Tempo Brasileiro/UFRJ, 1995.

FERREIRA-BRITO, L. Similarities and Differences in Two Sign Languages. *Sign Language Studies,* Silver Spring, v. 42, p. 45-46, 1984.

FROMM, G. O uso de *corpora* na análise linguística. *Revista Factus,* São Paulo, v. 1, n. 1, p. 69-76, 2003.

GALISSON, R.; COSTE, D. *Dicionário de didáctica das línguas*. Coimbra: Livraria Almedina, 1983.

HANKE, T. (ed.). *ViSiCAST Deliverable D5-1:* interface Definitions. Hamburg: University of Hamburg 2000. Available from: https://www.researchgate.net/publication/302152848_ViSiCAST_Deliverable_D5-1_Interface_Definitions . Access on: May. 14, 2016.

IPHAN. *Guia de pesquisa e documentação para o INDL*: patrimônio cultural e diversidade linguística. Brasília, DF: Instituto do Patrimônio Histórico e Artístico Nacional, 2016.

KARNOPP, L. B. *Aquisição do parâmetro configuração de mão dos sinais da LIBRAS:* estudo sobre quatro crianças surdas filhas de pais surdos. 1994. 180f. Dissertação (Mestrado em Letras) – Instituto de Letras e Artes, PUCRS, Porto Alegre, 1994.

KARNOPP, L. B. *Aquisição fonológica na Língua Brasileira de Sinais:* estudo longitudinal de uma criança surda. 1999. 273f. Tese (Doutorado em Letras) – Instituto de Letras e Artes, PUCRS, Porto Alegre, 1999.

KENNEDY, G. *An Introduction to Corpus Linguistics.* London; New York: Longman, 1998.

LEESON, L.; SAEED, J.; BYRNE-DUNNE, D. *Moving Heads and Moving Hands:* Developing a Digital *Corpus* of Irish Sign Language. The 'Signs of Ireland' *Corpus* Development Project. *In*: INFORMATION TECHNOLOGY AND TELECOMMUNICATIONS CONFERENCE, 6th, 2006, Chengdu, China. *Proceedings* […]. Chengdu, China: IEEE, 2006. Available from: https://www.researchgate.net/publication/277186734_

Moving_Heads_and_Moving_Hands_Developing_a_Digital_Corpus_of_Irish_Sign_Language. Access on: Mary. 13, 2019.

LEITE, T. A. *A segmentação da língua de sinais brasileira (Libras):* um estudo linguístico descritivo a partir da conversação espontânea entre surdos. 2008. 280f. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2008.

LEITE, T. de A.; QUADROS, R. M. de. Línguas de sinais do Brasil: reflexões sobre o seu estatuto de risco e a importância da documentação. *In*: QUADROS, R. M.; STUMPF, M. R.; LEITE, T. A. (org.). *Estudos da Língua de Sinais II.* Florianópolis: Editora Insular, 2014. p. 15-27.

LUDWIG, C. R. *et al*. Inventário da Língua Brasileira de Sinais da Região de Palmas – Tocantins: Metodologia de Coleta e Transcrição de Dados. *Porto das Letras*, Porto Nacional, TO, v. 5, n. 1, p. 59-74, 2019. Available from: https://sistemas.uft.edu.br/periodicos/index.php/portodasletras/article/view/6489/14835. Access on: Dec. 14, 2020.

McENERY, T.; WILSON, A. *Corpus Linguistics.* Edinburgh: Edinburgh University Press, 1996.

MILLER, C. Some Reflections on the Need for a Common Sign Notation. *Sign Language and Linguistics,* Amsterdam, 4, n. 1/2, p. 11-28, 2001. DOI: https://doi.org/10.1075/sll.4.12.04mil

NISHIO, R. *et al*. Elicitation Methods in the DGS (German Sign Language) Corpus Project. *In*: DREUW, P.; EFTHIMIOU, E.; HANKE, T.; JOHNSTON, T.; MARTÍNEZ RUIZ, G.; SCHEMBRI, A. (ed.) *Corpora and Sign Language Technologies*. 4th Workshop on the Representation and Processing of Sign Languages. Paris: ELRA, 2010. p. 178-185

QUADROS, R. M. *As categorias vazias pronominais:* uma análise alternativa com base na língua de sinais brasileira e reflexos no processo de aquisição. 1995. 141f. Dissertação (Mestrado em Letras) – Instituto de Letras e Artes, PUCRS, Porto Alegre, 1995.

QUADROS, R. M. *Phrase Structure of Brazilian Sign Language*. 1999. 279f. Tese (Doutorado em Linguística) – Instituto de Letras e Artes, PUCRS, Porto Alegre, 1999.

QUADROS, R. M. Contextualização dos estudos linguísticos sobre a Libras no Brasil. *In:* QUADROS, R. M.; STUMPF, M. R.; LEITE, T. A. (org.). *Estudos da Língua Brasileira de Sinais I.* Florianópolis: Editora Insular, 2013. p. 15-36.

QUADROS, R. M.; LILLO-MARTIN, D.; CHEN PICHLER, D. Methodological Considerations for the Development and Use of Sign Language Acquisition Corpora. *In:* RASO, T.; MELLO, H. (ed.). *Spoken Corpora and Linguistic Studies*. Amsterdam: John Benjamins, 2014. p. 84-102.

QUADROS, R. M. Documentação da Libras. *In:* SEMINÁRIO IBERO-AMERICANO DE DIVERSIDADE LINGUÍSTICA, 2014, Foz do Iguaçu. *Anais* [...] Brasília: IPHAN – Ministério da Cultura, 2016a. p. 157-174.

QUADROS, R. M. A transcrição de textos do *corpus* de Libras. *Revista Leitura,* Maceió, n. 57, v. 1, p. 8-34, 2016b.

QUADROS, R. M. *et al. Língua Brasileira de Sinais:* Patrimônio Linguístico Brasileiro. Florianópolis: Editora Garapuvu, 2018.

QUADROS, R. M. Tecnologia para o estabelecimento de documentação de língua de sinais. *In:* CORRÊA, Y.; CRUZ, C. R. (org.). *Língua Brasileira de Sinais e tecnologias digitais.* Porto Alegre: Penso, 2019. p. 1-25.

QUADROS, R. M. *et al.* Brazilian Sign Language Documentation. *In:* QUADROS, Ronice Müller de. *Brazilian Sign Language Studies.* De Gruyter Mouton: Ishara Press, 2020. p. 9-32.

QUADROS, R. M. *et al.* Inventário Nacional de Libras. *Revista Fórum Linguístico*, Florianópolis, in press.

SCHEMBRI, A. C. The British Sign Language *Corpus* Project: open access archives and the observer's paradox. *In:* THE CONSTRUCTION AND EXPLOITATION OF SIGN LANGUAGE *CORPORA* WORKSHOP, 3th., 2008, Marrackech. *Proceedings* […]. Marrackech: Research Gate, 2008. p. 165-169.

SINCLAIR, J. *Corpus* and Text: Basic Principles. *In*: WYNNE, M. (ed.). *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford: Oxbow Books, 2005. p. 1-16. Available from: http://icar.cnrs.fr/ecole_thematique/contaci/documents/Baude/wynne.pdf . Access on: Out. 30, 2016.

STOKOE, W. Sign Language Structure: an Outline of the Visual Communication Systems of the American Deaf. *Journal of Deaf Studies and Deaf Education,* Oxford, v. 10, n. 1, p. 3-37, 2005. DOI: https://doi.org/10.1093/deafed/eni001

TAGNIN, S. E. O.; TEIXEIRA, E. D. Lingüística de Corpus e Tradução Técnica - Relato da montagem de um corpus multivarietal de culinária. *Tradterm*, [*S.l.*], v. 10, p. 313-358, 2004. DOI: 10.11606/issn.2317-9511.tradterm.2004.47184. Disponível em: https://www.revistas.usp.br/tradterm/article/view/47184. Acesso em: 12 May. 2019..

VERAS, E. C. *Procedimentos metodológicos para a compilação de um corpus de língua de sinais a partir da rede:* reflexões com base em um corpus piloto de gêneros na plataforma *YouTube*. 2014. 183f. Dissertação (Mestrado em Linguística) – Centro de Comunicação e Expressão, Universidade Federal de Santa Catarina, Florianópolis, 2014.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 829-858, 2021

Using machine translator as a pedagogical resource in English for specific purposes courses in the academic context

O uso do tradutor automático como recurso pedagógico na aula de inglês para propósitos específicos no contexto acadêmico

Débora Borsatti
Universidade de Santa Cruz do Sul (UNISC), Santa Cruz do Sul, Rio Grande do Sul / Brasil
deborsatti@gmail.com
http://orcid.org/0000-0003-1486-0047

Adriana Blanco Riess
Universidade de Santa Cruz do Sul (UNISC), Santa Cruz do Sul, Rio Grande do Sul / Brasil
adrianariess@unisc.br
http://orcid.org/0000-0002-0228-6028

Abstract: This paper presents a proposal for pedagogical use of MT in English for Specific Purpose (ESP) courses, aiming at investigating the efficiency of this technology as a support for reading scientific texts in English as a FL. The theoretical approach is on ESP, reading and comprehension and a proposal to use MT in ESP courses, aiming to understand the processing of MT and how this knowledge can raise benefits on reading comprehension for academic purposes. In addition, we discussed corpus linguistics and its relation to language teaching as well as its role in MT. The analysis shows that, due to the hybrid system that utilizes the rule-based system and the corpus-based system, Google Translate produces relatively understandable and readable texts. Despite its evident limitations, the tool can provide linguistic awareness when pedagogically explored by ESP teachers in academic context.

Keywords: Machine Translation; pedagogical tool; reading; English for Specific Purposes.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.829-858

Resumo: Este artigo apresenta uma proposta de uso pedagógico de tradução pela Máquina (MT) em cursos de inglês para fins específicos (ESP), com o objetivo de investigar a eficiência dessa tecnologia como suporte para a leitura de textos científicos em inglês como L2/ LE. A abordagem teórica é sobre ESP, leitura e compreensão e uma proposta de uso de MT em cursos de ESP, com o objetivo de entender o processamento da MT e como esse conhecimento pode trazer benefícios na compreensão da leitura para fins acadêmicos. Também, discute-se a linguística de corpus e sua relação tanto com o ensino de línguas quanto seu papel na MT. Por fim, a partir da análise que se faz, devido ao sistema híbrido que utiliza o sistema baseado em regras com o sistema baseado em corpus, o *Google Translate* produz textos relativamente compreensíveis e legíveis. Apesar de suas limitações evidentes, essa tecnologia pode fornecer consciência linguística quando explorada pedagogicamente pelos professores de ESP no contexto acadêmico

Palavras-chave: Tradutor Automático; ferramenta pedagógica; leitura; Inglês para Propósitos Específicos.

Submitted on August 04th, 2020

Accepted on October 07th, 2020

Introduction

In Brazil, English for Specific Purpose (ESP) courses are commonly offered in Universities, and they are basically focused on reading skill practice. Recently, a number of Higher Education Institutions (henceforth HEI) in the country has been developing actions for internationalization, which aim to allow Brazilian universities to take part in the international academic community, and part of this process involves foreign language learning, specially English, which is considered a *Lingua Franca* (*lato sensu*) in the scientific context. One of the actions worth mentioning is the *International Mobility Program Science Without Borders* which aimed at developing science in Brazil by funding undergraduate and graduate members of the academic community to study in universities abroad. In fact, a discussion on internationalization and language policies in Brazil has already been published by Sarmento; Baumvol, Martinez (2019) on a special issue of the journal *Organon* by the Federal University of Rio Grande do Sul-URFGS.

Among the methods for teaching ESP in the academic environment in Brazil, developing reading strategies is the most frequent

one because it is considered a need for undergraduate students (CELANI *et al*., 1988, 2005; RAMOS, 2009) and ESP focuses on the student's needs (BLOOR, 1997; HUTCHINSON; WATERS, 1987; ROBINSON, 1980, 1991). Reading comprehension is a complex task that involves various cognitive abilities and strategies and reading in a foreign language (FL) is even more complex and requires a certain level of proficiency to be accomplished. A beginner learner faces different linguistics barriers, such as language structure and most of all, lack of vocabulary knowledge.

Depending on the reading goal, skimming reading or knowing the general subject is enough, but if the purpose is studying and learning, for instance, is it necessary to go beyond the main ideas. Although the learner uses strategies such as focusing on cognate words and trying to infer the general meaning through previous knowledge, in order to have a deeper comprehension, at some point the reader will have to search for the meaning of words, which can be done by looking a word up at a dictionary, or even checking whole sentences by using an automatic online translator.

The Internet era has provided us with a wide range of easily accessible information on various subjects. Although the effectiveness of some online resources may be doubted from the educational perspective, the Internet is still one of the main sources for research and language learning. Using machine translation (MT), more specifically, Google translate, and having the information instantly translated instead of thinking and analyzing is a subject that might concern most language teachers. However, it is undeniable that most students resort to these tools when they are struggling with a meaning in a foreign language. In addition, what must be considered is that the contemporary world presupposes the use of technology and the internet has a major role in it.

There is a number of studies suggesting the use of MT in EF classes through different perspectives, such as the ones related to translation programs (LEWIS, 1997; MCCARTHY, 2004), comprehension and acceptation of translated texts (LEFFA, 1994; PETRARCA, 2002), the relationship between MT and English as a global language (CRIBB, 2000), techniques for detecting plagiarism and work produced by MT (LUTON, 2003), writing and post-editing via MT (NIÑO, 2004; KLIFFER, 2005), in addition to using MT as a reading strategy (RIESS, 2015; SCARAMUCCI, 1997).

This paper addresses the use of free online MT in ESP courses at universities in Brazil. For this purpose, previous investigations on the use of MT for FL teaching and learning are explored before discussing their implications for the language class along with some practical examples of using MT for language teaching for academic reading purposes. It is important to bear in mind that Machine translation is a resource for teaching grounded on a perspective defined by Larsen Freeman (2003) as *grammarian*, students do not learn by a set of rules, more than that they learn by doing grammar. In addition, using MT is only seen as feasible the moment technologies of information and communication have proved to offer advantages for teaching and learning. This is specially seen in the studies of Corpus linguistics (SARMENTO; TESSLER; BAUMVOL, 2019)[1] as well as Ergodic learning (LEFFA; BEVILÁQUA, 2019) where students count on resources that go beyond the didactic material because they stimulate both autonomy and learning styles.

In order to intertwine the objectives of teaching ESP and the advantages of using MT as a pedagogical support this paper is divided in three parts. The first one explores the area of English for Specific Purposes; it presents a brief overview on how the field was developed in Brazil up to the present times. The second part discusses Machine Translation and the conceptual foundations that support our belief it can be used to enhance learning. Finally, the third part describes the proposal itself on how MT is suitable for designing ESP lessons.

1 English for Specific Purposes

According to Hutchinson and Waters (1987) and Bloor (1997), the main issues related to ESP in teaching began in the 1960s, when the United States economy became dominant in the western world and English became an internationally accepted language in trade, business in general and in the academic context. Consequently, language learning needs began to have an important role to the design of language courses. Hutchinson and Waters (1987) defined ESP as a course in which "all the decisions as to content and method are based on the learner's reason for learning.

[1] English as a Medium of Instruction into Practice. Oral Presentation. 27th Annual Conference of the Brazilian Association for International Education, Cuiabá, 2015.

The concept of ESP was explained by Robinson (1991) as "an enterprise involving education, training and practice and drawing upon three major realms of knowledge: language, pedagogy and the students'/ participants' area of interest" (ROBINSON, 1991, p.1). The author cites some criteria for a course be considered as ESP, such as "goal directed", meaning that the student needs the language for work or for studies; the relevance of precision in needs analysis; establishing course length; and student profile - normally adults or young adults who usually have some background knowledge of the language. Another relevant aspect about ESP courses is that they must be designed based on the students' area of studies.

Methodological aspects of ESP were described by Dudley Evans and St. John (1998), who point that these courses normally take place in the working context or at university and the methodological choices may differ from General English courses, since the courses may be designed having a specific discipline in mind. Aiming at discussing the specificities of ESP in the Brazilian context the next sections present an overview of the area both in the first years of development and the current studies in progress.

1.1 A brief overview of ESP in Brazil

The origin of ESP courses in Brazil began with a project developed by a group of researchers from PUC-SP in the 1970's (CELANI *et al*., 1988; RAMOS, 2009). The project was initially under the coordination of the researchers from PUC-SP and involved most of the federal universities in Brazil. The focus of the studies was around the reading skill due to the results of the needs analysis which indicated that most university students at that time needed to read texts and books in English on their specific area of studies. This is the reason why ESP courses in Brazil are associated with the teaching and practicing of reading.

In Brazil, ESP is often related to instrumental reading, which comprises reading practice for comprehension and interpretation of texts. From the conception of reading as an active process, Farrell (2003) explains that, after establishing the focus of the class and selecting the texts (based on student needs), the next step is the teaching of reading fundamentally anchored in reading strategies. According to the author, the success of reading activity is directly related to the use of appropriate and effective strategies.

The process of internationalization of HEI has highlighted the importance of English as a global language. Hamp-Lyons (2011) points out that the role of EL has become increasingly stronger, which forces researchers to publish their studies in English. In addition, as stated by Hyland (2006) researchers find more reference to their studies when they are published in EL. One of the reasons ESP has been an area of interest in the Brazilian context is the importance of the country in an international scenario. It seems to be obvious that linguistic skills play a key role in this panorama where sociocultural, economical and political powers are at stake.

Along the late decades new approaches for ESP have been discussed by various researchers, such as Ramos (2004) who introduced the pedagogical importance of implementing genre in the EFL classroom. More recently, Sarmento, Tessler, and Baumvol (2015) have discussed the idea of English as Medium of Instruction (EMI) related to internationalization in the Brazilian HEI and the need to introduce other skills into the ESP courses, such as speaking, listening and writing. This approach, which aims to assure that the student will be able to participate in international academic events, as well as being able to publish their paper in international journals, is called English for Academic Purposes (EAP) and it is currently offered in universities all over the world. However, there is a long way to go in the EAP field in Brazil, since it is a process that has just began to be discussed and analyzed in the country.

English for Academic Purposes (EAP) courses are similar to any other ESP courses, which are based on the student's needs. In this case, the need is related to academic interests, therefore, EAP is a course designed for teaching English to assist learners in target language studies or research (FLOWERDEW; PEACKOCK, 2001; JORDAN, 2012). Consequently, the content "currently focuses on seeking language specificities used in academia [...] incorporating and going beyond the communicative context" (HYLAND, 2006, p. 2).

In this paper, the term ESP was chosen due to the focus on reading skill, which is a reality in courses offered at different universities in Brazil, but it may be extended to other skills practice in the future. However, developing new methods of teaching English in the academic context is important given its relevance for science within the contemporary global scenery, and using machine translation (MT) is part of the globalized world. Therefore, including this tool in the ESP courses might give another perspective for EFL students in terms of reading comprehension.

1.2 The state of the art of ESP in Brazil

As explored in the previous section teachers and professors in Brazilian universities have already adopted ESP lessons. The first experiences have proved to be fruitful so as to enrich teaching with new methodologies and resources. The present section describes current studies that represent the state of the art of ESP in Brazil and point to the future of the area. In general terms, technology plays a key role in this development, either because of electronic corpora or because of MT itself.

It was already mentioned that there are many examples of universities in Brazil which have introduced English in the academic context. However, it is worth talking about some recent studies which can be considered "hands on" in terms of teaching. For instance, Maciel and Vergara (2019) discuss the role of English teaching in a Medical school; Sarmento and Baumvol (2016) also investigate multilingual students learning through Integrated Language and Content Instruction (ILCI). Moreover, Kirsch (2019) analyzes how teachers of English were trained to perform in light of internationalization/globalization.

Basically, in all the work by the authors mentioned above the idea is centered in earlier fundamentals of teaching, specially the view of ESP teachers as also researchers. Dudley-Evans and St. John (1998) explain it is necessary to raise adequate textual information to design instruction and that needs assessment for creating and adapting didactic materials. In addition, there must be a further thought about developing reading strategies during ESP lessons. There are several types of strategies that can turn reading into an active process in a foreign language class; these strategies have long been investigated, such as the useful list seen in Anderson (1991) and Koda (2005), as well as the discussion about translation in Grabe (2009). The teacher who uses language learning strategies, according to Oxford (1990) is no longer at the center of teaching and becomes a facilitator, helping students to identify and use the appropriate strategies. It is undoubted for these authors that students become more autonomous in learning (or acquiring) a language when reading strategies are at stake.

In terms of teaching strategies, Coxhead (2000) suggest that in ESP students should be exposed to high frequency academic words, most of times gathered in ready-made lists such as *the academic word list online*. Regarding grammar, Larsen Freeman (2003) talks about a concept called *grammarian* from the perspective that grammar and vocabulary are

addressed as integrated with the 4 skills, rather than stand-alone issues. Consequently, grammar cannot be seen as a set of rules to be memorized, but as a phenomenon of *doing grammar*. As a whole, an updated ESP course has to consider the use of an academic English language corpora available online. In addition, it has to develop general academic skills and have teachers explaining all the existing learning styles, strategies, as well as make students be part of their learning process by discussing their preferences.

Examples of ESP courses are the ones described by Berber Sardinha (2006) that envisaged instructional material production using corpora. Moreover, Barbosa (2004) prepared units on a Foreign trade course at a distance founded on an experimental approach (KOHONEN, 2001). Leffa (2019) explains that language pedagogies must evolve from three perspectives: instructionism to constructivism and then to connectionism (HEICK, 2017; MATTAR, 2018). It all starts from linguistic exposition (linguistic corpora) and it moves to active learning, which demands the student to use accessible resources to learn. This sequence moves up to ergodic learning; this concept is grounded on the idea that the student not only builds his/her own learning according to the resources he/she uses, as well as he/she modifies the learning dimension.

While we can talk about the state of the art of ESP in Brazil today, it is not less important to discuss MT, once this is the center of this paper. It is not obvious, neither it is accepted at all ESP courses should allow students to use online translators, such as Google translate. The next section presents the area of Machine translation because it is believed that the limitations as well as the strengths of this tool enable instructional designers or teachers to take advantage of them for more accurate use.

2 Machine translation

If we are to investigate the introduction of MT in ESP classes we have to keep in mind that Foreign Language tutors do not necessarily need to know much about the working MT software, but they should know these tools are available, how to use them, and their general strengths and weaknesses. Therefore, it is important for this paper to present some of the basis of the systems for MT, because the teacher as a researcher may be limited by the performance of the tool. As it was stated previously in this study, MT may provide linguistic awareness as a source of knowledge much more efficiently once it is explored pedagogically.

The introduction of MT into language learning contexts has been compared to the advent of the calculator by some researchers (LUTON, 2003, p. 770; GROVES; MUNDT, 2015, p. 120). Nevertheless, "while there seems to be general agreement that children should learn to do basic arithmetic without a calculator before moving on to more advanced operations in which the calculator can be used as a shortcut, it is possible that the parallel to MT and language learning does not extend as far". Therefore, many teachers seem to be resistant to using MT in class, which could compromise the way students think and develop language learning. What has to be reinforced is that linguistic skills and mathematics are different kinds of knowledge, for this reason it is not clear that a mere comparison between the two of them is a benefit.

One of the proposals of this study is to present teaching English in the university context and the role of the teacher as developing linguistic awareness, not only language itself. Tomitch (2009) explains that teaching reading in a foreign language is much more than teaching language itself. For this reason, it is believed that through the analysis of translation errors related to specific lexical items and syntactic structures, students will be able to understand how translation works as well as develop language awareness. We consider the lexicon brings information related to cultural differences, for example, that go beyond the language itself. Error analyses and language transfer are not new in the area of Second language Acquisition (henceforth SLA); however, they have a broader view here. In order to give a hint of what is meant in this proposal, the example that follows can clarify. Although students are not quite familiar with terms such as polysemy, lexical ambiguity, and structural ambiguity, examples in English using MT may illustrate that polysemy can create lexical ambiguity, and word order can determine the level of structural ambiguity in a sentence. An example could be seen in *We need to see that house* and *I will house the club in a new building,* in the first sentence the word *house* is a noun and in the second it is a verb; syntax in this sense differentiates the meaning. Students, then, are displayed with an obvious differentiation in meaning that translation can easily determine.

2.1 What MT is and how it works

According to the definition by Hutchins and Somers (1992), MT are "computer systems responsible for the production of translations from one natural language into another, with or without human assistance"

(HUTCHINS; SOMERS, 1992, p. 3). Hutchins (2003) explains that in fully automatic translation, the system translates the entire text without the intervention of the human translator, producing a raw translation, commonly known as 'informative translation'.

MT systems can be programmed for two languages (bilingual systems) or for more languages (multilingual systems). In terms of processing, MT systems can be grouped into two categories: rule-based system and corpus/statistical-based systems. Rule-based systems work based on rules for/of morphology, syntax, lexical selection and rules transference. They operate by filtering source text input such as morphological analyzers, part-of-speech taggers and bilingual dictionaries and then they transfer rules and reorder them, whereas statistical MT systems are based on "machine-learning technologies" and rely on "large volumes of parallel human-translated texts from which the MT engine can learn" (STEDING, 2009, p. 184).

Rule-based systems contain three main approaches. The first is the direct approach which, according to Hutchins (2003) definition, operates as a large bilingual dictionary, in which the source text is translated word by word, not considering the analysis of its syntactic structures or relations between words in the sentence.

The second is the transfer approach, which is more widely used and comprises three phases: analysis, transfer and generation. Hutchins (2003) explains that transfer systems may have separate programs for lexical transfer (selection of equivalent words in vocabulary) and structural transfer (transformation of source language structures into appropriate target language structures), differently from the direct approach, the transfer approach takes structure/syntax into consideration.

Developed in the 1980s, interlingua is the third approach, which is based on the assumption that it is possible to convert target language (TL) texts into common syntactic-semantic representations in different languages. Interlingua is based on the principles of universal linguistics. Thus, translation occurs into two phases: from source language to interlingua and from interlingua to target language.

Corpus-based approach is more recent and can be divided into two categories: statistical-based MT and example-based MT. Somers (2003) explains that this model is based essentially on the alignment of words, expressions and word sequences in a parallel bilingual corpus as well as on probabilities calculus of words/expressions in a given

sentence to correspond to one or more words in the equivalent sentence in the target language.

The word "língua" in Portuguese can be used as an example of what was mentioned in the previous paragraph because it represents ambiguity meaning both language and tongue, which are different words in English. Using Google translate, when inserting the sentence: "Eu mordi minha língua", the result is "I bit my tongue", but when the sentence is "Eu não falo sua língua", the translation is "I don't speak your language". This means that the relation between the verb and noun is calculated through probability. The verb "*morder* (bite)" combined to "*língua* (tongue)" and the verb "*falar* (speak)" is combined to "*língua* (language)". Somers (2003) ensures that it involves a great complex probability calculation.

Unlike the statistical-based system, example-based MT operates by comparing the input with a corpus of representative examples already translated, drawing the closest correspondences as a translation model for the target text. As Somers (1998) points out, this approach resembles the way human translators work, since it solves new problems based on solutions used for similar previous problems and this is why the translation outcome is more fluent and less literal.

Many MTs are actually hybrids, their basic design and main mode of operation put them in one of the two categories. Google Translate exemplifies the latter approach, as its website explains in the following terms: "By detecting patterns in documents that have already been translated by human translators, Google Translate can make intelligent guesses as to what an appropriate translation should be" (Google).

Bowken (2002) makes a connection between this approach and output quality, noticing that because statistical MT reflects a "better understanding of the strengths of machines" than earlier methods, errors are "less common and considerably less outrageous" than in the past (BOWKEN, 2002, p. 3).

2.2 Corpus linguistics and MT

One of the investigations that is also part of the theme "Machine translation" as a whole is *Corpus linguistic*, more specifically, *Electronic Corpora*. This is a branch of Computational linguistics that deals with the processing, storage and analyses of great amounts of linguistic data that

are machine readable. The Latin term *corpus* refers to a body because it is formed by a variety of relevant linguistic information which display both oral and written language behavior.

In this study that discusses the use of online *Google translate* it is implicit we are talking about electronic *corpora,* since information is changed into digits that are machine readable, as we stated before. A traditional *corpus* can be a collection of physical texts, for instance, the indigenous talk of some Amazon region annotated by ethnographic research. As for *Google translate* it utilizes all the written texts that were translated by humans and published on the web. For this reason, its database is giant, as such, it can extract linguistic items which are able to generate new translation in the target language.

It is also worth saying in the case of MT that Google corpora can be either Comparative or Parallel. They both approximate; however, comparative corpora contrast the linguistic items of each language, they deal with at least two languages (2 monolingual texts) that are contrasted. An example of this corpus is *Compara* seen at *http://portugues.mct.pt/COMPARA/*, it is made up of literary work in its original language by the side of its translation that was once published by a human translator. A slight difference to Parallel corpora is that there is a linguistic *corpus* in the first language and in parallel the engine displays encountered translations (MCENERY; XIAU; TONO, 2006). Basically, Google deals with parallel corpora, but as it will be studied next it actually uses a hybrid system.

In terms of language teaching, which is the focus of this article, parallel corpora have been much used to a diversity of pedagogical objectives. They can be used to teach technical vocabulary in ESP classes; Berber Sardinha (1998)[2] for example, presents the Business English corpus that is fruitful if the teacher needs a good source for instructional material. In addition, Riess e Gabriel (2019) analysed lexical disambiguation during reading in English as FL/L2 using the *Webcorp,* a linguistic corpus of general English. The authors investigated how a reader disambiguated the word *mind* in its different contexts of meaning.

In 2015 a special issue of the Brazilian journal *Domínios da linguagem* was published in which the editors focused on linguistic

[2] Size of a representative corpus. Summary of discussion on CORPORA email discussion list, 26 August 1998.

corpora, methods, and interfaces. In the presentation section they mention the development of this area in Brazil and present other journals which were also dedicated to related subjects, such as *Veredas* in the year of 2009 and *Revista Brasileira de Linguística Aplicada* (2011). Almost 10 years ago these journals already pointed Corpus Linguistic as a promising field to be explored by researchers of language.

What is also important to mention here refers to corpus linguistics and statistical language learning. In fact, an electronic translator such as Google needs to be fed with a great deal of data; however, it works on the basis of statistics. That means the more frequent the translation of a word or segment appears, the more it will be used next time. Charniak (1996) explains that the objective of statistics is to count how often the translation appears, so that the preferable translated word or segment becomes a prototype. Consequently, programming can calculate the probabilities of an occurence.

2.3 MT as a teaching resource in ESP class

Incorporating technology in foreign language classroom without compromising students' practice is a challenge for teachers all over the world. The relevance of this subject is highlighted by Hall (2001) when he states that "How well we prepare learners of additional languages to meet the social, political, and economic challenges of the next several decades will depend partly on our success in integrating technology into the foreign language curriculum". Teaching students to evaluate these tools may provide them with a model for criticizing as well as reflecting on other technologies that work as pedagogical resources.

In fact, technology can offer excellent tools for language teachers and MT has been discussed as one of them. Different perspectives of studies have been conducted by researchers in various parts of the world. These works promote the use of MT in language teaching for distinct purposes, such as for developing critical thinking, promoting electronic Literacy and language awareness (CORREA, 2014; WILLIAMS, 2006) for correcting/editing grammar, training future professional translators (KLIFFER, 2005) and even for preventing plagiarism (STEDING, 2009). Post-editing is the most common practice, and it involves the correction of raw MT output into an acceptable text for a particular purpose (BELAM, 2002; LA TORRE, 1999; NIÑO, 2004).

A study conducted with FL students at Duke University in 2011 and 2012 evidenced the student's preference for Google Translate through the significant percentage of 81% that reported using it to support their language learning (CLIFFORD *et al*., 2013, p. 111). This tool is described as a "free translation service that provides instant translations between dozens of different languages" (Google).

Ana Niño's (2009) survey about students' and professors' perceptions of MT has served as the basis for investigation at Duke, which addresses her observation that a survey of a broad spectrum of language students (not just advanced learners) would provide a more accurate picture of the state of MT use in the language classroom.

One of the first authors to research on the role of MT in language classes was Ball (1989). The author suggested the correction of errors presented in computer-produced translations as teaching application of MT for language students. Later, Somers (2003, p.327) stated that this practice could "bring out subtle aspects of language differences" as well as "reinforce learners' appreciation of both L1 and L2 grammar and style".

Other authors who recommend teaching practices using MT in FL contexts include McCarthy (2004), who wrote 12 "solutions" for dealing with the inevitability of MT use based on the discussions with the students; Williams (2006), who suggested using MT websites to improve students' electronic literacy; Steding (2009), whose work was concerned about preventing MT-based cheating and Niño (2009) proposing "good practices" and "bad practices" for MT use in LF class (p. 247-248).

Although at least one study (GASPARI; SOMERS, 2007) discusses the need for discouraging students from using MT for single word-lookup, the majority of studies that problematize the use of MT by students are concerned with the translation of longer texts and the belief that this is a form of cheating in FL class (CASE, 2015). Rather than looking only at the possible misuses of this relatively new electronic tool, however, we may wish to examine it further for its potentially positive applications in the study of foreign languages (WILLIAMS, 2006, p. 566-567).

Most studies in this field work with MT through two types of activities: post-editing and contrastive analysis. Post-editing exercises involve translating a text into the target language using MT and using one's skills in the target language to "correct" the "errors" made by the computer (BELAM, 2002; GROVES; MUNDT, 2015; KLIFFER , 2005;

NIÑO, 2008; SOMERS, 2003; ZANETTIN, 2009). Contrastive analysis involves translating from the target language to the students' native language, so that students can see the kinds of errors are produced in order to highlight differences in language structure, idioms, and collocations (ANDERSON, 1995; SOMERS, 2001, 2003). Most studies have focused on advanced learners or translators in training, (BELAM, 2002; NIÑO, 2008; O'BRIEN, 2002); however, there are some researchers who argue or investigate their use with beginners (CORNESS, 1985; GARCÍA (2010).

Garcia and Pena (2011) claim that MT helped their beginner-level students identifying MT as a type of scaffolding that together with the other digital tools and online activities could "support students in generating authentic language while interacting and collaborating in an enjoyable learning environment, with technology as the facilitator and stimulator of communication" (PENA, 2011, p. 66).

Another author that supports using MT as an advantage for students is Williams (2006), who suggests that they "force students to think about language as a communication tool, not as a set of decontextualized vocabulary words or phrases" (p. 574). In this sense, MT is not in itself "bad," of course, and there is an educational place for it in the classroom according to Case (2015), but many language instructors probably will encounter it as an uninvited guest.

3 Reading comprehension and a proposal for using MT in ESP classes

When it comes to the concept of reading comprehension, we need to discuss reading models. The interactive model described by the seminal work of Rumelhart, (1985) proposes an integration of the bottom-up processing (GOUGH, 1972), which understands reading from a linear sequence from decoding letters to sounds, words, sentences, and finally to meaning, and top-down processing (GOODMAN, 1967, 1970), in which the reader uses his prior knowledge to interpret the text and create hypothesis in order to understand it. From the interactive perspective, comprehension is not a final product, but a process that is developed along the reading according to the strategies and resources that the reader uses to attribute meaning to the text. The integration of both models (bottom-up and top-down) is defended by a number of authors based on the idea that comprehension depend on both, the information that is in the mind of the reader and what is printed in the text. Therefore,

reading is a dynamic process of meaning construction. Proficient reading is defined by Perfetti, Landi and Oakhill (2008), according to two components: word recognition and textual comprehension. It is assumed that comprehension cannot be achieved without the efficiency of the most basic processes involved in word recognition. Thus, understanding how reading comprehension occurs requires the study of the variables involved in word identification.

According to Perfetti (1985, 1999); Perfetti and Hart (2001), the limited capacity of working memory (WM) makes reading difficult when several processes that require attention need to be activated simultaneously. Based on this assumption, less proficient readers first struggle with the lowest elements in the hierarchy, which are language spelling rules and lexical knowledge, and then they move to the higher elements, which are syntactic and semantic knowledge. Although it is a useful strategy in the early stages of reading development, contextual exploration is just a stage and should not be the only support for compensating for vocabulary lacking.

During reading, the activation of prior knowledge in memory is essential for the construction of meanings, which may occur implicitly or explicitly through inferencing (DELL'ISOLA, 2001). In other words, meaning is produced as a result from the recovery of knowledge and values that are activated through various cognitive strategies, among which the production of inferences is fundamental.

The definition of inference is explained by Koch (1997) as "what is used to establish a non-explicit relationship in the text, between two elements of the text. The inferences arise from need and from the reader's world knowledge" (KOCH, 1997, p.70 translated by the author). The concept of inference varies according to different authors. However, it is a consensus among authors that inferences fill in the gaps in the text, since the text itself does not comprise all the necessary information for the reader. Therefore, comprehension issues may result from this incompleteness or lack of knowledge, since miscomprehension may occur depending on the information that has been activated.

Nuttall (1996, p. 75) stated that "to infer the reader must have sufficient clues. Lexical inference will not help readers if all or most of the words are unknown. If, moreover, the context does not offer sufficient clues, inference becomes impossible." As a consequence, less proficient readers in FL may be blocked in terms of bottom-up processing, which

will make it difficult to access prior knowledge in order to fill in the text gaps. Thus, the use of MT can be beneficial for beginner learners of FL, providing a general idea of the text so that the reader will be able to raise hypotheses about the text.

Another important aspect pointed by Grabe (1991) is that L1 readers usually have a wide vocabulary before they begin to read, while FL readers generally have a restricted vocabulary added to less experiences in the target language. The author argues that, although the reader has a good master of syntax in the FL, he is unlikely to be familiar with pragmatic and cultural knowledge, which are related to social interaction that is common for native speakers and it can hinder the perception of these aspects in the texts.

Due to the complexity involved in reading comprehension, there are many strategies that can be used for reading successfully. Strategies are behaviors that are consciously selected to facilitate understanding (NORDIN; RASHID; ZUBIR; SADJIRIN, 2013). For example, readers may decide how much time to spend looking at a word, whether to reread a section or to skip a section. They must decide when to summarize, question the text, or make predictions and in order to do so, readers depend on their executive control abilities (ARRINGTON; KULESZ; FRANCIS; FLETCHER; BARNES, 2014; CARTWRIGHT, 2012).

However, reading strategies are not only conscious and they can be divided into cognitive or metacognitive (KATO, 2007). They work as support to construct the textual coherence through the relations established between the elements of the text, such as syntactic segmentation strategies, and anaphoric retrieval. Proficient readers tend to use strategies more automatically, but when some new aspect arises interrupting the comprehension process, it makes the reader act consciously, slowing down his reading process in a reflected or metacognitive way.

These strategies function as fault detection mechanisms and result from increased processing capacity effort. Perception of reading failure is a part of comprehension monitoring, as the reader needs to know what to do when the failures occur, and this is where strategic decisions must be made. Furthermore, it has been postulated that good readers are more metacognitively aware of their own strategies than less proficient readers, as they tend to monitor comprehension better, being more aware of the characteristics of the text and the strategies they use while reading (NUTTALL, 1996).

There are several different types of reading strategies that are not going to be explored in this paper because it is not the main point. In summary, reading strategies play a special role in both NL and FL reading, as they are cognitive mechanisms for comprehension and therefore can be taught or develop language awareness, especially in reading.

One of the goals of ESP in Brazil, in the HEI context, is to assure that students will be able to read and comprehend academic texts in English as a foreign language. This paper proposes the use of MT in ESP courses for reading academic texts in EFL. Based on the studies cited in this paper it is possible to bring these practices into the reality of ESP in Brazil and adapt the activities suggested as well as design new possibilities.

Another relevant aspect of ESP courses is that dictionary use is an example of a conscious reading strategy for solving a lexical difficulty. From the same perspective, MT can be a support reading strategy. Riess (2015), in her PhD dissertation on reading strategies and the use of Google translator, discusses the use of MT as a reading strategy. The author claims that:

> The idea is not to exclude the tool from the class, on the contrary, it is to include it and point out to failures and strengths of translation. We suggest the use of Google translate as a strategy that benefits reading comprehension, because the reader can search for translation to what is unknown, at the various levels- from the lexicon to the sentence till the whole text. (RIESS, 2015, p.104)

The idea is to include MT in class by pointing out the linguistic items the translation has either failed or succeeded. In this sense, it is suggested that using Google Translate is a strategy that benefits reading comprehension, since translation can be used at the lexical, sentence or text level, depending on the reader's proficiency. Riess mentions authors such as Perfetti and Hart (2001) and Scaramucci (1997), based on the idea that by increasing vocabulary, reading can flow better, because words are better understood, which makes comprehension occur. Also, in her study, Riess (2015) found that students prefer translating words more than the whole text. However, the number of participants may not be statistically enough to conclude such a behavior. In this sense, the use of Google translate would not be that different from the dictionary. However, participants from Eastern languages (mostly Arabians) demonstrated to

translate sentences more than words. In the verbal protocols they say word order is one of the most difficult for them.

In this paper, we propose working with MT in ESP courses that focus on academic reading because, in Brazil, using dictionary is a common support strategy in these courses, since the majority of students show low level proficiency and need this resource for comprehension. Thus, MT can be an important tool as a reading strategy. Therefore, meanings can be searched through online dictionaries, but also through Google Translate, which is proved to be widely explored among students. Therefore, ESP teachers can use it for pre-editing and postediting activities using MT in different levels of search, such as word meaning, sentences or even entire texts. Abstracts can be used as an example of text, being short texts, which summarize scientific papers, these texts are suitable for academic reading classes. It is important to emphasize that scientific language can be quite predictable, being more direct by using objective language and avoiding language aspects which normally appear in literary texts and usually result in mistakes for MT such as metaphors and other figures of speech.

In this article we suggest teachers three tasks to be used in ESP lessons. They are theoretically founded in the subjects already described previously, they are 1) Translating and discussing, 2) Reading and translating, 3) Checking mistakes. In the first case an abstract translated from English to the student's native language could be read and discussed in class focusing on the linguistic issues. The teacher, then, counts on the student´s linguistic awareness, because discussions will probably be raised by the student´s view. The second activity is reading an abstract in the FL using strategies and then have the Google translation to check the ideas. In such a circumstance monitoring reading is at stake because students have to compare their comprehension either by using strategies or by the language translated by the machine. The third suggestion we give has to do with the efficiency of machine translation. This is because discussing the possible mistakes made by MT is a different way of teaching grammar rules and structures. In addition to these linguistic items, teachers can also explore vocabulary. The role of the teacher is more active in this case, because it is him/her who points out where idiosyncrasies are. It is different from using parallel bilingual corpora, for example, because in this situation the student counts more on his/ her intuition to infer meaning, whereas with MT the teacher is, to a certain

extent, the mediator of knowledge. In fact, this has to do with explicit teaching rather than implicit, because teachers will locate where errors are and display them in order to call student's attention.

While using Google translate the student's attention is focused on the translated text, thus, by using MT for formal instructions, teachers can provide the students with a chance to pay attention to the original text, which is the purpose of ESP courses. Additionally, abstracts are usually short, and the general language corpus are very similar, which creates an opportunity to work with scientific language, and prepare students for future academic writings. This view agrees with the idea of integrated language and content language instruction already discussed by those involved in the Science without borders program described previously (SARMENTO TESSLER; BAUMVOL, 2015).

Final remarks

Technology has been used in education all over the world in different types of subjects and there are several online tools available for pedagogical practices in language learning. Among various resources for foreign language studies through technology, this study demonstrates that MT systems have been improved over the last decades and that new methodologies are being employed on a linguistic and interdisciplinary basis. Research show that Google translate is the most accessed free online MT tool by students. Google translate is processed by a hybrid system that combines rule-based and corpus-based systems, in quite coherent texts.

Reading comprehension requires lexical knowledge, consequently, low proficient readers in a foreign language usually struggle to comprehend FL texts due to lack of vocabulary and language structure. ESP courses in universities in Brazil are generally focused on reading skill development. Assuming that English is a global language in science, undergraduate and graduate students should be able to read in English in order to access international research.

Based on a number of studies conducted by researchers from different parts of the world, MT can be used in FL classes for pedagogical purposes. There are different approaches and activities that can be carried out with learners, such as post-editing and analytical tasks that have been proved to help students to raise language awareness and knowledge.

The methodology applied in ESP courses involve not only the integration of content (the specific field) as well as teaching reading strategies, which includes dictionary use. Therefore, this paper proposes using MT as a reading strategy and as a pedagogical tool for teaching grammar, vocabulary and develop language awareness.

Finally, it is important to emphasize that further studies need to be conducted, especially in relation to ESP in academic settings, since most of the studies are regarded to general EF language courses. Furthermore, research should be designed in order to develop methodologies that can quantify the comprehension degrees of an automatically translated text and the cognitive aspects involved in this process, as well as investigations on the effectiveness of using Google Translator as a pedagogical instrument. We believe a combined quantitative and qualitative study can show more concrete outcomes to both experience and intuition on the use of MT for fruitful pedagogical practices.

Acknoledgements

We gratefully acknowledge the support of CAPES for the split doctorate both authors were funded at the University of Pittsburgh at the Learning Research and Development Center-LRDC.

Authors' Contributions

Débora Ache Borsatti is a PhD student at the University of Santa Cruz do Sul (UNISC). The author has been researching about the use of machine translation in ESP courses, focusing on reading in English for Academic Purposes. This paper was written as an assignment for the course "Second Language Acquisition", lectured by author 2.
Adriana Blanco Riess is a co-author in the paper. As an assistant professor at UNISC, Adriana Adriana read, reviewed, and added important contributions for the proposed discussion.

References

ANDERSON, N. J. Individual Differences in Strategy Use in Second Language Reading and Testing. *Modern Language Journal*, [*S.l.*], v. 75, n. 4, p. 460-472, Winter 1991. DOI: https://doi.org/10.2307/329495

ANDERSON, D. D. Machine Translation as a Tool in Second Language Learning. *CALICO Journal*, [*S.l.*], v. 13, n. 1, p. 68-97, 1995. DOI: 10.1558/cj.v13i1.68-97

ARRINGTON, C. N.; KULESZ, P. A.; FRANCIS, D. J.; FLETCHER, J. M.; BARNES, M. A. The Contribution of Attentional Control and Working Memory to Reading Comprehension and Decoding. *Scientific Studies of Reading*, [*S.l.*], v. 18, n. 5, p. 325-346, 2014. DOI: https://doi.org/10.1080/10888438.2014.902461

BALL, R. V. Computer-assisted translation and the modern languages curriculum. *The CTISS File*, [*S.l.*], v. 8, p. 52-55, 1989.

BARBOSA, M. E. D. C. *Material didático para o ensino de inglês instrumental online*: uma abordagem experiencial baseada em corpus, gênero e tarefa. 2004. Dissertação (Mestrado em Linguística Aplicada e Estudos da Linguagem) - LAEL, Pontifícia Universidade Católica de São Paulo, PUCSP, 2004.

BELAM, J. Teaching Machine Translation Evaluation by Assessed Project Work. *In*: EAMT WORKSHOP TEACHING MACHINE TRANSLATION, 6th, Manchester, UK. *Proceedings* […]. Manchester: European Association for Machine Translation, 2002. p. 131-136.

BERBER SARDINHA, Tony. Preparação de material didático para aprendizagem baseada em tarefas com WordSmith Tools e corpora. *Calidoscópio*, São Leopoldo, RS, v. 4, n. 3, p.148-155, set./dez. 2006. DOI: https://doi.org/10.4013/6001

BLOOR, M. The English Language and ESP Teaching in the 21st Century. *In*: MEYER, F.; BOLIVAR, A.; FEBRES, J.; SERRA, M. B. (ed.). *ESP in Latin America*. Bogotá: Universidad de los Andes; CODEPRE, 1997.

BOWKEN, L. *Computer-Aided Translation Technology*: A Practical Introduction. Ottawa: University of Ottawa Press, 2002.

CARTWRIGHT, K. B. Insights from Cognitive Neuroscience: The Importance of Executive Function for Early Reading Development and Education. *Early Education and Development*, [*S.l.*], v. 23, p. 24-36, 2012. DOI: https://doi.org/10.1080/10409289.2011.615025

CASE, M. Machine Translation and the Disruption of Foreign Language Learning Activities. *eLearning Papers*, [*S.l.*], n. 45, p. 4-16, 2015.

CELANI, M. A. A. *et al*. *ESP in Brazil:* 25 Years of Evolution and Reflection. Campinas: Mercado de Letras; São Paulo: EDUC, 2005.

CELANI, M. A. A. *et al*. *The Brazilian ESP Project:* An Evaluation. São Paulo: EDUC, 1988.

CHARNIAK, C. Statistical Language Learning. Cambridge: MIT Press, 1996.

CLIFFORD, J.; MERSCHEL, L.; MUNNÉ, J. Surveying the Landscape: What is the Role of Machine Translation in Language Learning? *The Acquisition of Second Languages and Innovative Pedagogies,* Durham, NC, n. 10, p. 108-121, 2013.

CORNESS, P. The ALPS Computer-Assisted Translation System in an Academic Environment. *Translating and the Computer*, [*S.l.*], v. 7, p. 118-127, 1985.

CORREA, M. Leaving the "Peer" Out of Peer-Editing: Online Translators as a Pedagogical Tool in the Spanish as a Second Language Classroom. *Latin American Journal of Content and Language Integrated Learning*, Cundinamarca, Colombia, v. 7, n. 1, p. 1-20, 2014. DOI: doi:10.5294/laclil.2014.7.1.1

CORREA, M. Academic Dishonesty in the Second Language Classroom: Instructors' Perspectives. *Modern Journal of Language Teaching Methods*, [*S.l.*], v. 1, n. 1, p. 65-79, 2011.

COXHEAD, A. A New Academic Word List. *TESOL Quarterly*, [*S.l.*], v. 34, p. 213-238, 2000. DOI: https://doi.org/10.2307/3587951

CRIBB, V. M. Machine Translation: The Alternative for the 21st Century? *TESOL Quarterly*, [*S.l.*], v. 34, n. 3, p. 560-569, 2000. DOI: https://doi.org/10.2307/3587744

DELL'ISOLA, R. L. P. *Leitura:* inferências e o contexto sociocultural. Belo Horizonte: Formato Editorial, 2001.

DUDLEY-EVANS, T.; ST. JOHN, M. J. *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press, 1998.

FARREL, T. S. C. *Planejamento de atividades de leitura para aulas de idiomas.* Tradução de Itana Summers Medrado. São Paulo: SBS, 2003.

FLOWERDEW, J.; PEACOCK, M. Issues in EAP: A preliminary perspective. *In*: _____. (org.). *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press, 2001. p. 177-194. DOI: https://doi.org/10.1017/CBO9781139524766.004

GARCÍA, I. Can Machine Translation Help the Language Learner? *In*: ICT FOR LANGUAGE LEARNING, 3rd., 2010, Florence. *Proceedings* […]. Florence: Libreria Universitaria, 2010. p. 1-4. Available on: https://conference.pixel-online.net/conferences/ICT4LL2010/common/download/Proceedings_pdf/TRAD02-Garcia.pdf. Retrieved at: Nov. 17, 2019.

GARCIA, I.; PENA, M. I. Machine Translation-Assisted Language Learning: Writing for Beginners. *Computer Assisted Language Learning*, [*S.l.*], v. 24, n. 5, p. 471-487, 2011. DOI: https://doi.org/10.1080/09588221.2011.582687

GASPARI, F.; SOMERS, H. Making a Sow's Ear out of a Silk Purse: (Mis) Using Online MT Services as Bilingual Dictionaries. *In*: TRANSLATING AND THE COMPUTER, 29, 2007, London *Proceedings* […], London: Aslib/IMI, 2007. p. 29-30.

GOODMAN, K. S. Dialect Rejection and Reading: A Response. *Reading Research Quarterly*, [*S.l.*], v. 5, p. 600-603, 1970. DOI: https://doi.org/10.2307/747199

GOODMAN, K. S. Reading: A Psycholinguistic Guessing Game. *Journal of the Reading Specialist*, [*S.l.*], v. 6, n. 4, p. 126-135, 1967. DOI: https://doi.org/10.1080/19388076709556976

GOOGLE. *Inside Google Translate*. Available on: http://translate.google.com/about. Retrieved at: Aug. 28, 2019.

GOUGH, P. B. One Second of Reading. *In*: KAVANAGH, J. F.; MATTINGLY, I. G. (ed.). *Language by Ear and By Eye:* The Relationship Between Speech and Reading. Cambridge: MIT Press, 1972. p. 291-320.

GRABE, W. Current Developments in Second Language Reading Research. *TESOL Quarterly*, [*S.l.*], v. 25, n. 3, p. 375-406, 1991. DOI: https://doi.org/10.2307/3586977

GRABE, W. *Reading in a Second Language*: Moving from Theory to Practice. New York: Cambridge University Press, 2009. DOI: https://doi.org/10.1017/CBO9781139150484

GROVES, M.; MUNDT, K. Friend or Foe? Google Translate in Language for Academic Purposes. English for Specific Purposes, [S.l.], v. 37, p. 112-121, 2015. DOI: https://doi.org/10.1016/j.esp.2014.09.001

HALL, J. K. *Methods of Teaching Foreign Languages*: Creating a Community of Learners in the Classroom. Upper Saddle River: Prentice Hall, 2001.

HAMP-LYONS, L. English for Academic Purposes: 2011 and Beyond. *Journal of English for Academic Purposes*, [*S.l.*], v. 10, n. 1, p. 2-4, 2011. DOI: http://dx.doi.org/10.1016/j.jeap.2011.01.001

HEICK, T. The Difference Between Instructivism, Constructivism, and Connectivism. Teachthought, 2017. Available from: https://www.teachthought.com/learning/thedifference-between-instructivism-constructivism-and-connectivism/. Retrieved at: Sept. 5, 2020.

HUTCHINS, J. Commercial Systems: The State of the Art. *In*: SOMERS, H. L. (ed.). *Computers and Translation:* A Translator's Guide. Amsterdam: John Benjamins Publishing Company, 2003. p. 161-174

HUTCHINS, W.; SOMERS, H. *An Introduction to Machine Translation*. London: Academic Press, 1992.

HUTCHINSON, T.; WATERS, A. *English for Specific Purposes:* A Learning - Centered Approach. Cambridge: Cambridge University Press, 1987. DOI: https://doi.org/10.1017/CBO9780511733031

HYLAND, K. *English for Academic Purposes:* An Advanced Resource Book. New York: Routledge, 2006. DOI: https://doi.org/10.4324/9780203006603

JORDAN, R. R. *English for Academic Purposes:* A Guide and Resource Book for Teachers. 13. ed. Cambridge: Cambridge University Press, 2012.

KATO, M. A. *O aprendizado da leitura.* São Paulo: Martins Fontes, 2007.

KIRSCH, W. Processos de internacionalização e seus legados involuntários: o caso da formação de professores de inglês como língua adicional dos centros de língua inglesa do programa idiomas sem fronteiras. *Organon*, Porto Alegre, v. 34, n. 66, p. 1-16, 2019. DOI: https://doi.org/10.22456/2238-8915.91293

KLIFFER, M. D. An Experiment in MT Post-Editing by a Class of Intermediate/Advanced French Majors. *In*: EAMT - ANNUAL CONFERENCE, 10th., 2005, Budapest. *Proceedings* […]. Budapest: EAMT, 2005. p. 160-165.

KOCH, I. G. V. *O texto e a construção dos sentidos*. São Paulo: Contexto, 1997.

KODA, K. *Insights into Second Language Reading*: A Cross Linguistic Approach. Cambridge: Cambridge University Press, 2005. DOI: https://doi.org/10.1017/CBO9781139524841

KOHONEN, V. Towards Experiential Foreign Language Education. *In:* KOHONEN, V.; JAATINEN, R.; KAIKKONEN P.; J. LEHTOVAARA, J. (ed.). Experiential Learning in Foreign Language Education. London: Pearson Education, 2001. p. 8-60. DOI: https://doi.org/10.4324/9781315840505-2

LARSEN-FREEMAN, D. *Teaching Language*: From Grammar to Grammaring. 1st ed. Michigan: Heinle Elt, 2003.

LA TORRE, M. D. A Web-Based Resource to Improve Translation Skills. *ReCall Hull*, Cambridge, v. 11, n. 3, p. 41-49, 1999.

LEFFA, V. J. Machine-Translated Text: Is It Comprehensible to Proficient Readers? *System*, [*S.l.*], v. 22, n. 3, p. 391-399, 1994. DOI: https://doi.org/10.1016/0346-251X(94)90024-8

LEFFA, V. J.; BEVILÁQUA, A. F. Aprendizagem ergódica: a busca do hipertexto responsivo no ensino de línguas. *Revista Língua & Literatura*, Frederico Westphalen, RS, v. 21, n. 38, p. 99-117, 2019.

LEWIS, D. Machine Translation in a Modern Languages Curriculum. *Computer Assisted Language Learning,* [*S.l.*], v. 10, p. 255-271, 1997. DOI: https://doi.org/10.1080/0958822970100305

LUTON, L. If the Computer Did My Homework, How Come I didn't get an "A"? *French Review*, Marion, IL , v. 76, p. 766-770, 2003.

MACIEL, R. F.; VERGARA, V. S. Internacionalização como prática local: um olhar situado sobre o papel da língua no english club e no curso de medicina. *Organon*, Porto Alegre, v. 34, n. 66, p. 1-17, 2019. DOI: https://doi.org/10.22456/2238-8915.91066

MATTAR, J. Constructivism and Connectivism in Education Technology: Active, Situated, Authentic, Experiential, and Anchored Learning. *RIED. Revista Iberoamericana de Educación a Distancia*, Madrid, v. 21, n. 2, p. 201-217, 2018. DOI: https://doi.org/http://dx.doi.org/10.5944/ried.21.2.20055

MCCARTHY, B. Does Online Machine Translation Spell the End of Take-Home Translation Assignments? *CALL-EJ Online*, [*S.l.*], v. 6, n. 1, p. 26-39, 2004. Available on: http://callej.org/journal/6-1/mccarthy.html. Retrieved at: Dec.10, 2019.

MCENERY, T., XIAO, R.; TONO, Y. Corpus-based Language Studies: An Advanced Resource Book. London: Routledge; 2006.

NIÑO, A. Machine Translation in Foreign Language Learning: Language Learners' and Tutors' Perceptions of Its Advantages and Disadvantages. *ReCALL*, Cambridge, v. 2, n. 2, p. 241-258, 2009. DOI: https://doi.org/10.1017/S0958344009000172

NIÑO, A. Evaluating the Use of Machine Translation Post-Editing in the Foreign Language Class. *Computer Assisted Language Learning*, [*S.l.*], v. 21, n. 1, p. 29-49, 2008. DOI: https://doi.org/10.1080/09588220701865482

NIÑO, A. Recycling MT: A Course on FL Writing Via MT Post-Editing. *In*: CLUK (COMPUTATIONAL LINGUISTICS UNITED KINGDOM) ANNUAL RESEARCH COLLOQUIUM, 7th., 2004, Birmingham. *Proceedings* [...]. Birmingham: University of Birmingham, 2004. p. 179-187.

NORDIN, N. M.; RASHID, S.; ZUBIR, S. I. S.; SADJIRIN, R. Differences in Reading Strategies: How ESL Learners Really Read. *Procedia-Social and Behavioral Sciences*, [*S.l.*], v. 90, p. 468-477, 2013. DOI: https://doi.org/10.1016/j.sbspro.2013.07.116

NUTTALL, C. *Teaching Reading Skills in a Foreign Language*. Oxford: Heinemann English Language Teaching, 1996.

O'BRIEN, S. Teaching Post-Editing: A Proposal for Course Content. *In*: EAMT WORKSHOP TEACHING MACHINE TRANSLATION, 6th., 2002, Manchester, UK. *Proceedings* [...]. Manchester: UMIST, 2002. p. 99-106.

OXFORD, R. L. *Language Learning Strategies:* What Every Teacher Should Know? New York: Newbury House Publishers, 1990.

PENA, M. I. C. The Potential of Digital Tools in the Language Classroom. *International Journal of the Humanities*, v. 8, n. 11, p. 57-68, 2011. DOI: https://doi.org/10.18848/1447-9508/CGP/v08i11/43047

PERFETTI, C. *Reading Ability*. New York: Oxford University Press, 1985.

PERFETTI C. Comprehending Written Language. A Blue Print of the Re. *In*: HAGOORT, P.; BROWN, C. (ed.) *Neurocognition of Language Processing*. Oxford: Oxford University Press, 1999. p. 167-208. DOI: https://doi.org/10.1093/acprof:oso/9780198507932.003.0006

PERFETTI, C.; HART, L. The Lexical Bases of Comprehension Skill. *In*: GORFIEN, D. (ed.). *On the Consequences of Meaning Selection*. Washington: Psychology American Association, 2001. p. 67-86. DOI: https://doi.org/10.1037/10459-004

PERFETTI, C. A.; LANDI, N.; OAKHILL, J. The Acquisition of Reading Comprehension Skill. In: SNOWLING, M. J.; HULME, C. (org.). *The Science of Reading*: A Handbook. Oxford: Blackwell Publishing, 2008. p. 227-247. DOI: https://doi.org/10.1002/9780470757642.ch13

PETRARCA, M.P. Machine Translation: A Tool for Understanding Linguistic Challenges Facing the Second Language Student. 2002. 240f. Thesis (Ph.D) – Indiana University of Pennsylvania, Ann Arbor, 2002. Available on: https://www.learntechlib.org/p/120926/. Retrieved at: Sept. 18, 2019.

RAMOS, R. ESP in Brazil: History, New Trends and Challenges. *In*: KRZANOWSKI, M. (org.). *English for Academic and Specific Purposes in Developing, Emerging and Least Developed Countries*. Reading: Garnet Publishing, 2009. p.68-84.

RAMOS, R. C. G., Gêneros textuais: uma proposta de aplicação em cursos de inglês para fins específicos. *The ESPecialist*, São Paulo, v. 25, n. 2, p 108-128, 2004.

RIESS, A. B. As *Estratégias de Leitura sem e com o uso do Google tradutor*. 2015. 220f. Tese (Doutorado em Letras) – Faculdade de Letras, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2015.

RIESS. A. B.; GABRIEL, R. A desambiguação lexical durante a compreensão leitora em inglês como língua estrangeira. *Letras de Hoje*, Porto Alegre, v. 54, n. 2, p. 181-190, abr.-jun. 2019. DOI: http://dx.doi.org/10.15448/1984-7726.2019.2.32529

ROBINSON, P. *ESP* - English for Specific Purposes. Oxford: Pergamon Press, 1980.

ROBINSON, P. *ESP Today:* A Practitioner's Guide. Hertfordshire: Prentice Hall International, 1991.

RUMELHART, D. Toward an Interactive Model of Reading. *In*: SINGER, H.; RUDDELL, R. (ed.). *Theoretical Models and Processes of Reading*. Newark DE: International Reading Association, 1985.

SARMENTO, S.; BAUMVOL. L. A Internacionalização em casa e o uso de Inglês como Meio de Instrução. *Echoes*, Florianópolis, s. n., p. 66-82, 2016.

SARMENTO, S.; BAUMVOL, L. K.; MARTINEZ, R. O papel das línguas na internacionalização da educação. *Organon*. Porto Alegre, v. 34, n. 66, p. 1-3, 2019.

SCARAMUCCI, M. A competência lexical de alunos universitários aprendendo a ler em inglês como língua estrangeira. *DELTA*, São Paulo, v. 13, n. 2, p. 215-246, 1997. DOI: https://doi.org/10.1590/S0102-44501997000200003

SOMERS, H. Machine Translation: Latest Developments. *In*: MITKOV, R. (org.). *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press, 2003. p. 513-528.

SOMERS, H. Further Experiments in Bilingual Text Alignment. *International Journal of Corpus Linguistics*, Amsterdam, v, 3, p. 1-36, 1998. DOI: https://doi.org/10.1075/ijcl.3.1.06som

STEDING, S. Machine Translation in the German Classroom: Detection, Reaction, Prevention. *DieUnterrichtspraxis/Teaching German*, [*S.l.*], v. 42, n. 2, p. 178-189, 2009. DOI: https://doi.org/10.1111/j.1756-1221.2009.00052.x

TOMITCH, L. M. B. Aquisição de leitura em língua inglesa. *In*: LIMA, D.C. (org.). *Ensino e aprendizagem de língua inglesa*: conversas com especialistas. São Paulo: Parábola Editorial, 2009. p. 191-201.

WILLIAMS, L. Web-Based Machine Translation as a Tool for Promoting Electronic Literacy and Language Awareness. *Foreign Language Annals,* [*S.l.*], v. 39, n. 4, p. 565-578, 2006. DOI: https://doi.org/10.1111/j.1944-9720.2006.tb02276.x

ZANETTIN, F. Corpus-Based Translation Activities for Language Learners. The Interpreter and Translator Trainer, [*S.l.*], v. 3, n. 2, p. 209-224, 2009. DOI: https://doi.org/10.1080/1750399X.2009.10798789

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 859-907, 2021

CC BY

An investigation of linguistic problems in automatic multi-document summaries

Uma investigação de problemas linguísticos em sumários automáticos multidocumento

Márcio de Souza Dias
Universidade Federal de Goiás (UFG), Catalão, Goiás / Brasil
marciosouzadias@ufg.br
http://orcid.org/0000-0003-1116-6965

Ariani Di Felippo
Universidade Federal de São Carlos (UFSCar), São Carlos, São Paulo / Brasil
arianidf@gmail.com
http://orcid.org/0000-0002-4566-9352

Amanda Pontes Rassi
Redação Nota 1000 Ltda., São Paulo, São Paulo / Brasil
amandarassi85@gmail.com
http://orcid.org/0000-0001-5314-1868

Paula Christina Figueira Cardoso
Universidade Federal de Lavras (UFLA), Lavras, Minas Gerais / Brasil
paula.cardoso@ufla.br
http://orcid.org/0000-0003-3621-8960

Fernando Antônio Asevedo Nóbrega
Samsung, São Paulo, São Paulo / Brasil
fernandoasevedo@gmail.com
http://orcid.org/0000-0002-1129-0133

Thiago Alexandre Salgueiro Pardo
Universidade de São Paulo (USP), São Carlos, São Paulo / Brasil
taspardo@icmc.usp.br
http://orcid.org/0000-0003-2111-1319

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.859-907

Abstract: Automatic summaries commonly present diverse linguistic problems that affect textual quality and thus their understanding by users. Few studies have tried to characterize such problems and their relation with the performance of the summarization systems. In this paper, we investigated the problems in multi-document extracts (i.e., summaries produced by concatenating several sentences taken exactly as they appear in the source texts) generated by systems for Brazilian Portuguese that have different approaches (i.e., superficial and deep) and performances (i.e., baseline and state-of-the art methods). For that, we first reviewed the main characterization studies, resulting in a typology of linguistic problems more suitable for multi-document summarization. Then, we manually annotated a corpus of automatic multi-document extracts in Portuguese based on the typology, which showed that some of linguistic problems are significantly more recurrent than others. Thus, this corpus annotation may support research on linguistic problems detection and correction for summary improvement, allowing the production of automatic summaries that are not only informative (i.e., they convey the content of the source material), but also linguistically well structured.

Keywords: automatic summarization; multi-document summary; linguistic problem; corpus annotation.

Resumo: Sumários automáticos geralmente apresentam vários problemas linguísticos que afetam a sua qualidade textual e, consequentemente, sua compreensão pelos usuários. Alguns trabalhos caracterizam tais problemas e os relacionam ao desempenho dos sistemas de sumarização. Neste artigo, investigaram-se os problemas em extratos (isto é, sumários produzidos pela concatenação de sentenças extraídas na íntegra dos textos-fonte) multidocumento em Português do Brasil gerados por sistemas que apresentam diferentes abordagens (isto é, superficial e profunda) e desempenho (isto é, métodos *baseline* e do estado-da-arte). Para tanto, as principais caracterizações dos problemas linguísticos em sumários automáticos foram investigadas, resultando em uma tipologia mais adequada à sumarização multidocumento. Em seguida, anotou-se manualmente um corpus de extratos com base na tipologia, evidenciando que alguns tipos de problemas são significativamente mais recorrentes que outros. Assim, essa anotação gera subsídios para as tarefas automáticas de detecção e correção de problemas linguísticos com vistas à produção de sumários automáticos não só mais informativos (isto é, que cobrem o conteúdo do material de origem), como também linguisticamente bem-estruturados.

Palavras-chave: sumarização automática; sumário multidocumento; problema linguístico; anotação de *corpus*.

Submitted on April 29th, 2020

Accepted on July 1st, 2020

1 Introduction

Multi-document Summarization (MDS) is an important area of Natural Language Processing (NLP). It aims at automatically producing a unique summary for a set of source texts on the same topic (MANI, 2001; NENKOVA; MCKEOWN, 2011). It currently has attracted a lot of attention in the scientific community because of the increasing incredible amount of available textual information nowadays, mainly on the web.

It is a consensus that a good summary should contain the most relevant information in the texts, and the area has achieved significant progress in producing summaries that are more informative. The progress is the result of both linguistically poor and rich summarization methods, such as the empirical/statistical approaches (see, e.g., ANDO *et al*., 2000; CARBONELL *et al*., 1997; HAGHIGHI; VANDERWENDE, 2009; MIHALCEA; TARAU, 2005; RIBALDO *et al*., 2016) and the deep ones (CARDOSO; PARDO, 2016; CASTRO JORGE; PARDO, 2010; MCKEOWN; RADEV, 1995; RADEV, 2000; ZHANG *et al*., 2002).

Automatic summaries must also present the information to the reader in a cohesive and coherent way. According to Koch (1998), cohesion is related to the surface organization of a text. It may be expressed by successive links among elements in the superficial structure of the text. For example, anaphoric pronouns, which refer back to textual antecedents, are elements of cohesion. Coherence is related to the meaning of a text; related to the possible interpretation of the text (KOCH; TRAVAGLIA, 2002). Beaugrande and Dressler (1981) claim that the continuity of meaning is what keeps the text coherent. Thus, coherence is the combination of concepts and relations of textual elements and, sometimes, it is necessary to make use of world knowledge and knowledge about the interlocutors and the situation itself for the text to make sense. For example, coherence can be created between sentences through repetition of words, which helps to reiterate the same ideas.

Although current summarization methods are still limited on such aspects, since most of the systems only produce extractive[1] instead of abstractive summaries[2] (which are still hard to achieve and not fully understood, systematized and formalized). Trying to evaluate

[1] Summaries produced by concatenating sentences taken exactly as they appear in the source texts.

[2] Summaries that allow rewriting operations over the original material.

the linguistic quality (LQ) of summaries through numeric scores using lexical, syntactic and/or semantic features (see, e.g., CONROY *et al.*, 2011; GIANNAKOPOULOS; KARKALETSIS, 2011; LIN *et al.*, 2012; OLIVEIRA, 2011; PITLER *et al.*, 2010) or to identify certain problematic linguistic aspects (see, e.g., CRISTINI; DI-FELIPPO, 2019; FONSECA *et al*., 2019; FRIEDRICH *et al.*, 2014; PITLER *et al.*, 2010), the summarization literature has revealed that automatic extracts present several problems that affect their LQ.

In order to propose specific solutions for improving the LQ of automatic summaries or more sophisticated MDS methods that tackle such issues it is necessary to identify and to characterize the problems in a corpus of automatic summaries.

In this paper, we investigate the types of LQ problems that affect multi-document summary quality. Initially, we reviewed the main approaches in the literature of linguistic problems in automatic summaries, resulting in a typology more suitable for the multi-document scenario. Next, we used the typology to annotate a corpus of extractive multi-document summaries in Brazilian Portuguese[3] produced by systems with different performances, from both superficial (that use little linguistic knowledge) and deep approaches (which are based on sophisticated linguistic knowledge, as semantics and discourse), including baseline and state-of-the art methods. Finally, with the annotated corpus, we systematized and characterized the problems that the systems produce and show that some problems are significantly more recurrent than others.

In Section 2, we present an overview of basic concepts in multi-document summarization, focusing on the methods used to produce the summaries that we evaluated. Section 3 presents the linguistic problems that are available in the literature, resulting in a typology of problems. In Section 4, we present our corpus of summaries used in the annotation. In Section 5, we detailed the annotation of linguistic problems in the multi-document summaries in Portuguese. Section 6 shows the results and the analysis of the error annotation. In Section 7, the final remarks will be presented.

[3] The LQ problems are generic and may be applied to any language.

2 Automatic Summarization

In this section, we present an overview of basic concepts in Automatic Summarization and methods developed specifically for generating summaries in Brazilian Portuguese.

2.1 Basic concepts

According to Mani (2001), a summary is a shorter version of one or more texts. Depending on the number of documents to summarize, the automatic process is defined as single or multi-document summarization. While the first dates back to the 50s, the latter, which is the focus of this paper, consists in a more recent initiative that officially started in the 90s, bringing new challenges to the Automatic Summarization area.

There are several possible classifications for summaries (see, e.g., MANI; MAYBURY, 1999). Summaries may be informative, indicative or critical. Informative summaries include the main facts of the source documents organized in a cohesive and coherent way. These summaries can be read in place of the original texts. Indicative summaries, differently from the informative ones, do not substitute the original texts, but only indicate what the texts are about. For example, indexes may be classified as indicative summaries. Critical summaries bring the authors' opinions or points of view about the source texts. Examples of critical summaries are book reviews.

Summaries are also classified according to the intended audience. Generic summarization does not take into account any specific interest of the reader, producing general-purpose summaries. On the other hand, summarization focused on the interest of the reader uses information based on his/her prior knowledge and interests. For example, a layman may need a summary with more contextual information about the subject, while a reader with a good knowledge about the subject may expect that the summary presents additional or new information.

Summaries may be classified as extractive or abstractive. Extractive summaries are formed by pieces of non-modified text, with copy and paste operations (from the source texts to the summaries), basically. Abstractive summaries make use of rewriting operations, i.e., there is some or full modification in the structure and/or in the writing of the source text passages for building the corresponding summaries. Currently, most of the available automatic summarizers are extractive since abstraction is still considered a very difficult task.

The construction of summaries may follow two linguistic approaches: superficial/shallow and deep approaches (MANI, 2001). Shallow approaches use little or no linguistic knowledge at all to produce summaries. The main advantage of the shallow approach is its robustness[4] and scalability,[5] but it may produce worse summaries than the ones resulting from deep approaches. Deep approaches use linguistic knowledge, theories and formal language models in the creation of summaries, as lexicons, wordnets, grammars, and syntactic-semantic and discourse analysis. This approach is considered the most complex one, because of the number of linguistic variables. Its application is usually limited since systems of this approach are mostly developed for specific domains. Shallow and deep approaches may also be merged, resulting in the hybrid approach.

Finally, another important concept in summarization is the amount of information that will be included in the summaries, which is determined by the compression rate, i.e., the ratio between the size of the summary and the size of the source texts (MANI, 2001), usually measured in number of words.

In this paper, we conduct our investigation with extractive, informative and generic summaries (which consist in the most usual configuration in the area), produced by both shallow and deep approaches for Portuguese. We briefly introduce the main characteristics of the summarization methods that we used in what follows.

2.2 Summarization methods for Portuguese

There are several multi-document summarization systems for Portuguese, following different content selection strategies, using both classical and state of the art methods in the area. For this investigation, we have selected four of them, trying to get a sample of summaries of different performances, which represent the main available approaches.

One of them was GistSumm (GIST SUMMarizer) (PARDO *et al*., 2003; PARDO, 2005). This summarizer follows a simple shallow approach, and, to the best of our knowledge, it was the first one made available for Portuguese. Its approach is based on the gist of the source

[4] In this case, a robust method is applicable to very different testing data, e.g., different genre or domain.

[5] The scalability represents the ability of the method to deal with large amount of data.

texts, i.e., the main idea intended to be conveyed or understood by the reader. The gist is the most important segment of the source texts, commonly expressed by only one sentence. The most widely applied technique for detecting it has been simple word frequency measures. Once identified, the gist serves as guide for identifying and selecting other sentences to compose the final extract. Figure 1 shows a summary generated by GistSumm.

FIGURE 1 – Summary generated by GistSumm

[S1] The crimes happened in the city of Muttur, in which during the last two weeks, there were severe conflicts between the troops of the Sri Lanka army and the guerrillas of the Liberation Tigers of Tamil Eelam (LTTE).

[S2] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO "did not suffer a similar loss in over 25 years of existence."

[S3] The violent conflict started on July 26, when government air troops bombed positions of the guerrillas after the rebels blocked a dam located in its territory for more than a week, hindering the supply of water in places under the government control.

[S4] The special envoy for the peace in Sri Lanka from Norway, Jon Hanssen-Bauer, arrived in the island last week and met the two parties, attempting to reduce the tension and to avoid a new start of the civil war.

[S5] The crimes happened in the city of Muttur, in which, during the last two weeks, there were severe conflicts between the troops of the Sri Lanka army and the guerrillas of the Liberation Tigers of Tamil Eelam (LTTE).

[S6] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO "did not suffer a similar loss in over 25 years of existence."

[S7] The special envoy for the peace in Sri Lanka from Norway, Jon Hanssen-Bauer, arrived in the island last week and met the two parties, attempting to reduce the tension and to avoid a new start of the civil war.

[S8] Fifteen local employees of a French charity institution in Sri Lanka were found dead in the city of Muttur in the north of the country.

One may see that the summary has several problems, such as redundant information (S1 with S5, S2 with S6, and S4 with S7), noun phrases without explanation (e.g., "the crimes" in S1 is not specified or explained), and acronyms without explanation ("ACF" and "NGO" in S2). Such problems occur due to the simplicity of GistSumm, whose

method is considered a baseline method for Portuguese. It was included in this investigation for historical reasons and to evidence improvements and remaining problems that the best current methods show.

The RSumm summarizer (RIBALDO *et al.*, 2012, 2016) is based on classical graph-based methods, which use the relationship map approaches of Salton *et al.* (1997) adapted for MDS. According to the authors, graphs/maps are built from a set of documents on the same topic, where each vertex represents a sentence and the edges indicate the lexical similarity between the sentences. The best method groups topic-related sentences and select the most relevant one from each subtopic to compose the summary. Figure 2 shows an example of a summary generated by RSumm for the same source texts of the summary in Figure 1. One may see that problems still happen in the summary, mainly related to the proper introduction of noun phrases. However, it is clear that this summary is much better than the one produced by GistSumm.

FIGURE 2 – Summary generated by RSumm

[S1] The special envoy for the peace in Sri Lanka from Norway, Jon Hanssen-Bauer, arrived in the island last week and met the two parties, attempting to reduce the tension and to avoid a new start of the civil war.

[S2] Fifteen local employees of a French charity institution in Sri Lanka were found dead in the city of Muttur in the north of the country.

[S3] The crimes happened in the city of Muttur, in which, during the last two weeks, there were severe conflicts between the troops of the Sri Lanka army and the guerrillas of the Liberation Tigers of Tamil Eelam (LTTE).

Cardoso and Pardo (2015, 2016) presented a deep method for MDS. They assume that the relevance of a sentence is influenced by its salience in its source text, which is given by Rhetorical Structure Theory (RST) (MANN; THOMPSON, 1987), using the method proposed by Marcu (1999), and its salience in the set of texts, given by Cross-document Structure Theory (CST) (RADEV, 2000). The method is referred by RC-4 (which stands for the "**4**th combination of **R**ST and **C**ST information"). Figure 3 shows a summary generated by RC-4.

FIGURE 3 – Summary generated by RC-4

[S1] Fifteen volunteers from the French NGO "Action Contre la Faim" (ACF) were killed in northeastern Sri Lanka today, said a spokeswoman

[S2] According to a representative of the group Action Contre la Faim, the bodies were found in the organization office.

[S3] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO did not suffer a similar loss in over 25 years of existence.

[S4] Up to now, the Sri Lankan authorities did not confirm the deaths or clarified what happened in the city of Muttur.

[S5] The rebels said that they will consider a new bombing of the army.

This summary is much better than the others, but it still presents some problems, such as lack of connection between the S5 content and the rest of the summary, and occurrence of the noun phrases "The rebels" and "a new bombing of the army" that do not have their respective referents in the summary.

The last summarizer is based on a statistical method (CASTRO JORGE, 2015). It captures summarization patterns by estimating the occurrence probability of some features in human summaries, including, e.g., discourse (following the RST and CST models) and sentence position information. The features represent strategic characteristics that indicate the salience of a sentence among a set of sentences. The probabilistic model is based on a generative learning approach (the noisy-channel framework), where the task is formulated with probabilistic components, including probabilities for content selection during the transformation process and for coherence of the produced summary, and a decodification step (i.e., the production of the final summary). This summarization method is referenced by MTRST-MCAD (Method of Transformation with RST and Model for Coherence evaluation After Decodification). Figure 4 shows an example of a summary created by the MTRST-MCAD method.

FIGURE 4 – Summary generated by MTRST-MCAD

[S1] It is unclear who committed the murders of the employees of the French organization.

[S2] The rebels said that they will consider a new bombing of the army.

[S3] Up to now, the Sri Lankan authorities did not confirm the deaths or clarified what happened in the city of Muttur.

[S4] "We tried to send a team to Muttur to check what is going on, but the soldiers did not allow us to enter the city, which is totally blocked", he said.

[S5] The director of ACF in Sri Lanka, Benoit Miribel, confirmed the death of its employees and said that the NGO did not suffer a similar loss over 25 years of existence.

One may see that the summary also has some problems that affect its quality, such as the lack of connection between S2 content and the rest of the summary, and the occurrence of the definite noun phrases "the murders of the employees" and "the French organization" in S1 that do not have their respective referents. The same occurs with the definite noun phrase "The rebels" and "the army" in S2. Besides these problems, the explanations for the "ACF" and "NGO" acronyms in S5 are not present in the summary.

The RC-4 system (in the deep approach) is currently the best method for Portuguese, followed very closely by RSumm (in the shallow approach). With some distance, we have MTRST-MCAD and, finally, GistSumm. The evaluations of these methods have so far been guided by summary informativeness criteria, mainly using ROUGE (LIN, 2004), a standard n-gram-based measure that is automatically computed, allowing for fast and easily reproducible evaluation. Despite the importance of informativeness, the examples in this section show that this criterion is not enough for assuring that good summaries are produced and provide evidence that the systems need to treat problems that affect the LQ of their summaries, as they severely harm the summary quality. For this, we believe that the definition and the identification of problems related to LQ will guide the summarizers in possible solutions for these problems.

In what follows, we present and discuss important issues and previous initiatives related to defining and characterizing linguistic problems in summaries, proposing, in the end, a synthesized and comparative view of them. This forms the basis of the study that we conduct in our corpus.

3 Definition and characterization of linguistic problems

Some works have tried to find and deal with linguistic problems in summaries for improving their quality. Although some identified problems are similar, some approaches are much more refined than others and there is great variation in the error catalogues. To the best of our knowledge, we briefly list and discuss the main initiatives in what follows.

3.1 The revision of linguistic quality issues in automatic summaries

Otterbacher *et al.* (2002) studied the problems related to the cohesion of extractive multi-document summaries and suggested revisions (solutions) to improve cohesion. The authors presented a corpus-based analysis of automatically generated extractive multi-document summaries, produced by the MEAD summarizer (RADEV *et al.*, 2003), which is one of the most popular summarization systems for English. The authors discussed the feasibility of automatically improving the summaries and they created a taxonomy of problems related to cohesion.

According to them, the taxonomy is divided into five pragmatic categories related to textual cohesion in multi-document summaries: *Discourse*, *Identification of Entities*, *Temporal Expressions*, *Grammar*, and *Location Settings*. In what follows, we detail these problems and some of their main related problems, showing examples.

The *discourse* category focuses on the relationships among the sentences of the summary (inter-sentence level) and on the relationships among textual elements inside sentences (intra-sentence level). The authors considered some aspects in this category that may cause cohesion problems in multi-document summaries: *Topic Shift*, *Lack of Purpose*, *Contradiction*, *Redundancy*, and *Conditional Sentences*.

The *Topic Shift,* which is the fast change of one subject by another, has the highest occurrence (45%). In order to solve the problem, an addition of a transitional sentence or phrase may be necessary, as illustrated in Figure 5. The underlined segment is a possible example of transitional phrase in a *Topic Shift.*

FIGURE 5 – Example of solution for *topic shift* problem

[S1] <u>In a related story</u>, the government of Hong Kong announced a proposal to require all drug rehabilitation centers...

Source: Otterbacher *et al*. (2002)

Another common problem in summaries is sentences with *lack of purpose*, which may be solved by the addition of sentences or phrases that motivate a purpose in the problematic segment. Figure 6 shows this situation.

FIGURE 6 – Example of solution for *lacked purpose*

[S1] In order to assist the ongoing investigation as the cause of the crash, the U.S. team from the National Transportation Safety Board will join experts...

Source: Otterbacher *et al*. (2002)

Contradiction is related to some information in a given sentence that contrasts with one or more previous sentences. In such cases, a discourse marker such as "however" or "in contrast" may help. Figure 7 shows an example of contradiction.

FIGURE 7 – Example of *contradiction* that was solved

[S1] However, according to reports on CNN, the control tower was concerned with the speed and altitude of the plane and had discussed these concerns with the pilot.

Source: Otterbacher *et al*. (2002)

Redundancy occurs when a sentence contains previously reported information. For Otterbacher *et al*. (2002), a possible action to solve this problem is to delete the redundant constituent (non-head element of NPs, PPs, or the entire relative clause or phrase). Figure 8 shows an example of this scenario, where the underlined passage must be removed.

FIGURE 8 – Example of *redundancy* that may be solved

[S1] The crash of flight 072 that killed 143 people…
[S2] The plane, which was carrying the 143 victims, was headed for Bahrain from Egypt.

Source: Otterbacher *et al*. (2002)

According to the authors, sometimes events in a given sentence are *conditioned* by events in another sentence. Thus, a good action is to modify the sentences, using the structure "IF (sentence 1), (sentence 2)".

Besides this, the verb tenses may be changed to represent the condition. Figure 9 is an example of this use.

FIGURE 9 – Example of *conditional sentence* with improved cohesion

[S1] If the proposed measures were implemented, they would ensure broadly the same registration standard to be applied to all drug treatment centers.

Source: Otterbacher *et al*. (2002)

The *identification of entities* category requires the resolution of referential expressions, since the reader needs to identify each entity mentioned in a summary. According to Otterbacher *et al*. (2002), 9 problems were found in summaries related to this category, which were: *Underspecified Entity*, *Misused Quantifier*, *Overspecified Entity*, *Repeated Entity*, *Bare Anaphora*, *Misused Definite Article*, *Misused Indefinite Article*, *Missing Article*, and *Missing Entity*. The *underspecified entity* problem was the most frequent in this category, in 38% of the cases.

The authors also use some revisions to solve problems related to the identification of entities. For example, one possible solution to solve an *underspecified entity* (a newly mentioned entity that has no description, or the presence of an acronym without explanation) is the addition of a full name, a description or a title for the new entity, or expanding the acronym if this is the case. Figure 10 shows an example of this revision.

FIGURE 10 – Example of *underspecified entity* revision

[S1] Mrs. Clarie Lo, the Commissioner of Narcotics, said the proposal would be introduced to non-medical drug treatment centers.

Source: Otterbacher *et al*. (2002)

The *misused definite article* problem may also be solved by adding a definite article if the entity has already been mentioned, or an indefinite article if the entity is new. Figure 11 shows part of a text with the addition of the indefinite article "*a*", since the entity "*second eruption"* is new in the text.

FIGURE 11 – Example of *misused definite article* revision

[S1] On Thursday, a second eruption appeared to be smaller than anticipated.

Source: Otterbacher *et al*. (2002)

The *temporal* category is related to the right temporal relationships among events. The authors identified five types of possible problems that fall into this category: *Temporal Ordering*, *Time of Event*, *Event Repetition, Synchrony* and *Anachronism*. The *temporal ordering* problem represented 89% of all errors found in this category.

Temporal ordering is related to the establishment of correct temporal relations among events. If there is a problem, the authors recommend, e.g., to add time expressions, to add ordinal numbers, to delete inappropriate time expressions, or to modify an existing time expression. Figure 12 shows an example of a temporal ordering problem that was revised.

FIGURE 12 – Example of revision for *temporal ordering* error

[S1] Two days later, a second eruption appeared to be smaller than scientists had anticipated.

Source: Otterbacher *et al*. (2002)

The *event repetition* problem may be solved by simply adding an adverb such as "again". Figure 13 shows an example of such revision.

FIGURE 13 – Example of *event repetition* problem revision

[S1] Mount Pinatubo is likely to explode again in the next few days or weeks.

Source: Otterbacher *et al*. (2002)

Some problems in *grammar* category have also been identified in the corpus used by Otterbacher *et al*. (2002) Among these problems are: *Run-on Sentence*, *Mismatched Verb*, *Missing Punctuation*, *Awkward Syntax*, *Parenthetical*, *Subheadings/Titles*, and *Misused Adverb*. The *run-on sentence* problem was the most frequent one, representing 35% of these errors.

For the authors, a *run-on sentence* is a very long sentence. Thus, the authors recommend splitting long sentences into two separate sentences and deleting the conjunction. Figure 14 shows a long sentence that was revised.

FIGURE 14 – Example of *run-on sentence* problem revision

[S1] Lt. Col. Ron Rand announced at 5 a.m. Monday that all personnel should begin evacuating the base. [S1] Meanwhile, dawn skies over central Luzon were filled…

Source: Otterbacher *et al*. (2002)

Parenthetical is a problem related to the inappropriate use of parenthesis. Thus, the authors simply suggest deleting the parenthesis symbols. Figure 15 shows an example of inappropriate use of parenthesis.

FIGURE 15 – Example of a *parenthetical* problem revised

[S1] (Volcanoes such as Pinatubo arise where one of the earth's crust plates is slowly diving beneath another.)

Source: Otterbacher *et al*. (2002)

The *location settings* category includes a type of revision related to the correct location of events, in order for the text to be improved. These settings may be: *Location of Event*, *Collocation*, *Change of Location*, and *Place/Source Stamp*.

Location of event specifies where an event takes place. Thus, the authors suggest adding a prepositional phrase that indicates place (city, state, or country). Figure 16 shows a type of *location of event* setting that was revised.

FIGURE 16 – Example of a *location of event* setting revision

[S1] Three bodies were lain before the faithful in the Grand Mosque <u>in Manama, Bahrain</u> during a special prayer…

Source: Otterbacher *et al*. (2002)

Collocation is related to two or more events that occur in the same place. Thus, the authors suggest adding a prepositional phrase or an adverb that indicates the collocation. An example is shown in Figure 17.

FIGURE 17 – Example of revision for *collocation*

[S1] Meanwhile, <u>in the same area</u>, search teams sifted through the wreckage.

Source: Otterbacher *et al*. (2002)

Generally, according to the authors, the *discourse* category corresponded to 34% of all the problems found in the corpus, followed by the categories *identification of entities* (with 26%), *temporal expressions* (22%), *grammar* (12%), and *location settings* (6%).

Friedrich *et al*. (2014) presented a corpus of multi-document summaries (called LQVSumm) which was manually annotated with several types of LQ errors. These summaries were automatically created in the TAC (Text Analysis Conference) 2011 shared task on Guided Summarization (OWCZARZAK; DANG, 2011). The authors identified two classes of problems: one considering entity mentions and another happening at the level of clauses. The first is related to reference or coreference problems. The last involves grammar or redundancy errors.

For the authors, in the level of entity, the problem types are: *First mention without explanation*, *Subsequent mention with explanation*, *Definite noun phrase without reference to previous mention*, *Indefinite noun phrase with reference to previous mention*, *Pronoun with missing antecedent*, *Pronoun with misleading antecedent*, and *Acronyms without explanations*.

The *first mention without explanation* problem is assigned to the first mention of an entity for which there is not a clear reference to the reader. For example, in the sentence "Paul bought toys to the poor children", there is no sufficient introduction for the entity "Paul".

The *subsequent mention with explanation* problem is related to entity mentions that have already been referenced in the text and present an inappropriate extra explanation. For example, consider sentences S1 and S2 in Figure 18. In sentence S2, there is an additional explanation related to the entity Taylor, but the entity has already been referenced in sentence S1.

FIGURE 18 – Example of *subsequent mention with explanation* error

[S1] Taylor's attorney could not be reached for comment Friday night.
[S2] Tony Taylor, 34, of Hampton, Va., has a plea-agreement hearing scheduled for 9 a.m.

Source: Friedrich *et al*. (2014)

The *definite noun phrase without reference to previous mention* problem occurs when a definite noun phrase is used to refer to the first

mention of an entity in the text. For example, "the Petrobras Company" should be used in a summary in which "a company" has been mentioned before.

The *indefinite noun phrase with reference to previous mention* error occurs when an indefinite noun phrase is used for an entity already mentioned in the discourse. For example, the noun phrase "*a company*" is not appropriate if the same company has already been mentioned in the summary.

The *pronoun with missing antecedent* problem occurs when there is no possible antecedent that matches with the pronoun. Figure 19, for example, shows a beginning of an automatic multi-document summary where the pronoun "he" does not have a possible antecedent.

FIGURE 19 – Example of *pronoun with missing antecedent*

[S1] The renouncement may not stop the investigation because the process was already started.
[S2] He will establish the process against the deputies involved with the Sanguessugas Mafia.

Source: Cardoso *et al*. (2011)

The *pronoun with misleading antecedent* error occurs when an anaphoric expression refers to a misleading antecedent and its right antecedent is not in the summary. For example, Figure 20 shows part of a summary about soccer. In this case, the pronoun "he" (in the second sentence) apparently refers to the soccer player Kaká (in the first sentence), but, in the source text, the pronoun refers to Robinho, who is not introduced in the summary.

FIGURE 20 – Example of *pronoun with misleading antecedent*

[S1] At the 27 minutes, Kaká kicked the ball and Ronaldinho diverted the kick.
[S1] 20 cm from the end line, he gave two humiliating dribbles in the Ecuadorian defender and crossed the ball to Elano, who scored the fourth goal, at 37 minutes.

Source: Cardoso *et al*. (2011)

The *acronyms without explanations* problem occurs when acronyms are not previously known and are not explained in the first time they are introduced.

Friedrich *et al*. also proposed the annotation at the clause level. This was made on arbitrary spans, from single tokens to complete sentences. According to the authors, the clause level errors are: *Incomplete sentence*, *Inclusion of datelines*, *Other ungrammatical form*, *No semantic relatedness*, *Redundant information*, and *No discourse relation*.

An *incomplete sentence* problem usually results from segmentation errors in sentence compression (or truncation), which aims at reducing the length of candidate sentences to generate summaries with the desirable size pre-defined by the compression rate. For example, the following sentence is incomplete, since the name of the person was lost in the end of the sentence: "One was killed in a bedroom and others were murdered in a classroom, according to the head of the campus police, W."

For the authors, the *inclusion of datelines* in summaries is not desired and should be avoided. For example, a summary with the information "GEORGETOWN, Pennsylvania 2006-10-05 16:53:53 UTC" must be annotated with this problem.

The *other ungrammatical form* error considers all other ungrammaticality cases, such as missing spaces and wrong punctuation.

The *no semantic relatedness* problem occurs when sentences do not show plausible semantic relations. In Figure 21, for example, S1 and S2 are apparently not related.

FIGURE 21 – Example of *no semantic relatedness* problem

[S1] It is popularly known as the 'pink city' because of the ochre-pink hue of its old buildings and crenellated city walls.
[S2] He said there was no justification for such killings.

Source: Friedrich *et al*. (2014)

The *redundant information* problem occurs when two or more sentences express the same information. For example, in Figure 22, sentences S1 and S2 are partially redundant.

FIGURE 22 – Example of summary with *redundant information*

[S1] The suspect apparently called his wife from a cell phone shortly before the shooting began, saying he was "acting out in revenge for something that happened 20 years ago", Miller said.
[S2] The gunman, a local truck driver Charles Roberts, was apparently acting in "revenge for an incident that happened to him 20 years ago.

Source: Friedrich *et al*. (2014)

The *no discourse relation* problem, in particular, may happen when an explicit discourse connective (e.g., "and", "but", "even though" and "because") is no longer appropriate in the new context in the summary, does not being suitable for signaling the corresponding discourse relation. For example, this is the case for the connective "and" in the second sentence in Figure 23.

FIGURE 23 – Example of *no discourse relation* error in a summary

[S1] Taylor's attorney could not be reached for comment Friday night.
[S2] And the person who cooperates first gets the biggest reward.

Source: Friedrich *et al*. (2014)

It their conclusions, the authors show that there are relationships between the types of problems they defined and the summary readability evaluation performed at TAC, which we introduce in what follows.

In the mono-document summarization, Kaspersson *et al*. (2012) investigated linguistic problems that occur in summaries extracted from single texts. The focus was on discourse problems, such as referring expressions with missing antecedents and fragments, and how text units in the summaries are connected. In addition, the authors have investigated how the different size of summaries and different genres influence the occurrence of types of problems. The authors considered texts of three different genres in their study: Swedish newspapers, popular Swedish science texts, and authority texts from the Swedish Social Insurance Administration.

The problems found by the authors were grouped into three categories: *Erroneous anaphoric reference*, *Absent cohesion or context*, and *Broken anaphoric reference*. *Erroneous anaphoric reference* is related to an anaphoric expression in the summarized text that refers to an erroneous antecedent, given that the correct antecedent was not extracted from the source text of the summary. This category occurs for the following cases: *Noun phrases*, *Proper names*, and *Pronouns*. *Absent cohesion or context* is a self-explanatory error, related to the lack of cohesion or necessary context in summaries. *Broken anaphoric reference* happens when an anaphoric expression presented in a summary does not have its antecedent because this antecedent was not extracted

from the source text. This category also occurs for the following cases: *Noun phrases*, *Proper names*, and *Pronouns*.

The authors report that the most significant problems are: *Erroneous anaphoric reference related to pronoun*, *Absent cohesion or context*, *Broken anaphoric references related to noun phrases* and *Broken anaphoric references related to pronouns*.

For evaluating summaries in summarization contests, TAC (DANG, 2005) developed classical guidelines to evaluate LQ in summaries related to 5 features: *Grammaticality*, *No Redundancy*, *Referential Clarity*, *Textual Focus*, and *Textual Structure and Coherence*.

Grammaticality verifies whether there are format and grammar problems in the summaries, including capitalization (e.g., whether proper names start with a capital letter). In relation to *no redundancy*, a good summary should present the maximum amount of unique information that is possible in respect to the compression rate. Thus, a summary is weighted by the unnecessary repetition of information. This analysis must happen in different levels, such as the redundant data/fact of an event, sentences, and names (entities should be, whenever possible, referenced by pronouns). A summary presents *referential clarity* when text references are not ambiguous. A summary has *focus* when all sentences are related to the addressed issue. The last feature of TAC suggests that a summary is also evaluated by its good *structuring and coherence*. For example, a summary should not present divergent information on the same fact or event.

These 5 criteria that were proposed in TAC (actually, when it was named Document Understanding Conference (DUC)) are widespread in the area and used by most of the works that attempt to check LQ in summaries.

3.2 A synthesized view of linguistic quality issues

In section 4.1, we reviewed the more important sets of LQ problems in automatic summaries defined by previous research. Such sets present similarities and differences in several aspects, such as (i) coverage, since some problem sets are more complete than others; (ii) types of problems, (iii) generality of the problems (since some problem sets are more fine-grained than others), and (iv) purpose (some errors are tailored for single summarization, others are for MDS, and others are

more agnostic). This shows the relevance and the complexity of these studies, which support summarization and other tasks.

In Table 1, we synthesized the LQ problem sets, showing the similarities and differences based on 5 classes: (i) errors related to inappropriate formatting and metadata inclusion; (ii) problems with grammatical origin; (iii) inadequacies that come from style/grammar choices; (iv) problems related to inadequacies in the use of entities and, therefore, also related to cohesion; and (v) errors related to discourse and coherence. We indicate with an "X" when a study treats the respective LQ issue.

It is clear that some problem types cause problems in other levels (e.g., a grammar error of missing subject/agent in a sentence also results in lower cohesion), but we focused on the origin of the problems when categorizing them. It is also interesting to notice that such categorization may not be completely fair to the listed works, as they report different problem specificity levels: while Otterbacher *et al*. (2002) and Friedrich *et al*. (2014) present much more refined error catalogues, Kaspersson *et al*. (2012) and Dang (2005) are more worried with general level problems.

TABLE 1– Synthesis of LQ problems in summaries

LQ problems	**Otterbacher *et al*. (2002)**	**Kaspersson *et al*. (2012)**	**Friedrich *et al*. (2014)**	**Dang (2005)**
Formatting, metadata				
inclusion of subheading/titles	x			
inclusion of place/source stamp	x			
inclusion of datelines			x	x
inclusion of system-internal formatting				x
Grammar				
missing subject/agent	x		x	x
mismatched verb	x		x	x
missing punctuation	x		x	x
wrong parenthetical	x		x	x
incomplete sentence			x	x
wrong capitalization				x

Grammar, style				
run-on sentence	x			
awkward syntax	x			
missing/omitted article	x			
Entities, cohesion				
first mention without explanation	x	x	x	x
acronyms without explanations	x	x	x	x
subsequent mention with explanation	x		x	x
repeated entity			x	x
definite noun phrase without reference to previous mention	x	x	x	x
indefinite noun phrase with reference to previous mention	x	x	x	x
misused quantifier	x	x		x
pronoun with missing antecedent	x	x	x	x
noun phrase with missing antecedent		x		x
proper noun with missing antecedent		x		x
pronoun with misleading antecedent		x	x	x
noun phrase with misleading antecedent		x		x
proper noun with misleading antecedent		x		x
not clear identification of who or what the pronouns and noun phrases are referring to				x
Discourse, coherence				
occurrence of redundancy	x		x	x
occurrence of contradiction	x			x
not explicit conditional sentences	x	x		x
lack of purpose for a sentence	x	x		x
lack of place specification for an event (including collocation, change of location)	x	x		x

lack of time specification for an event (including anachronism, temporal ordering, synchrony, repetition of event)	x	x		x
abrupt topic shift	x	x		x
no semantic relatedness	x	x	x	x
misused word/discourse marker	x	x	x	x

For multi-document processing tasks (as MDS), the last two problem types ("Entities, cohesion" and "Discourse, coherence") look more worthy of identification and treatment, as they are more frequent errors and cause more serious problems. Thus, as described in section 5 (specifically in section 5.1), we have based our corpus annotation on these LQ problems, looking for a more appropriate and informative error set for MDS.

In next section, we introduce the summarization corpus that we used to conduct our investigation of linguistic problems, over which we ran the above summarization methods and performed the corpus analysis.

4 The Corpus

The corpus used in this work was the CSTNews corpus (CARDOSO *et al.*, 2011). This corpus has been specially created for multi-document summarization. It is composed of 140 texts (with an average of 334 words and 14.9 sentences per text) distributed in 50 sets/clusters of news texts written in Brazilian Portuguese[6] from various domains. Each cluster has 2 or 3 texts from different sources that address the same topic. These sources are important Brazilian online newspapers, as *Folha de São Paulo*, *Estadão*, *O Globo*, *Jornal do Brasil*, and *Gazeta do Povo*.

According to the authors, the choice of these news agencies was due to their popularity, to publish the main current news, to the use of a clear and everyday language, and because they make available different

[6] The adoption of a corpus in Portuguese was due to the facts that (i) it was possible to have access to the several different summaries that we needed for this investigation and (ii) the annotators were native speakers of this language, which allowed for a more refined and reliable annotation.

versions of the same facts, which is important for a multi-document corpus.

Besides the original texts, the corpus contains several linguistic annotation layers, manually produced by experts, with satisfactory annotation agreement results. The manual annotations include single and multi-document summaries, text-summary alignments, the identification of temporal expressions, RST and CST annotation, noun and verb senses, segmentation of the source texts in subtopics, and semantic annotation of informative aspects in summaries, among other annotations. There are also some automatic annotations, which include morphosyntactic and syntactic analyses, with the best parser for Portuguese, and multi-document summaries.

For the annotation task, 200 multi-document summaries have been used since each of the four automatic summarizers generated one extract for each cluster of the CSTNews. Table 2 shows the average of words and sentences per summary generated by each summarizer.

TABLE 2 – Basic counts for the corpus of automatic summaries

System	**Average of words**	**Average of sentences**
GistSumm	362	11
RSumm	134	4
RC-4	132	4
MTRST-MCAD	139.78	7.92

According to the table, the average of words and sentences in the summaries from GistSumm is higher than the summaries produced by the other summarizers. This happens because GistSumm compression rate is computed in a different way in relation to the other summarizers. It is computed over all the source texts, which are concatenated. For the other summarizers, the compression rate is 30% of the largest text of each cluster of the CSTNews corpus. We kept GistSumm in the comparison because we considered it interesting to see how the summary size variance affects the occurrence of LQ problems.

5 Annotation of linguistic problems in multi-document summaries

In this section, we describe the methodology that we used for the annotation of LQ problems in our corpus of automatic multi-document summaries in Portuguese. Such annotation allowed us to understand and categorize the linguistic problems, to check the quality of the automatic summaries and to guide the future development of automatic methods that judge the LQ of multi-document summaries and, consequently, of automatic summarizers.

Given the 50 clusters of the CSTNews corpus, the automatic summarizers GistSumm (PARDO *et al*., 2003; PARDO, 2005), RSumm (RIBALDO *et al*., 2012, 2016), RC-4 (CARDOSO; PARDO, 2015, 2016) and MTRST-MCAD (CASTRO JORGE, 2015) were used to generate individual extractive summaries for each cluster. Therefore, 200 automatic multi-document summaries were produced and manually annotated.

Based on the related literature and the analysis in section 4.2, we synthetically list the linguistic problems of interest in three categories: (i) *Entity Level*, (ii) *Clause Level*, and (iii) *Others* (see TABLE 3). In general, the problems we adopted are strongly based on those of Friedrich *et al*. (2014), extended with some more information and problem types that were necessary for our corpus annotation.

All errors were identified in the corpus with XML markers. The markers have the format <e TYPE=*(error name)*>*(Text Passage)*</e>. For some markers, there is additional information placed after the error name, and this will be explained along with their respective errors. The "*error name*" field is filled with the name of the error identified in the "*text passage*" field, which may contain full sentences or sentence fragments that show the error.

In what follows, the errors are explained once more, now adapted to this work and accompanied by the markup strategy and actual examples of our corpus.

5.1 The LQ problems typology

For the investigation of the problems in automatic multi-document extracts in Portuguese, we organized the linguistic problems of interest in 3 categories: (i) *Entity Level*, (ii) *Clause Level*, and (iii)

Others (errors that are different from the two first categories) (TABLE 3). The *Entity* and *Clause* categories have several types.

TABLE 3 – The typology of LQ problems

Level	Problem type	Tag
Entity	*First mention without explanation*	1M-EXP
	Subsequent mentions with explanation	SM+EXP
	Definite noun phrase without reference to the previous mentions	DNP-REF
	Indefinite noun phrase with reference to the previous mentions	INP+REF
	Pronouns without antecedent	PRO-ANT
	Pronouns with misleading antecedents	PRO_MIS
	Acronyms without explanation	ACR-EXP
Clause	*Redundant information*	RED
	Contradiction	CONTR
	Incomplete sentences	INC_SENT
	No semantic relationship	No_SEM
	Connective/discursive marker without appropriate context	DM
Other	*Errors that are different from the two first categories*	OTHER

5.1.1 Problems in the entity level

Based on Table 3, one sees that the errors in the *entity level* present 7 subcategories: 1M-EXP, SM+EXP, DNP-REF, INP+REF, PRO-ANT, PRO_MIS, and ACR-EXP.

First mention without explanation (1M-EXP) is identified in a summary when the first mention of an entity is not properly introduced. In Figure 24, there is a problem of 1M-EXP in the third sentence (S3) of the summary. In this case, the first mention of entity "Tepco" was annotated because the reader does not know what this entity is, i.e., there is not a clear introduction to this entity in its first mention.

FIGURE 24 – Annotation of a1M-EXP problem

[S3]<e TYPE=1M-EXP>Tepco</e> has declared the earthquake did not cause leaks, but, afterwards, it revealed that 1,200 liters of water with radioactive material from the factory have leaked to the sea.

Subsequent mentions with explanation (SM+EXP) are identified in summaries when entities have already been mentioned in the text, but they still appear with an inappropriate (usually, extra) explanation. For illustration, consider sentences S1 and S2 in Figure 25.

FIGURE 25 – Annotation of a SM+EXP problem

[S1] The president of the Ethics Council of the Senate, Leomar Quintanilha (PMDB-TO), said to be contrary to the unification of the processes against the Senator Renan Calheiros (PMDB-AL).

[S2] <e TYPE=SM+EXP SENT=S1 TEXT= "The president of the Ethics Council of Senate, Leomar Quintanilha (PMDB-TO)"> The president of the Ethics Council of Senate, Leomar Quintanilha (PMDB-TO)</e>, said that he is against the union of representations, however that he will propose to a vote.

The entity "Leomar Quintanilha (PMDB-TO)" is explained in sentence S1 as "The president of the Ethics Council of Senate", and sentence S2 contains the same entity with a repeated explanation, characterizing a type SM+EXP problem. This problem is annotated in the second occurrence of the entity with explanation, as shown in Figure 25. The SENT field contains the identification of the sentence in which the first mention of the entity that was specified in the field TEXT occurs.

Definite noun phrase without reference to previous mentions (DNP-REF) is identified in summaries when a definite noun phrase does not refer to any entity mentioned earlier. For example, consider sentences S1, S2 and S3 in Figure 26.

FIGURE 26 – Example of a DNP-REF problem

[S1] At least 17 people died after the crash of a passenger plane in the Democratic Republic of Congo.

[S2] According to an ONU spokeswoman, the plane, Russian-made, was trying to land in the Bukavu airport in the midst of a storm.

[S3] <e TYPE=DNP-REF>The spokesman</e> informed that the plane, a Soviet Antonov-28 of Ukrainian-made and owned by a Congolese company, Trasept Congo, also carried a cargo of minerals.

The error <e TYPE=DNP-REF> in sentence 3 is due to the definite noun phrase "The spokesman", for which there is no reference to any entity mentioned earlier.

The *indefinite noun phrase with reference to previous mentions* (INP+REF) problem is identified in summaries when an indefinite article is used together with an entity already mentioned in the discourse (that, therefore, should be introduced in another way). For example, S2, in Figure 27, includes the indefinite noun phrase "an Airbus A320", which was already introduced in S1 ("The Airbus-A320"), causing inconsistency in the summary.

FIGURE 27 – Example of the INP+REF problem

[S1] In São Paulo, on Tuesday (17), the Airbus-A320 of TAM presented a defect in the reverse of the right turbine for the last 13 days.

[S2] The problem would have been detected by the electronic system of the plane, but the plane, <e TYPE=INP+REF SENT=S1 TEXT= "the Airbus-A320"> an Airbus A320</e>, continued flying with the right reverse off.

Pronoun without antecedent (PRO-ANT) is identified when a pronoun does not have a possible antecedent in the summary. For example, the first sentence of the summary in Figure 28 contains the pronoun "he" without a possible antecedent for it.

FIGURE 28 – Example of the PRO-ANT problem

[S1] Hospitalized in a hospital in Buenos Aires, <e TYPE = PRO-ANT>he</e> relapsed and started to feel pain again due to acute hepatitis, according to his personal doctor, Alfredo Cahe.

[S2] "Maradona had a relapse in acute hepatitis. Now, he is stable. Although he improved on Sunday, it is expected that he continues in hospital," Cahe declared to "La Nación".

Pronoun with misleading antecedent (PRO_MIS) is identified when an anaphoric expression refers to a misleading antecedent and its correct antecedent is not present in the summary. In this annotation task, the annotators could check the source text to identify the correct antecedent. In the example in Figure 29, the pronoun "he" (in S2) seems to connect to the entity "Kaká" (in S1). However, in the source text, the pronoun refers to the soccer player "Robinho", who is not cited in the summary.

FIGURE 29 – Example of the PRO_MIS problem

[S1] At 27 minutes, Kaká kicked from far away and Ronaldinho diverted the kick.
[S2] 20 cm from the end line <e TYPE=PRO_MIS ANT="Kaká, Ronaldinho">he</e> dribbled the Ecuadorian defender and crossed the ball to Elano, who scored the fourth goal at 37 minutes.

Besides identifying the type of error in the TYPE tag, the misleading antecedents must also be listed in the ANT tag. This allows the recovery of the problems in future studies.

Acronyms without explanation (ACR-EXP) are identified in a summary by their "non expanded form" or when they are not explained. For example, in the sentences in Figure 30, the "Deic" and "PF" acronyms have no proper introduction.

FIGURE 30 – Example of the ACR-EXP problem

[S1] The other suspect is graffiti man and, according to <e TYPE=ACR-EXP>Deic</e>, he has been arrested for theft, but has already been released.
[S2] The <e TYPE = ACR-EXP CS = "Federal Police"> PF </ e> did not know how to inform if this kind of reward is paid to law enforcement agencies.

Some acronyms are considered to be common sense, such as abbreviations of states and national (Brazilian) political parties. Such cases was annotated with the CS tag, which contains the common sense meaning of the acronym, as shown in the annotation of the error in Figure 30. In this work, common sense was used when the majority of the annotators had the same knowledge about the acronym. Differently from us, Friedrich *et al*. (2014) considered as common sense entities that are in a pre-compiled list of well-known acronyms.

5.1.2 Problems in the clause level

Based on Table 3, the *clause* category has 5 types of problems, which are: RED, CONTR, INC_SENT, No_SEM, and DM.

Redundant information (RED) (in total or partial levels) negatively affects the informativity of summaries. As an example, it is possible to see that sentence S2 in Figure 31 contains information from sentence S1, i.e., it is a repetition. Due to this, we marked this problem as a RED error in the TYPE tag, and we indicated the first sentence where the original information was present.

FIGURE 31 – Example of the RED problem

[S1] A homemade bomb was thrown against the building of the Public Ministry, in the center of the capital, but nobody was injured.

[S2] <e TYPE=RED SENT=S1> A homemade bomb exploded outside the building of the State Public Ministry and nearby shops were hit by shrapnels. </e>

Contradiction (CONTR) is identified when there is a conflict of information between two sentences. In Figure 32, sentences S1 and S2 have contradictory information in relation to the number of injured and dead people. Thus, we marked the sentence that presented the contradiction as CONTR, and we identified the sentence that presented the contradiction in the SENT tag.

FIGURE 32 – Example of the CONTR problem

[S1] The Egyptian Minister of Health Hatem, El-Gabaly, said on Monday that 57 people died and 128 were injured in the collision between two passenger trains in the Nile Delta, north of Cairo.

[S2] <e TYPE=CONTR SENT=S1> At least 80 people died and over 165 were injured on Monday after the collision of two passenger trains in the Nile Delta, north of Cairo, according to the police and the medical sources. </e>

Incomplete sentence (INC_SENT) is identified when there are no punctuation marks, space or complement of a sentence. For example, in the summary in Figure 33, sentence S2 finished with a comma, i.e., this sentence is considered incomplete.

FIGURE 33 – Example of the SENT_INC problem

[S1] As expected, the athlete Fabiana Murer won the gold medal in the pole vault at the Pan American Games in Rio, on Monday, at the João Havelange Stadium.

[S2] <e TYPE=INC_SENT>Murer won the highest place of the podium with the 4m60 mark against 4M40 of the American April Steiner.</e>

No semantic relationship (No_SEM) is identified when adjacent sentences do not present proper semantic relationship. As an example, Figure 34 contains a summary, in which there is not a clear relation between S2 and S1.

FIGURE 34 – Example of the No_SEM problem

[S1] Abadia was arrested in a residence located in a luxury condominium of Aldeia da Serra, in São Paulo.
[S2] <e TYPE=No_SEM>Four safes were also sealed</e> [...]

Connective/discursive marker without appropriate context (DM) is identified when the use of explicit discourse markers (e.g., "but", "because", "however") are considered inappropriate in the context of the summary. In the summary in Figure 35, the discourse marker "But" does not relate to the previous sentence. This happens due to the extractive nature of the summaries, which may include sentences without their contexts of occurrence. In the annotation of this error, we used the CONEC tag to identify the marker that is inappropriately used.

FIGURE 35 – Example of the DM problem

[…] [S4] Until the end of the game, Bruno and Anderson did not enter the court anymore.
[S5] <e TYPE=DM CONEC="But">But, after that, everybody in the gymnasium screamed the lifter name.</e> [...]

5.1.3 Other problems

In case of problems that were not listed in the previous categories, we labeled them as *Other* and the "EXPLANATION" tag contains the explanation of the error. For example, Figure 36 presents a summary that is problematic because it uses terms in different languages referring to the same entity/event (the "championship"), i.e., "Brasil Open" in sentence S1 and "Aberto do Brasil 2013" in sentence S2.

FIGURE 36 – Example of the *Other* problem: inappropriate references

[S1] In addition to Rafael Nadal, the tournament will have three more athletes among the 20 best of ATP ranking: the Spanish Nicolás Almago (11th place and 3 times champion of Brasil Open), the Argentinian Juan Mônaco (12th) and the Swiss Stanilas Wawrinka (17th).
[S2] The organization of <e TYPE=Other EXPLANATION="reference in Portuguese for the term introduced in English">Aberto do Brasil 2013</e> announced this Tuesday morning that the Spanish Rafael Nadal will be returning to the tournament to be disputed in February.

Problems as “Metadata inclusion” and “Distinct spelling for the same entity” are also considered as belonging to *Other*. Figures 37 and 38 show the respective examples for these problems.

FIGURE 37 – Example of the *Other* problem: metadata inclusion

<e TYPE=Other EXPLANATION=”Metadata inclusion”>FIRST -</e>Murer jumps to break the Pan American record; first gold medal in Athletics.

FIGURE 38 – Example of the *Other* problem: distinct spelling for the same entity

[S1] Israeli military forces in south of Lebanon also reported that, on Sunday, 30 militants of Hesbollah were killed, while an officer and two soldiers were wounded in Oiled.

[S2] The Israeli air force attacked 150 targets early this morning in Lebanon as the Jewish state soldiers killed 10 <e TYPE=Other EXPLANATION=”Distinct spelling for the same entity”>Hezbollah</e> militiamen in the Bint Djebeil and Kafr Hula Lebanese villages, according to military sources.

5.2. The task of linguistic problem annotation

The goal of the annotation was to identify the linguistic errors of the typology described in section 5.1 (see Table 3) in summaries that were automatically generated by the 4 cited automatic summarizers.

The task was carried out by a group of experts in a face-to-face process, i.e., it happened every day at a specific time and place for 1 hour. We believe that: 1 hour a day made the task less exhausting for the annotators and this may have positively influenced the annotation quality; everyday annotation, in turn, creates commitment to the task. The task was also better managed with all annotators in the same place.

We used some days to train the 6 annotators (2 linguists and 4 computational scientists) and to refine the guidelines with them. These annotators have been chosen because of their experience in NLP and with annotation tasks.

Due to the subjectivity of the task, the linguistic problems were only marked after a consensus among the annotators or when the majority of them agreed. This strategy is interesting because it produces a more consistent and correct annotation, allowing a more robust annotation with high linguistic error coverage. On the other hand, the annotation time is longer in comparison to the traditional strategies, in which each annotator

works with different summaries per day. In this work, the duration of the annotation task was approximately 150 days.

Some problems are interesting to comment. The *No semantic relationship* error was the error that required more attention and refinement in its interpretation, due to the high degree of subjectivity involved in this problem identification. Thus, this interpretation involved discussions among annotators until the reconciliation process, i.e., the final decision for marking the problem, as suggested by Hovy and Lavid (2010). The *Acronym without explanation* problem required that every annotator had the same background knowledge in order to fill the CS (common sense) field. This background knowledge may be different among the annotators and this may cause the inadequate identification of the problem. Therefore, the annotation approach used in this work may have avoided this type of problem.

Even with all the annotators working together, we periodically verified the agreement among them. In such case, each annotator separately worked with the same summaries, and, after this, we calculated the agreement by the Kappa measure (CARLETTA, 1996). Kappa is a classic agreement measure in NLP, which indicates the correlation between annotators while it discounts the agreement by chance. In the literature, there are some suggestions that guide the decision on the minimum agreement value that is expected: a value less than 0.4 may indicate an unreliable annotation; if it is between 0.4 and 0.75, the annotation is satisfactory; and if it is higher than 0.75, it is very good. This value, however, changes according to the subjectivity of the phenomenon and the difficulty of the annotation task. We consider our annotation task as a very difficult and subjective one. Thus, we expect lower kappa values.

We present the results of the annotation in the following section.

6 Results and discussion

6.1 Performance of the summarizers

For the 4 multi-document summarizers considered in this task (GistSumm, RSumm, RC-4, and MTRST-MCAD), 1,359 linguistic problems were identified. Table 4 shows the quantity of errors by summarizer.

TABLE 4 – Total of problems annotated for each summarizer

Systems	**Annotated problems**	
	Quantity	**%**
GistSumm	521	38.33
MTRST-MCAD	421	30.97
RC-4	220	16.20
RSumm	197	14.50

As expected, Table 4 shows that there are more problems in the summaries produced by GistSumm than in the summaries of the others, which looks natural, given that GistSumm is a very simple summarizer and produces longer summaries than the other systems, running more risk to commit problems.

The statistics computed from our annotation show that *redundant information* (RED) is the most recurrent error, with a total of 261 occurrences in the summaries of the four summarizers (see TABLE 5). This result confirms that detection and properly treatment of redundancy are problematic issues in MDS. Together with *acronyms without explanation* (ACR-EXP) and *definite noun phrase without reference to the previous mentions* (DNP-REF), RED accounted for more than 50% of the problems.

TABLE 5 – Problems by subcategory

Problem subcategory	**Qty.**	**%**
Redundant information (RED)	261	19.20
Acronyms without explanation (ACR-EXP)	255	18.76
Definite noun phrase without reference to the previous mentions (DNP-REF)	182	13.39
*Subsequent mentions with explanation (*SM+EXP)	152	11.18
No semantic relationship (No_SEM)	136	10.00
Other	123	9.05

First mention without explanation (1M-EXP)	103	7.57
Contradiction (CONTR)	41	3.01
Connective/discursive marker without appropriate context (DM)	37	2.72
Pronouns without antecedent (PRO-ANT)	30	2.20
Indefinite noun phrase with reference to the previous mentions (INP+REF)	25	1.83
Incomplete sentence (INC_SENT)	11	0.80
Pronouns with misleading antecedents (PRO_MIS)	3	0.29

Table 6 illustrates the quantity of *redundant information* (RED) error for each summarizer, as it is the most recurrent problem.

TABLE 6 – Total of *redundant information* (RED) problems

Systems	**Problems**	
	Quantity	**%**
GistSumm	160	61.30
RC-4	55	21.08
MTRST-MCAD	23	8.81
RSumm	23	8.81

Redundancy errors may also directly increase the problems of the *entity* category as redundancy may cause repetitions and introduction of entities in an inappropriate way. For example, Figure 39 shows part of a summary with *redundant information* (RED) and problems related to the *entity* category embedded in the redundant sentences.

The sentences with repeated information (as in S5, S7 and S17) present errors of the *entity* category. In this case, for each redundant information error, there is one INP+REF error. This also certainly contributes to the high amount of annotated errors in the summaries produced by GistSumm.

FIGURE 39 – A GistSumm summary with *redundancy* caused by an *entity* category error

[S1] A new series of criminal attacks was recorded early on Monday, the 7th, in São Paulo and municipalities in the countryside of the State of São Paulo.

[S2] A homemade bomb was thrown against the building of the Public Ministry, in the state capital.

[S3] The criminal actions may have been ordered by the leaders of the Primeiro Comando da Capital (PCC), which had promised to return the attacks in São Paulo on Father's Day on Sunday.

[S4] At ABC Paulista, at least ten buses were set on fire - seven in Mauá and three in Santo André.

[S5] <e TYPE=RED SENT=S2><e TYPE=INP+REF SENT=S2 TEXT="A homemade bomb"> A homemade bomb </e> was thrown against <e TYPE= SM+EXP SENT=S2 TEXT = "the Public Ministry"> the Public Ministry (MP)</ e> headquarters. </e>

[S6] The building of the Treasury secretary, in the center, was hit by three homemade bombs.

[S7] <e TYPE=RED SENT=S3> The leaders of the criminal gang PCC had promised <e TYPE=INP+REF SENT=S1 TEXT="A new series of criminal attacks"> A new wave of attacks </e> will happen if the Public Ministry of São Paulo deny the temporary exit of prisoners because of Father's Day. </e> [...]

[S17] <e TYPE=RED SENT=S1,S3> Members of PCC had promised <e TYPE=INP+REF SENT=S1 TEXT="A new series of criminal attacks"> a new wave of attacks </e> will happen if the Public Ministry of São Paulo deny the temporary exit of prisoners because of Father's Day.</e> [...]

In relation to the quantity of problems by category, Table 7 synthesizes the achieved results. The *entity* category included the most frequent problems, which occurred 750 times. The fact that this category had the highest amount of problems was expected, since there are more entities than sentences in a summary. As an example, the summary in Figure 40 was generated by RSumm, and it does not present errors of the *clause* category. However, five annotated errors are related to the *entity* category, and 1 to the *other* category.

According to Table 7, the RC-4 and RSumm summarizers, which make use of more linguistic knowledge, present a lower quantity of errors than the others. In particular, the RSumm summarizer had the lowest quantity of annotated errors in two of the three categories; in the remaining category, it was outperformed by the RC-4 system only.

TABLE 7 – Quantity of problems by category

Systems	***Entity Level***	***Clause Level***	***Other***
GistSumm	239	*221*	61
MTRST-MCAD	252	129	40
RC-4	123	83	14
RSumm	136	53	8
Total	**750**	**486**	**123**

FIGURE 40 – Example of Rsumm summary with more problems of the *entity* category

[S1] <e TYPE=DNP-REF>In the second round</e>, the vote intentions for President Lula fell from 53% in June to 50% in July, while candidate Alckmin increased from 29% to 36%.

[S2] <e TYPE = ACR-EXP>CNI</e> explains that the research does not provide a comparison with the previous survey for the first round, because it is the first time that <e TYPE=ACR-EXP> Ibope </e> uses the official list of candidates for president.

[S3] Although it does not allow comparisons, it is worth remembering that, in June, Lula had 48% of the votes; Alckmin 18% and <e TYPE=1M-EXP>Heloisa Helena </e> 5%.

[S4] The margin of error is two percentage points upwards or downwards.

[S5] <e TYPE=Other EXPLANATION="Phrase with ambiguous referent">The research</e> was held between 29 and 31 July and was registered in <e TYPE=ACR-EXP>TSE</e> under number 12197/2006.

Some important data is also presented in Table 8, such as the percentage of the problems that were found in the summaries generated by each of the four summarizers. We show in bold some of the main errors for each system.

According to the table, *redundant information* is the main problem in 2 of the 4 summarizers of different approaches, i.e., the GistSumm (of the shallow approach) and the RC-4 summarizer (of the deep approach). The *acronyms without explanation* problem had the greatest occurrence in the RSumm summarizer. In the MTRST-MCAD summarizer, 25.42% of the identified problems were related to *definite*

noun phrase without reference to the previous mentions, being the most frequent error for this summarizer.

Except for the *pronouns with misleading antecedents* problem, which was not identified in the summaries generated by MTRST-MCAD and RSumm systems, all the other errors happened at least in 1 summary of each summarizer. This shows that the summarizers did not treat or inadequately treated the problems that affect LQ.

TABLE 8 – Occurrence of each problem in the *corpus* per summarizer

Problems	**Systems**			
	MTRST-MCAD	**GistSumm**	**Rsumm**	**RC-4**
RED	5.46%	**30.71%**	11.68%	**25.00%**
ACR-EXP	12.83%	18.62%	**27.92%**	22.27%
DNP-REF	**25.42%**	3.45%	18.78%	9.09%
SM+EXP	5.23%	16.51%	6.60%	14.09%
No_SEM	19.95%	4.22%	8.63%	5.91%
Other	9.50%	11.71%	4.06%	6.36%
1M-EXP	10.69%	4.03%	12.18%	5.91%
CONTR	0.95%	4.80%	1.02%	4.55%
DM	3.56%	1.73%	4.57%	1.82%
PRO-ANT	4.75%	0.77%	1.52%	1.36%
INP+REF	0.95%	2.30%	2.03%	2.27%
INC_SENT	0.71%	0.96%	1.02%	0.45%
PRO_MIS	0.00%	0.19%	0.00%	0.91%

Generally, the results of the annotation showed that the summarizers with the best summary informativeness evaluation in the area (RSumm and RC-4) also had a lower quantity of problems, but these summarizers still need to be improved, as there are LQ problems to be tackled.

It is interesting to notice that some of the error types in this work may be directly related to the classical ones of TAC. For example, the

clause category has problems such as*redundant information*, *connective/ discursive marker without appropriate context*, and *incomplete sentences*, which are directly related to *grammaticality* and to *no redundancy* in TAC.

Besides, the *no semantic relationship* problem of this category affects the *textual focus* of TAC, because a summary without semantic relationship among its sentences does not have a defined focus. The *referential clarity* of TAC is directly related to the *entity* category by means of the problems *definite noun phrase without reference to the previous mentions*, *indefinite noun phrase with reference to the previous mentions*, and *pronouns without antecedent*, for example. The *textual structure and coherence* errors are the merge of all errors that were considered.

The main problem observed in multi-document summaries in Friedrich *et al*. (2014) was *incomplete sentence*. On the other hand, the *redundant information* problem was the main problem in this work. However, these two problems are in the clause level, which may indicate that this is an important issue for future research.

In the experiments from Otterbacher *et al*. (2002), the *temporal ordering* problem was the most frequent one. This problem is related to the identification of correct temporal relationships between the events described in a summary. This problem is also at the clause level, which, in this work, happens in the *no semantic relationship* (No_SEM) error, when the temporal order of an event is not respected in the selection of sentences from the source texts to compose a summary.

6.2 Annotation agreement

As commented before in this paper, the annotation was made by a group, but we decided to measure the agreement among the annotators to check the understanding of the errors and the problem annotation process itself. For this, we calculated the Kappa measure and the percent agreement of the majority for 4 clusters of the CSTNews corpus (in particular, cluster C12, C22, C32 and C42). Notice that each cluster has 1 summary generated by each summarizer (GistSumm, RSumm, RC-4 and MTRST-MCAD), i.e., 4 summaries in each cluster.

Table 9 shows the Kappa scores for the agreement among annotators in each cluster for the simple indication of errors (in a binary decision).

TABLE 9 – Kappa measure for simply indicating a problem

Cluster	**Kappa**
C12	0.409
C22	0.641
C32	0.578
C42	0.324
Average	**0.488**

Cluster 22 had the best agreement result. However, due to the difficulty of the task, this result is not so high. The subjectivity causes different understandings and this is demonstrated when the annotators do the annotation in isolation. This behavior is repeated in Table 10, when we measure Kappa for the indication of the problem category. The Table 10 shows that the Kappa for the *Other* category had the best values. The agreement was the most significant in cluster 12 for the *Other* problem category.

TABLE 10 – Kappa measure for problem category

Cluster	**Kappa for *Entity Level***	**Kappa for *Clause Level***	**Kappa for *Other***
C12	0.356	0.560	1.000
C22	0.670	0.537	0.902
C32	0.552	0.616	0.627
C42	0.606	0.418	0.751
Average	**0.546**	**0.533**	**0.760**

Considering the relatively low results of Kappa measure, the percent agreement by majority was also relevant in order to better judge the task. In this case, the percentage of sentences in all clusters that the majority of the annotators agreed was calculated. For example, in the summaries of cluster 12, at least 4 of the 6 annotators marked the occurrence of an error in all sentences (100% of the sentences, therefore)

of these summaries. Table 11 shows the results of the agreement by majority, considering the occurrence of a problem in a certain sentence.

Table 11 shows that the majority of the annotators agreed in marking an error in all the sentences in the summaries of clusters C12 and C22. Clusters C32 and C42 also presented a good percentage of agreement.

TABLE 11 – Percent agreement (by majority) for the indication of a problem in a sentence

Clusters	**% of sentences**
C12	100
C22	100
C32	91.89
C42	81.25
Average	**93.28**

We also used the agreement by majority for categories of problems. We calculated the percentage of sentences for which the majority of the annotators marked an error of a specific category. Table 12 shows the results obtained by this measure of agreement.

The majority of annotators agreed 100% for the sentences in the summaries of clusters C12 and C22, regarding the occurrence of all the problem categories. In cluster C42, the *clause* category was the only one in which the majority of the annotators agreed below 90%. These results showed that the majority of the annotators understood well all the linguistic problem categories identified in the summaries.

TABLE 12 – Percent agreement (by majority) for the indication of a problem category in a sentence

Clusters	**% of sentences with problems**		
	Entity Level	***Clause Level***	***Other* problems**
C12	100	100	100
C22	100	100	100
C32	94.59	91.89	100
C42	90.62	71.87	93.75
Average	**96.30**	**90.94**	**98.43**

To confirm this, Table 13 shows the percentage of sentences for which all annotators agreed in the identification of the linguistic problems.

TABLE 13 – Agreement for 100% of annotators for each problem

Problems	**% of sentences**			
	C12	**C22**	**C32**	**C42**
1M-EXP	54.54	90.00	91.89	81.25
SM+EXP	81.81	76.66	84.48	90.62
DNP-REF	63.63	93.33	83.78	53.12
INP+REF	-	-	89.18	-
PRO-ANT	-	-	-	96.87
ACR-EXP	100	96.66	94.59	93.75
No_SEM	81.81	76.66	75.67	75.00
DM	-	-	91.89	96.87
RED	-	83.33	89.18	81.25
CONTR	-	86.66	94.59	-
Other	-	96.66	81.08	90.62

According to Table 13, over half of the sentences in the summaries had 100% of agreement among the annotators. All the sentences with the *acronyms without explanation* (ACR-EXP) problem were marked by all annotators for the first cluster. The hyphen (-) in Table 13 means that the error was not identified by any of the annotators. The *pronouns with misleading antecedents* (PRO_MIS) and *incomplete sentence* (INC_SENT) problems were not identified in the clusters used in the agreement and, for this reason, are not listed in the Table 13.

With the reported agreement results, we may conclude that the annotation task was well understood and the annotation is reliable. We believe that our well-defined typology of LQ problems was an important reason for the reported agreement scores.

7 Final remarks

This paper reported the study, an annotation task and the characterization of linguistic problems in multi-document summaries automatically produced by systems of varied paradigms, from shallow to deep approaches, including classic and state of the art methods. The corpus consisted of summaries composed by four automatic summarizers, and it was possible to verify that (i) some problems deserve more attention from the automatic summarizers, as problems related to redundancy and introduction of definite noun phrases and acronyms, which accounted for more than 50% of the errors, and (ii) that the summarizers with the best summary informativeness results (according to standard informativeness measures) also produce a lower quantity of problems. Our results may be used as a guide to treat errors in future summarizers.

The literature review and organization and the methodology used for the problem annotation process are also contributions to the area. In particular, the annotation strategy was interesting because the problem annotation involves difficult and fuzzy aspects as subjectivity and world knowledge, which may affect the consistency of the annotation. The agreement values confirmed that such annotation strategy is worthy following.

As future work, we consider to study error correlation in the summaries, as well as automatic methods for detecting and properly dealing with them, improving the summary quality.

For the interested reader, the corpus that was produced, the summarization systems that we used and other related information about this work may be found at the SUCINTO project webpage.[7]

Acknowledgements

The authors are grateful to FAPESP (*Fundação de Amparo à Pesquisa do Estado de São Paulo*), USP Research Office (PRP 668) and Federal University of Goiás for supporting this work.

Contributions of the authors

Márcio de Souza Dias (1rst author): responsible for organizing the corpus annotation task, writing the introductory sections of the paper and pulling together the sections written by the co-authors.
Ariani Di-Felippo (2nd author): responsible for describing and analyzing the linguistic problems annotated in the corpus of automatic summaries.
Amanda Pontes Rassi (3rd author): responsible for writing the description of the linguistic problems of the literature.
Paula Christina Figueira Cardoso (4th author): responsible for describing the summarization methods and the reference corpus in Portuguese for Automatic Summarization.
Fernando Antônio Asevedo Nóbrega (5th author): responsible for computing and describing all the annotation statistics.
Thiago Alexandre Salgueiro Pardo (6th author): responsible for describing the basic concepts of the Automatic Summarization domain.

References

ANDO, R.; BOGURAEV, B.; BYRD, R.; NEFF, M. Multi-document Summarization by Visualizing Topical Content. *In*: ANLP/NAACL WORKSHOP ON AUTOMATIC SUMMARIZATION, 2000, New Brunswick. *Proceedings* […]. New Brunswick: Association for Computational Linguistics, 2000. p. 79-88. DOI: https://doi.org/10.3115/1117575.1117584

[7] Available on: http://conteudo.icmc.usp.br/pessoas/taspardo/sucinto/resources.html. Retrieved at: Feb. 10, 2019.

BEAUGRANDE, R.; DRESSLER, W. U. *Introduction to Text Linguistics.* 1. ed. London: Longman, 1981.

CARBONELL, J.; GENG, Y.; GOLDSTEIN, J. Automated Query-Relevant Summarization and Diversity-Based Reranking. *In*: IJCAI Workshop on AI in Digital Libraries, 1997, Nagoya. *Proceedings* […]. Nagoya: [s.n.], 1997. p. 12-19.

CARDOSO, P. C. F; MAZIERO, E.; JORGE, M.; SENO, E.; DI-FELIPPO, A.; RINO, L.; NUNES, M.; PARDO, T. A. S. CSTNews: A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. *In*: RST BRAZILIAN MEETING, 3., 2011, Cuiabá. *Proceedings* […]. Cuiabá: Sociedade Brasileira de Computação, 2011. p. 88-105.

CARDOSO, P. C. F.; PARDO, T. A. S. Joint Semantic Discourse Models for Automatic Multi-Document Summarization. *In*: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 10., 2015, Natal. *Proceedings* […]. Natal: Sociedade Brasileira de Computação, 2015. p. 81-90.

CARDOSO, P. C. F.; PARDO, T. A. S. Multi-Document Summarization Using Semantic Discourse Models. *Procesamiento de Lenguaje Natural,* Jaén, Espanha, v. 56, n. 1, p. 57-64, 2016.

CARLETTA, J. Assessing Agreement on Classification Tasks: The Kappa Statistic. *Computational Linguistics*, Cambridge, v. 22, n. 2, p. 249-254, 1996.

CASTRO JORGE, M. L. R. *Modelagem gerativa para sumarização automática multidocumento*. 2015. 151f. Tese (Doutorado em Ciência de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2015.

CASTRO JORGE, M. L. R.; PARDO, T. A. S. Experiments with CST-based Multidocument Summarization. *In*: ACL WORKSHOP: GRAPH-BASED METHODS FOR NATURAL LANGUAGE PROCESSING, 5., 2010, Uppsala, Sweden. *Proceedings of TextGraphs-5* […]. Uppsala: Association for Computational Linguistics, 2010. p. 74-82.

CONROY, J. M.; SCHLESINGER, J. D.; KUBINA, J.; RANKEL, P. A.; O'LEARY, D. P. CLASSY 2011 at TAC: Guided and Multilingual Summaries and Evaluation Metrics. *In*: TEXT ANALYSIS

CONFERENCE, 4., 2011, Maryland. *Proceedings* […]. Maryland: NIST, 2011. p. 1-8.

CRISTINI, L. F.; DI-FELIPPO, A. Violações linguísticas em referências a entidades do tipo "pessoa" em extratos automáticos multidocumento. *In*: WORKSHOP ON PORTUGUESE DESCRIPTION, 6., 2019, Salvador. *Proceedings* […]. Salvador: [s.n], 2019. p. 244-252.

DANG, H. T. Overview of DUC 2005. *In*: DOCUMENT UNDERSTANDING CONFERENCE, 2005, Vancouver. *Proceedings* […]. Vancouver: NIST, 2005 p. 1-12. Available on: https://duc.nist.gov/pubs.html#2005. Retrieved at: January. 2015.

FONSECA, H. P. A.; DIAS, M. S.; SILVA, N. F. F. Identificação automática de erros em sumários multidocumento. *In*: SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 12., 2019, Salvador. *Anais*... Salvador: Brazilian Computer Society, 2019. p. 395-399.

FRIEDRICH, A.; VALEEVA, M.; PALMER, A. LQVSumm: A Corpus of Linguistic Quality Violations in Multi-Document Summarization. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 9., 2014, Reykjavik. *Proceedings* […]. Reykjavik: European Language Resources Association, 2014. p. 1591-1599.

GIANNAKOPOULOS, G.; KARKALETSIS, V. AutoSummENG and MeMoG in Evaluating Guided Summaries. *In*: TEXT ANALYSIS CONFERENCE, 4., 2011, Maryland. *Proceedings* […]. Maryland: NIST, 2011. p. 1-10.

HAGHIGHI, A.; VANDERWENDE, L. Exploring Content Models for Multi-Document Summarization. *In*: HUMAN LANGUAGE TECHNOLOGIES: THE ANNUAL CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ACL, 2009, Boulder. *Proceedings* […]. Boulder: NACL, 2009. p. 362-370. DOI: https://doi.org/10.3115/1620754.1620807

HOVY, E. H.; LAVID, J. M. Towards a Science of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation Studies,* [*S.l.*], v. 22, n. 1, p. 13-36, 2010.

KASPERSSON, T.; SMITH, C.; DANIELSSON, H.; JÖNSSON, A. This Also Affects the Context – Errors in Extraction Based Summaries. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION, 8., 2012, Istanbul. *Proceedings* […]. Istanbul: European Language Resources Association, 2012. p. 173-178.

KOCH, I. G. V. *A coesão textual*. 10. ed. São Paulo: Contexto, 1998.

KOCH, I. G. V.; TRAVAGLIA, L. C. *A coerência textual*. São Paulo: Contexto, 2002.

LIN, C-Y. ROUGE: A Package for Automatic Evaluation of Summaries. *In*: ACL WORKSHOP ON TEXT SUMMARIZATION BRANCHES OUT, 2004, Barcelona. *Proceedings* […]. Barcelona: ACL, 2004. p. 74-81.

LIN, Z.; LIU, C.; NG, H. T.; KAN, M. Combining coherence models and machine translation evaluation metrics for summarization evaluation. In: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 50., 2012, Jeju Island. *Proceedings* […]. Jeju Island: ACL, 2012. p. 1006-1014.

MANI, I. *Automatic Summarization*. Amsterdam: John Benjamins Publishing, 2001.

MANI, I.; MAYBURY, M. T. *Advances in Automatic Text Summarization.* Cambridge: The MIT Press. 1999. DOI: https://doi.org/10.1075/nlp.3

MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: A Theory of Text Organization. *Technical Report ISI/RS-87-190*, 1987. Available on: https://www.sfu.ca/rst/05bibliographies/bibs/ISI_RS_87_190.pdf. Retrieved at: March. 2015.

MARCU, D. Discourse Trees Are Good Indicators of Importance in Text. *In*: MANI, I.; MAYBURY, M. T. (ed.). *Advances in Automatic Text Summarization*. Cambridge: The MIT Press, 1999. 123-136.

MCKEOWN, K.; RADEV, D. R. Generating Summaries of Multiple News Articles. *In*: ANNUAL INTERNATIONAL ACM-SIGIR CONFERENCE ON RESEARCH AND DEVELOPMENT IN INFORMATION RETRIEVAL, 18., 1995, Seattle. *Proceedings* […]. Seatle: Association for Computing Machinery, 1995. p. 74-82. DOI: https://doi.org/10.1145/215206.215334

MIHALCEA, R.; TARAU, P. An Algorithm for Language Independent Single and Multiple Document Summarization. *In*: INTERNATIONAL JOINT CONFERENCE ON NATURAL LANGUAGE PROCESSING, 2., 2005, Jeju Island. *Proceedings* […]. Jeju Island: ACL, 2005. p. 19-24. DOI: https://doi.org/10.1007/11562214

NENKOVA, A.; MCKEOWN, K. R. *Automatic Summarization*. Foundations and Trends in Information Retrieval. Hanover, MA: Now Publishers, 2011. DOI: https://doi.org/10.1561/1500000015

OLIVEIRA, P. C. F. de. CatolicaSC at TAC 2011. *In*: TEXT ANALYSIS CONFERENCE (TAC), 4., 2011, Gaithersburg. *Proceedings* […]. Gaithersburg: NIST, 2011. p. 1-3.

OTTERBACHER, J. C.; RADEV, D. R.; LUO, A. Revisions that Improve Cohesion in Multi-Document Summaries: A Preliminary Study. *In*: ACL-02 WORKSHOP ON AUTOMATIC SUMMARIZATION, 2002, Philadelphia. *Proceedings* […]. Philadelphia: ACL, 2002. p. 27-36. DOI: https://doi.org/10.3115/1118162.1118166

OWCZARZAK, K.; DANG T. H. Overview of the TAC 2011 Summarization Track: Guided task and AESOP task. *In*: TEXT ANALYSIS CONFERENCE, 3., 2011, Gaithersburg. *Proceedings* […]. Gaithersburg: NIST, 2010. Available on: https://tac.nist.gov/2011/Summarization/Guided-Summ.2011.guidelines.html. Retrieved at: January. 2015.

PARDO, T. A. S.; RINO, L. H. M.; NUNES, M. G. V. GistSumm: A Summarization Tool Based on a New Extractive Method. *In*: WORKSHOP ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE, 6., 2003, Faro, Portugal. *Proceedings* […]. Faro: Springer, 2003. p. 210-218. DOI: https://doi.org/10.1007/3-540-45011-4_34

PARDO, T. A. S. GistSumm - GIST SUMMarizer: extensões e novas funcionalidades. *Technical Report NILC-TR-05-05,* 2005. Available on: https://sites.icmc.usp.br/taspardo/NILCTR0505-Pardo.pdf. Retrieved at: January. 2015.

PITLER, E.; LOUIS, A.; NENKOVA, A. Automatic Evaluation of Linguistic Quality in Multi-document Summarization. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 48., 2010, Uppsala, Sweden. *Proceedings* […]. Uppsala: ACL, 2010. p. 544-554.

RADEV, D. R. A Common Theory of Information Fusion from Multiple Text Sources, Step One: Cross-document Structure. *In*: ACL SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1., 2000, Hong Kong. *Proceedings* […]. Hong Kong: ACL, 2000. p. 74-83. DOI: https://doi.org/10.3115/1117736.1117745

RADEV, D. R.; TEUFEL, S.; SAGGION, H.; LAM, W.; BLITZER, J.; CELEBI, A.; QI, H.; LIU, D.; DRABEK, E. Evaluation Challenges in Large-Scale Multi-Document Summarization. *In*: ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS, 41., 2003, Sapporo, Japan. Proceedings […]. Sapporo: ACL, 2003. p. 375-382. DOI: https://doi.org/10.3115/1075096.1075144

RIBALDO, R.; AKABANE, A. T.; RINO, L. H. M.; PARDO, T. A. S. Graph-based Methods for Multi-Document Summarization: Exploring Relationship Maps. Complex Networks and Discourse Information. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF PORTUGUESE, 10., 2012, Coimbra. *Proceedings* (Lecture Notes in Computer Science 7243) […]. Coimbra: Springer, 2012. p. 260-271. DOI: https://doi.org/10.1007/978-3-642-28885-2_30

RIBALDO, R.; CARDOSO, P. C. F.; PARDO, T. A. S. Exploring the Subtopic-Based Relationship Map Strategy for Multi-Document Summarization. *Journal of Theoretical and Applied Computing* (RITA), Porto Alegre, RS, v. 23, n. 1, p. 183-211, 2016. DOI: https://doi.org/10.22456/2175-2745.59104

SALTON, G.; SINGHAL, A.; MITRA, M.; BUCKLEY, C. Automatic Text Structuring and Summarization. *Information Processing & Management*, [*S.l.*], v. 33, n. 2, p. 193-207, 1997. DOI: https://doi.org/10.1016/S0306-4573(96)00062-3

ZHANG, Z.; GOLDENSHON, S. B.; RADEV, D. R. Towards CST-Enhanced Summarization. *In*: NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE, 18., 2002, Menlo Park, CA. *Proceedings* […]. Menlo Park: AAAI, 2002. p. 439-445.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 909-958, 2021

Procedimentos para construção do *Corpus* da Computação da Língua Inglesa (CoCLI) e cálculo do esforço na construção manual de *corpora*

Procedures for Corpus of Computing in English (CoCLI) construction and effort calculation in manual construction of corpora

Fernando Paulino de Oliveira
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil
fernandooliveira@ufu.br
http://orcid.org/0000-0002-7002-9664

Resumo: O presente trabalho tem como objetivo descrever os procedimentos metodológicos da pesquisa intitulada "*ToGatherUp*: um protótipo de ferramenta para a construção de corpora" que verificou o efeito da incorporação da ferramenta *ToGatherUp* no tempo e no esforço necessários para a construção manual de um *corpus* que elaboramos: o *Corpus* da Computação da Língua Inglesa (CoCLI). Para tanto, discorremos sobre como os autores da pesquisa desenvolveram um conjunto de métricas de medição de esforço – Esforço da Atividade (EA), Esforço Total de Coleta do Texto (ETCT) e Esforço Total do Projeto (ETP) – que serviram de base para a realização de um experimento estatístico comparativo entre os projetos de elaboração manual de duas versões idênticas do CoCLI que se diferenciam por em um deles utilizar o *ToGatherUp* e o outro não. O resultado do experimento demonstrou uma redução média de 7,47% no ETP do projeto em que o *ToGatherUp* foi incorporado em relação ao ETP do projeto em que a ferramenta não foi utilizada, o que corroborou a hipótese de que ela reduz o tempo e o esforço despendidos pelo pesquisador em projetos de elaboração manual de *corpora*.

Palavras-chave: Linguística de *Corpus*; construção manual de *corpus*; métricas de medição de esforço; *ToGatherUp.*

Abstract: The present work aims to describe the methodological procedures of the research entitled "*ToGatherUp*: a prototype of a tool for corpora construction" that verified the effect of incorporating *ToGatherUp* in necessary time and effort invested

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.909-958

in manual construction of *Corpus* of Computing in English (CoCLI). To this end, we discuss how the research authors developed a set of metrics for measuring effort – Activity Effort (EA), Total Effort for Text Collection (ETCT) and Total Project Effort (ETP) – which served as the basis for conducting a comparative statistical experiment between the manual elaboration of two identical versions of the CoCLI: which differ from each other by one of them using the *ToGatherUp* and the other one not using it. The experiment shows an average reduction of 7.47% in the ETP when using *ToGatherUp* compared to the ETP when not using the tool. This result corroborates the hypothesis that the tool reduces the time and effort spent by the researcher on manual elaboration projects of *corpora*.

Keywords: *Corpus* Linguistics; manual construction of *corpus*; effort measurement metrics; *ToGatherUp*.

Submetido em 25 de agosto de 2020

Aceito em 09 de novembro de 2020

1 Introdução

O desenvolvimento de pesquisas com base na observação empírica de dados da língua favoreceu o surgimento e o crescimento da Linguística de *Corpus*, doravante LC, que é "uma nova metodologia (que utiliza textos naturais e ferramentas informáticas para descrever a língua) e uma nova disciplina (no sentido de uma nova abordagem à descrição linguística)" (FRANKENBERG-GARCIA, 2012, p. 12). Conforme esclarece Berber Sardinha (2004), para que seja possível o uso prático da LC, o interessado precisa de "um ingrediente essencial: o *corpus*" (BERBER SARDINHA, 2004, p. 45).

A construção de *corpora* de pequenas extensões[1] pode não representar um desafio complicado, mas a de *corpora* compostos por grande volume de dados tem sido reportada como uma das partes mais difíceis do desenvolvimento de uma pesquisa (cf. KÜBLER; ASTON, 2010; ATKINS; CLEAR; OSTLER, 1992; BAKER, 2010; BIANCHI, 2012; EDWARD, 2015; MACMULLEN, 2003; MCENERY; HARDIE, 2011; MCENERY; XIAO; TONO, 2006; MINSHALL, 2013;

[1] A extensão ou o tamanho de um *corpus* representa o volume de dados linguísticos disponíveis para análise. Na seção Fundamentação teórica, discutimos sobre a extensão de *corpora*.

RENOUF, 2007; SEMINO; SHORT, 2004; VOORMANN; GUT, 2008; ZANETTIN, 2014). A principal reclamação dos linguistas refere-se à quantidade enorme de tempo e esforço necessária para a realização das atividades relativas à construção de um *corpus*. Além do tempo e esforço, Edward (2015) e Garretson (2008) afirmam que, ao começar a construção de um *corpus*, uma das primeiras barreiras enfrentadas pelos pesquisadores é encontrar ferramentas computacionais capazes de dar suporte[2] especializado às atividades do projeto.

Diante desses desafios, a proposta deste trabalho é contribuir com a prática de linguistas e pesquisadores de áreas afins por meio da apresentação dos procedimentos metodológicos adotados na pesquisa intitulada "*ToGatherUp*: um protótipo de ferramenta para a construção de *corpora*"[3] (OLIVEIRA, 2019). O objetivo dessa pesquisa - determinar os efeitos da incorporação do *ToGatherUp* no esforço necessário para a construção manual de *corpora* – levou seus autores a percorrerem um interessante e produtivo caminho que gerou uma proposta de sistematização do trabalho de construção manual de *corpora*, a criação de métricas de aferição de esforço e culminou na realização de um experimento que revelou a eficácia do *ToGatherUp* na redução do tempo e esforço investido na criação de *corpora*. Para alcançarmos nosso objetivo, nas próximas seções deste artigo, apresentaremos a fundamentação teórica, a metodologia, os resultados alcançados na pesquisa e nossas considerações finais sobre ela.

2 Fundamentação teórica

A Linguística é a área em que se desenvolve o estudo científico da linguagem humana com base em fatos linguísticos (MARTINET, 1978). De acordo com Widdowson (1996), de modo geral, os fatos linguísticos podem ser inferidos por meio da introspecção, da elicitação e da observação de dados provenientes do uso real da língua pelos seus usuários. Widdowson (1996) esclarece que os fatos linguísticos apreendidos por meio da introspecção e da elicitação não revelam o uso

[2] Do ponto de vista dos autores da pesquisa retratada por nós, as ferramentas que oferecem suporte à construção manual de *corpora* são aquelas que oferecem recursos que facilitam as atividades e o gerenciamento do projeto de construção manual de *corpora*.
[3] Disponível em: www.ileel.ufu.br/togatherup. Acesso em: 1 mar. 2019.

efetivo da língua, pois partem das intuições que os seus usuários têm sobre ela. Já a observação de dados linguísticos decorrentes do uso real da língua e que refletem o comportamento linguístico de seus usuários constitui-se como uma forma mais segura para a realização de inferências sobre a língua. Nesse sentido, as análises linguísticas com base na LC podem ser consideradas altamente confiáveis, uma vez que partem da observação de *corpora* compostos por dados linguísticos reais.

Sinclair (2005) afirma que a construção de um *corpus* deve ser realizada de acordo com critérios bem definidos e eficientes o bastante para que o seu delineamento final possa garantir que o conjunto de textos seja representativo. O conceito de representatividade na LC está associado à capacidade que um *corpus* tem de representar uma língua ou uma variedade dela e ao modo como foi construído. Podemos dizer que um *corpus* é representativo quando, a partir da análise do conjunto de textos provenientes das várias situações comunicativas reais de uma comunidade linguística, é possível obter conclusões, a respeito de suas propriedades, que permitam generalizações sobre a língua ou sobre a variedade de língua em estudo.

A fase de construção de um *corpus* em que são definidos os seus critérios tem sido referenciada pelos autores da LC como o "desenho do *corpus*".[4] Firmar o desenho de um *corpus* não é uma tarefa simples, pois, conforme Berber Sardinha (2004), não existem critérios objetivos para isso. Segundo Blecha (2012), a delimitação do desenho de um *corpus* deve ser orientada em consonância com os objetivos da pesquisa. Tagnin (2010) coaduna com Blecha (2012) e afirma que cabe ao criador do *corpus* a responsabilidade de definir os critérios que possam garantir sua representatividade. Dentre os critérios para a construção de *corpora*, na pesquisa aqui relatada, os fundamentos e implicações referentes à extensão do *corpus* ganham importância.

A extensão do *corpus* representa o volume de dados linguísticos que ele dispõe para análise. Na literatura da LC, não encontramos a definição exata do tamanho necessário para que um *corpus* seja representativo. No entanto, para estudos que consideram a chavicidade[5]

[4] Na literatura da LC, em língua inglesa, encontramos o termo *corpus design*.

[5] De acordo com Fromm (2007), a chavicidade (*keyness*) informa o quanto uma palavra se destaca na relação entre a sua frequência no *corpus* de estudo e no *corpus* de referência.

de palavras, encontramos recomendações e estimativas, como a de Berber Sardinha (2004), que afirma que a relação de tamanho entre os *corpora* de estudo e os *corpora* de referência influencia a quantidade de palavras-chave obtidas. O autor recomenda que um *corpus* deve ser o mais extenso possível.

Mesmo com a recomendação de Berber Sardinha (2004), cabe pontuar que a extensão de um *corpus* está sujeita à disponibilidade de dados que atendam às especificidades do desenho dele. A obtenção de dados suficientes para cada campo semântico de uma Árvore de Domínio, no caso das pesquisas terminológicas, ou para cada gênero textual que compõe um *corpus* de estudo do léxico, de modo que seja garantido o balanceamento[6] do *corpus*, é um exemplo dessa situação. Ademais, Fromm (2003) chama a atenção para o fato de que o desenvolvimento de um *corpus* extenso requer a participação de vários pesquisadores e auxiliares; caso contrário, a construção dele pode demorar anos para ser concluída. Nessa situação, há a questão do tempo que o pesquisador (ou a equipe de pesquisadores) tem para dedicar à obtenção de dados.

Berber Sardinha (2005) observa que, na prática, "o pesquisador coleta uma certa quantidade de dados de acordo com suas possibilidades, efetua a análise, mas não sabe se sua coleta foi além ou aquém do que seria teoricamente mais adequado" (BERBER SARDINHA, 2005, p. 188). Por essa razão, Nelson (2010) afirma que a criação de um *corpus* é "uma aceitação entre o que é o esperado e o que é possível" (NELSON, 2010, p. 30)[7] e Meyer (2004) explica que as mudanças no desenho inicial do *corpus* são naturais e inevitáveis (desde que não comprometam a integridade do *corpus*) diante dos obstáculos e complicações que podem surgir durante a sua compilação.

2.1 A organização do trabalho em projetos de construção manual de um *corpus*

A LC não oferece padrões pré-estabelecidos ou modelos sistematizados para a construção manual de um *corpus*. O que encontramos na sua literatura são abordagens que, embora obedeçam às

[6] Aluísio e Almeida (2006) definem o balanceamento como o equilíbrio entre as categorias atribuídas aos textos que compõem um *corpus*.

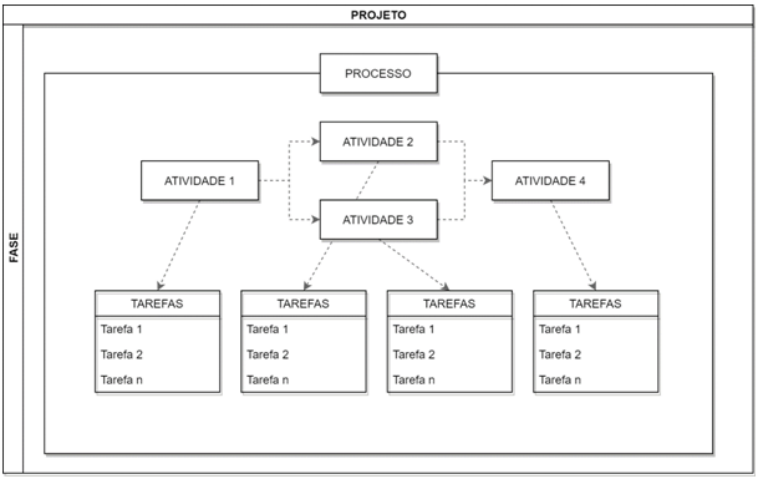
[7] Original: "any attempt at corpus creation is therefore a compromise between the hoped for and the achievable".

diretrizes criadas por Sinclair (2005), diferenciam-se entre si em aspectos de organização, uso de ferramentas e técnicas. Cabe destacar que, ao mesmo tempo em que a inexistência de um padrão promove a flexibilidade das práticas de elaboração de *corpora*, ela também gera problemas como a variação significativa dos nomes que são atribuídos às ações que envolvem a construção de um *corpus*. Nesse sentido, há autores que se referem ao trabalho de criação de *corpus* como um processo dividido em estágios (cf. ATKINS; CLEAR; OSTLER, 1992; ESCARTÍN, 2012; KENNEDY, 1998), em ciclos (cf. BIBER, 1993) ou em passos (cf. SANTOS, 2011). Além dessas denominações, é comum encontrarmos palavras, tais como: "tarefas", "atividades" e "procedimentos", sendo utilizadas com o mesmo sentido, isto é, remetendo-se às mesmas ações.

Por essa razão, os autores da pesquisa aqui retratada decidiram adotar os termos "fases", "processos", "atividades" e "tarefas" utilizados na área de Gerenciamento de Projetos para designar as partes do "ciclo de vida" de um projeto. A adoção dessa nomenclatura pelos autores foi feita por considerarem que a construção manual de um *corpus* equivale à realização de um projeto em conformidade com o conceito de projeto e, em partes, nos princípios da área de Gerenciamento de Projetos, expostos no guia *Project Management Body of Knowledge* (PMBOK), publicado em 2013 e considerado como a principal referência da área de Gerenciamento de Projetos. De acordo com o PMBOK, um projeto é "um esforço temporário empreendido para criar um produto, serviço ou resultado exclusivo" dentro de um "ciclo de vida" (PMBOK, 2013, p. 3). O ciclo de vida de um projeto corresponde à sequência de fases pelas quais ele passa ao longo do seu desenvolvimento.

Durante o ciclo de vida do projeto, cada fase pode comportar um ou mais processos. Estes, por sua vez, podem admitir uma ou mais atividades. Uma atividade pode relacionar-se com outra(s), de maneira lógica, de modo que seu início ou sua continuidade somente seja possível após a geração de um ou mais resultados (entregas) de outra(s) atividade(s). No nível mais baixo do ciclo de vida de um projeto, encontram-se as tarefas, que são as menores unidades de trabalho possíveis pertencentes ao escopo de uma atividade. A Figura 1 ilustra as relações dos componentes do trabalho de um projeto apresentadas neste parágrafo.

FIGURA 1 – Organização do trabalho de um projeto



Fonte: Oliveira (2019, p. 42.)

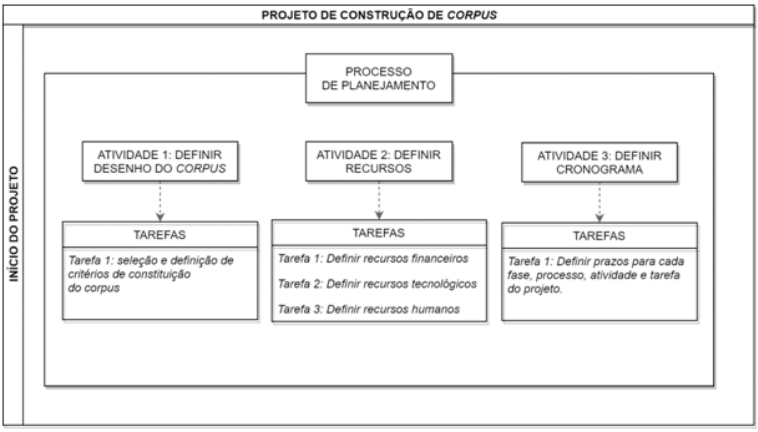
Além de resolver o problema da nomenclatura utilizada na construção de *corpora*, os autores da pesquisa perceberam que era possível transpor o modelo de organização do trabalho dos projetos da área de Gerenciamento de Projetos para os projetos de construção manual de *corpora*. A partir dessa percepção, eles procuraram sistematizar o trabalho relativo à construção manual de *corpora* de com base nesse modelo. Deste modo, propuseram que o projeto de construção manual de um *corpus*, de modo geral, pode ser dividido em três fases distintas: a) a inicial, em que há o planejamento do *corpus*; b) a intermediária, caracterizada pela obtenção, preparação e armazenamento dos dados do *corpus* e c) a de encerramento, na qual ocorre a distribuição dos dados do *corpus*. Nos próximos parágrafos, explicitamos a sistematização concebida pelos autores da pesquisa, situando-a com contribuições teóricas dos autores da LC.

A fase inicial do projeto de construção manual de um *corpus* é caracterizada pela execução das atividades de planejamento do *corpus*, de definição dos recursos necessários para elaborá-lo e de esquematização do cronograma de execução do projeto. De acordo com Nelson (2010, p. 53), há uma série de variáveis que precisam ser consideradas antes do início da compilação dos dados, a saber: o tamanho do *corpus*, o balanceamento dele, a estrutura conceitual em que os textos serão

organizados, o formato de armazenamento dos textos, a maneira como será feita a coleta dos textos, o padrão que será usado para a nomeação dos arquivos e o controle em relação à coleta e ao gerenciamento dos textos. Algumas dessas questões são analisadas durante o desenho do *corpus*, que é a primeira atividade do processo de planejamento do *corpus*. Para o estabelecimento do desenho do *corpus*, o seu criador precisa executar as tarefas de seleção e definição dos critérios que nortearão a constituição do *corpus*.

Ademais, Atkins, Clear e Ostler (1992, p. 3) mencionam o fato de que o planejamento do *corpus* deve prever o uso de recursos financeiros, tecnológicos e humanos necessários para garantir a conclusão do projeto. Santos (2011) complementa as exigências do planejamento do *corpus* ao afirmar que é necessário estabelecer o cronograma para a execução do projeto, pois várias decisões que precisam ser tomadas durante a elaboração do *corpus* estão vinculadas às restrições de tempo para a sua realização. A Figura 2 ilustra o processo de planejamento e as atividades da fase inicial do projeto de construção de um *corpus*.

FIGURA 2 – Processo de planejamento da construção de um *corpus*

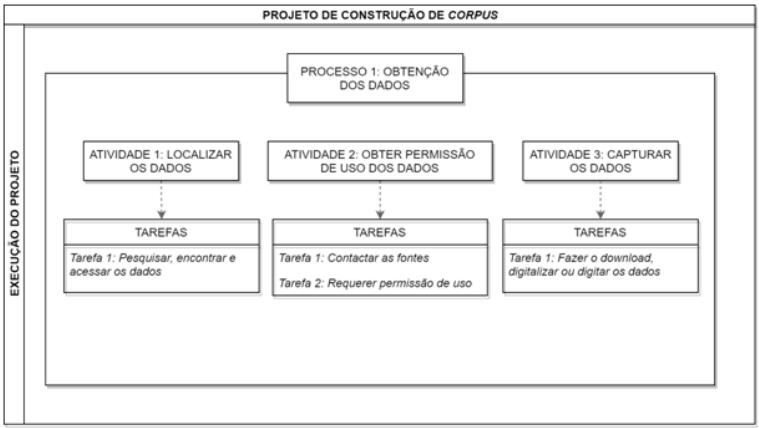


Fonte: Oliveira (2019, p. 44.)

Após planejar a construção do *corpus*, o pesquisador tem em mãos os parâmetros que guiarão a obtenção de dados linguísticos que irão compor o *corpus* e pode iniciar a fase de execução do projeto que consiste na realização das atividades relativas aos processos de obtenção,

preparação e armazenamento de dados do *corpus*. A Figura 3 exibe o processo de obtenção de dados para composição de um *corpus*.

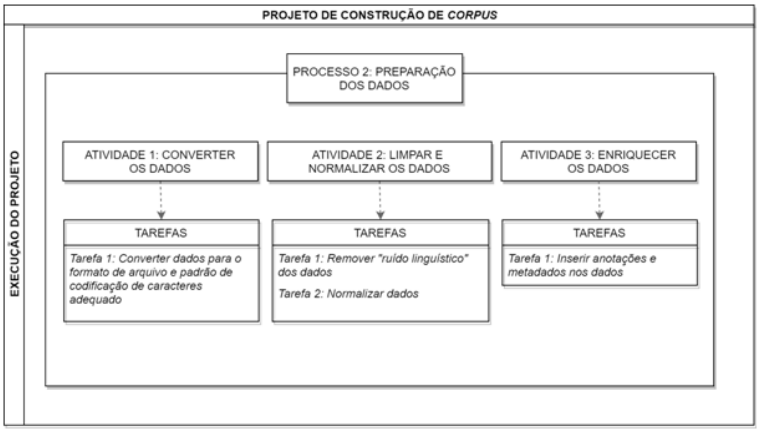
FIGURA 3 – O processo de obtenção de dados para composição de um *corpus*



Fonte: Oliveira (2019, p. 45.)

A primeira atividade desse processo consiste na pesquisa, localização e acesso aos materiais que comportam os dados desejados. Para Sinclair (1991), os dados podem ser encontrados, em suas versões originais, na forma eletrônica, impressa ou escrita à mão. Esse autor salienta que a obtenção de textos no formato eletrônico é a mais fácil e desejável comparada às demais, visto que exige um menor esforço do pesquisador no momento de adaptá-los para posterior processamento feito pelas ferramentas computacionais. Após a obtenção dos textos, o pesquisador precisa certificar-se de que eles são "úteis" para a inclusão em um *corpus*. A utilidade de um texto, para as pesquisas da LC, está associada, obrigatoriamente, à condição favorável dele para o processamento através de ferramentas computacionais e, opcionalmente, à integridade e ao enriquecimento dele. Esses aspectos configuram o processo de preparação dos dados do *corpus* e estão contemplados na Figura 4.

FIGURA 4 – O processo de preparação dos dados do *corpus*



Fonte: Oliveira (2019, p. 48.)

O processamento de um texto por meio de uma ferramenta computacional exige que ele esteja em um formato "compreensível pelos computadores" (*machine-readable*). Porém, os recursos que tornam um texto propício à interpretação humana podem ser prejudiciais ao processamento feito pelos computadores, já que estes ainda não possuem as mesmas capacidades de decodificação que os homens. Em virtude disso, na metodologia da LC, é indispensável a conversão das versões originais de textos para versões apropriadas ao trabalho que a máquina executa.

Após a realização da conversão, segundo Santos (2011), os arquivos podem apresentar resíduos como, por exemplo, números de páginas, informações de cabeçalhos e rodapés de páginas, anotações sobre a divisão das seções do texto, conteúdos de tabelas (que perdem o sentido ao serem desprovidos da estrutura da tabela) e erros de codificação de caracteres (resultantes da transposição de um padrão de codificação para outro). No contexto dos dados linguísticos, os resíduos presentes nos textos podem ser considerados como ruído linguístico e podem gerar problemas no que diz respeito à análise da frequência dos elementos linguísticos um *corpus*.

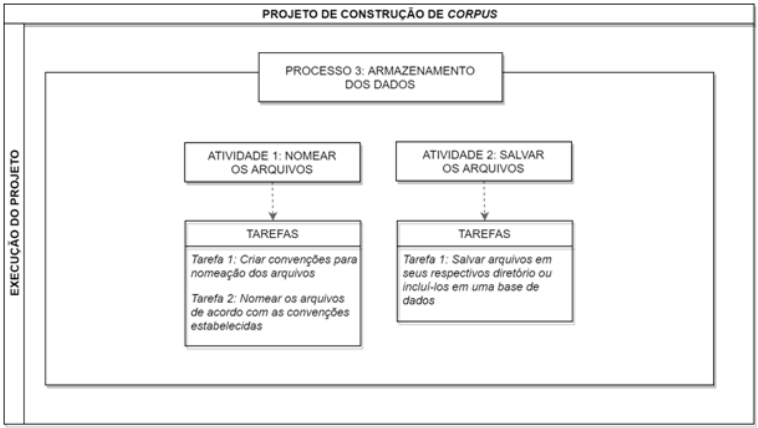
Conforme Gries (2009), a frequência de um elemento linguístico é base para a formação de listas de palavras que são utilizadas para a construção de listas de palavras-chave. As listas de palavras-chave, de

acordo com Tagnin (2015) e Edward (2015), apresentam os elementos linguísticos cujas frequências são estatisticamente relevantes a partir do resultado da comparação entre listas de palavras de um *corpus* de estudo e um *corpus* de referência. Pensando nessas relações, para que uma ferramenta computacional possa gerar listas com a frequência das palavras e, a partir disso, possa executar os cálculos que determinam a relevância dos elementos linguísticos, é necessário, em um primeiro momento, que ela identifique cada um dos elementos linguísticos presentes num texto. Essa identificação é realizada através do processo computacional conhecido na área de Processamento de Linguagem Natural como tokenização que consiste na segmentação das sentenças de um texto em elementos significativos, chamados *tokens*.

A tokenização de um texto que apresenta ruído linguístico pode gerar *tokens* sem nenhuma relação com qualquer elemento significativo da língua (por exemplo: *tokens* formados por partes de palavras que foram separadas incorretamente) e, consequentemente, gerar cálculos imprecisos sobre a frequência de um elemento, comprometendo a qualidade das listas provenientes da análise realizada por ferramentas computacionais. Por isso, uma das formas de reduzir ou de eliminar erros de tokenização, é a realização da limpeza e da normalização dos textos de um *corpus*. A limpeza de um texto, de acordo com Aluísio e Almeida (2006), consiste na remoção dos ruídos linguísticos. Já a normalização consiste na uniformização de palavras (em termos ortográficos), de siglas e de abreviaturas que possuem variações de escrita, a remoção de espaçamentos e de quebras de linhas desnecessários e a homogeneização de caracteres de pontuação do texto, como hifens, traços, aspas e apóstrofos.

Após a obtenção e preparação do *corpus*, os arquivos de texto que o constituem precisam ser nomeados e armazenados de modo que possam ser recuperados facilmente (NELSON, 2010). As atividades relacionadas à nomeação, ao armazenamento e à disponibilização dos dados de um *corpus* são realizadas no processo de armazenamento de dados, ilustrado na Figura 5.

FIGURA 5 – O processo de armazenamento dos dados do *corpus*

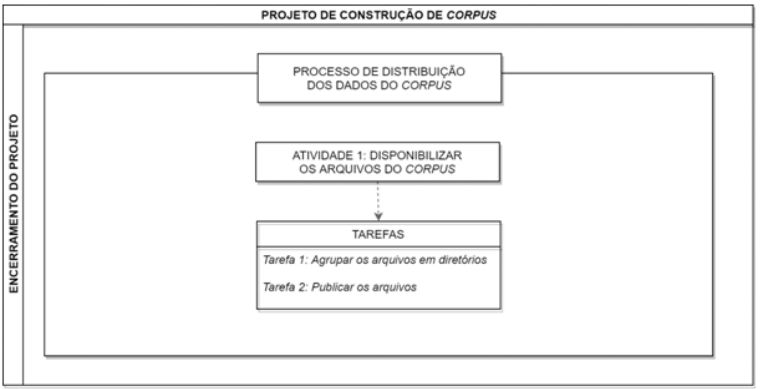


Fonte: Oliveira (2019, p. 59.)

A nomeação dos arquivos de texto para o armazenamento consiste na atribuição de um nome ao arquivo (*filename*) a fim de que ele possa ser identificado no sistema de arquivos do sistema operacional. Para facilitar a associação do nome de um arquivo ao seu conteúdo o pesquisador pode fazer uso de uma convenção de nomeação de arquivos, no inglês, *File Naming Conventions* (FNC). A convenção de nomeação de arquivos pode ser definida como um conjunto de regras que determina a estrutura da nomeação – constituída por diferentes segmentos que abrigam elementos informativos, ou seja, aqueles que fazem referência ao conteúdo, à descrição, ao contexto ou ao propósito dos arquivos. A ideia central de convencionar a nomeação é combinar informações suficientes para que a identificação do conteúdo do arquivo seja feita a partir de seu nome. Uma vez nomeado o arquivo de texto, comumente, é armazenado em um diretório organizado de acordo com a estrutura hierárquica adotada no projeto de construção do *corpus*, de modo que seja possível realizar uma associação significativa entre seu conteúdo e a categoria do diretório.

Depois das fases de planejamento e execução, a construção de um *corpus* segue para a fase de encerramento, que compreende o processo de distribuição dos dados do *corpus*, ilustrado na Figura 6, que trata da disponibilização dos dados do *corpus* com o propósito de serem processados pelas ferramentas computacionais, nas quais os métodos de análises e recuperação de informações são aplicados.

FIGURA 6 – O processo de distribuição dos dados do *corpus*



Fonte: Oliveira (2019, p. 64.)

A partir da sistematização apresentada, os autores da pesquisa debruçaram-se sobre a questão do tempo e do esforço na construção de *corpora*, procurando descortinar as principais dificuldades percebidas pela comunidade linguística quanto à execução das atividades de construção de *corpora*. No próximo tópico, encontramos as reflexões teóricas feita por eles sobre essas questões.

2.2 O tempo e o esforço na construção de um *corpus*

A introdução dos computadores no fazer linguístico e a Internet facilitaram a obtenção de dados linguísticos e o trabalho de construção de *corpora*. Porém, conforme mencionamos na introdução deste artigo, não é raro encontrarmos reclamações relacionadas à quantidade de tempo e esforço para se construir um *corpus*. De fato, a construção manual de *corpora* pode exigir bastante esforço e tempo do pesquisador, uma vez a sua intervenção, praticamente, é necessária em todas as fases do projeto (BAKER, 2010, p. 109), como podemos observar na descrição das atividades enumeradas a seguir:

1) **Definir o desenho do *corpus*, os recursos utilizados para lidar com ele e o cronograma referente ao projeto de construção do *corpus*:** exige que o pesquisador, basicamente, tome decisões teóricas e gerenciais;

2) **Localizar dados linguísticos:** a coleta manual de textos, em oposição à automática exige que o pesquisador busque fontes e verifique o acesso ao material (eletrônico, impresso ou em outro estado) que será utilizado em sua seleção. Os dados em formato eletrônico, por exemplo, são frequentemente localizados por meio de pesquisas feitas em sistemas de busca como o *Google*. Mesmo com a facilidade oferecida por esse tipo de sistema, a filtragem dos materiais encontrados (feita com base nos critérios do desenho do *corpus*) pode ser árdua devido ao grande volume e à qualidade das informações retornadas por *sites* como *Google*;
3) **Obter permissão de uso dos dados linguísticos:** para Santos (2011), essa atividade implica que o pesquisador precisará identificar a pessoa ou a entidade detentora dos direitos autorais de um texto, solicitar-lhe consentimento para usá-lo e, em seguida, aguardar retorno. O autor alerta que o consumo de tempo é ampliado nos casos em que o pesquisador precisa realizar várias tentativas de contato com o detentor para ter sucesso ou nas situações em que a solicitação de autorização tenha de partir de níveis hierárquicos superiores de uma instituição para que se obtenha uma resposta;
4) **Capturar os dados:** demanda a intervenção humana em uma escala que varia de acordo com o formato em que os dados estão quando são encontrados. Se estiverem no eletrônico, o pesquisador necessitará de intervir menos. Se os dados estiverem em materiais impressos ou escritos à mão, o pesquisador terá de convertê-los para o formato eletrônico, de preferência, por meio da digitalização com o auxílio de *scanners* e *softwares* de OCR[8] – uma das atividades que mais demandam esforço e tempo. Para Simske (2006), a precisão oferecida atualmente pelos OCRs na conversão de textos ainda é limitada e pode gerar erros referentes à troca, à inserção e à exclusão de caracteres, principalmente, quando a qualidade dos documentos originais é ruim. Isso faz com que autores como Nelson (2010), Kübler e Aston (2010), Santos (2011) e Bianchi (2012) preconizem a revisão manual cuidadosa dos dados resultantes da digitalização de textos por intermédio de *scanners* e OCRs para que se tenha certeza de que eles correspondem às suas versões originais;

[8] OCR é um *software* de reconhecimento ótico de caracteres. A sigla OCR vem do inglês *Optical Character Recognition*.

5) **Converter os dados:** exige pouco do pesquisador, apenas que ele manipule ferramentas computacionais que convertam os arquivos para o formato TXT e para o padrão *Unicode UFT-8*. Transformar um texto escrito em PDF para TXT, por exemplo, pode ser feito de forma gratuita por meio de serviços *on-line* de conversão, como o *Lightpdf* (lightpdf.com), entre outros. E a mudança para o padrão *Unicode UTF-8* pode ser efetuada por ferramentas como o *EncodeAnt* (ANTHONY, 2016);
6) **Limpar e normalizar os dados:** requer que o pesquisador proceda como auditor no que diz respeito aos dados do *corpus* para identificação e posterior eliminação ou correção das anomalias (ruído linguístico). A limpeza e a normalização estão diretamente relacionadas a algumas variáveis: volume e qualidade dos dados (resultante dos métodos de captura, conversão e codificação dos textos), finalidade (necessidades) da pesquisa e, por fim, métodos escolhidos para a execução da limpeza e da normalização. Vale lembrar que as tarefas em questão podem ganhar proporções gigantescas e, portanto, serem difíceis no caso de *corpora* compostos por grandes volumes de informação;
7) **Enriquecer os dados:** pressupõe que o pesquisador realize uma conferência no que alude à etiquetagem automática de *corpora*. Conforme Neumann e Hansen-Schirra (2012), o enriquecimento de *corpora* grandes depende da etiquetagem automática, pois o processamento manual de grandes volumes de dados é praticamente inviável. Semino e Short (2004) reforçam essa ideia ao afirmarem que até mesmo a etiquetagem manual de *corpora* pequenos é extremamente demorada. Contudo, a utilização de ferramentas computacionais para a etiquetagem não dispensa a intervenção manual do pesquisador (MEYER, 2004), uma vez que *taggers* e *parsers* não conseguem alcançar uma precisão total no processamento dos dados. Para Meyer (2004), a precisão dos etiquetadores, geralmente, é comprometida pela inconsistência dos dados (dados não limpos ou normalizados) e pela dificuldade que apresentam para lidar com as características idiossincráticas (GARSIDE; SMITH, 1997 *apud* MEYER, 2004) da linguagem humana. Em decorrência dos possíveis erros, o resultado da etiquetagem automática precisa ser conferido pelo pesquisador (*post-editing*) com o objetivo de corrigir e resolver possíveis

ambiguidades nas etiquetas (LEECH, 2005). Segundo Bianchi (2012), essa atividade é desenvolvida manualmente e exige muito tempo e esforço.

8) **Nomear arquivos:** prevê que o pesquisador atribua nomes aos arquivos do *corpus*, de preferência, após estabelecer uma convenção. Nessa atividade, o pesquisador poderá ter de checar como foi definida a estrutura da convenção quando for nomear cada arquivo do *corpus* caso não consiga memorizá-la. Ademais, ele precisará selecionar a informação mais adequada para compor cada segmento da estrutura do nome do arquivo;
9) **Salvar arquivos:** requer pouco do pesquisador quando é feito por meio da alocação dos arquivos em diretórios de um sistema de arquivos, de acordo com a hierarquia estabelecida em um projeto. Entretanto, segundo Sedlar (2005), em situações em que o pesquisador decida salvar os arquivos em uma base de dados, a execução da tarefa dependerá do uso de uma ferramenta computacional que ofereça a interface necessária para a inclusão dos arquivos no banco de dados. O pesquisador poderá optar pelo uso de uma ferramenta já existente ou pela criação de uma ferramenta customizada para o seu projeto. No primeiro caso, ele precisará de um esforço adicional para a escolha de uma ferramenta e para a aprendizagem do seu uso. No segundo, além do esforço para aprender a usar a ferramenta, ele investirá recursos financeiros, tempo e esforço por ter de contratar um profissional para desenvolver a aplicação ou por desenvolvê-la por conta própria;
10) **Disponibilizar arquivos:** demanda pouco esforço e tempo do pesquisador quando ele opta por apenas copiar os arquivos em dispositivos de armazenamento de dados. Já a publicação *on-line* do *corpus* pode requerer recursos financeiros (por exemplo, para a contratação de serviços de hospedagem ou de armazenamento de dados na nuvem) e, ainda, mais esforço e tempo do pesquisador, pois ele deverá se preocupar com questões: como a escolha de um local para a publicação, a compactação dos arquivos, a disponibilização de documentação sobre o *corpus* com informações suficientes para que a sua utilização seja feita por outros pesquisadores e a explicitação de uma licença de uso dos dados do *corpus*.

A fim de contornar as dificuldades relacionadas à coleta de dados, os pesquisadores podem adotar os *web corpora* ou *corpora ad-hoc*, que são *corpora* compostos por dados coletados da Internet de forma automática. Nesse caso, eles precisam lançar mão de ferramentas computacionais, como o *WebBootCat* (BARONI *et al*., 2006), o *WebCorp Linguist's Search Engine* (KEHOE; GEE, 2007) e o *Bootcat* (BARONI; BERNARDINI, 2004), que, segundo Aluísio e Almeida (2006, p. 168), utilizam motores de busca (*Google*, por exemplo) e um "pequeno conjunto de itens léxicos, denominados sementes (*seeds*)" para efetuarem a compilação.

Schäfer e Bildhauer (2013) consideram que a realização de inferências estatísticas a partir de *corpora* construídos com base em resultados de pesquisas de motores de busca não é uma boa prática de pesquisa, pois os buscadores privilegiam a precisão (*precision*) em detrimento da revocação (*recall*),[9] podem ser influenciados por fatores econômicos,[10] usam variáveis como a língua e a localização de quem fez a pesquisa e realizam alterações automáticas nas expressões fornecidas para a pesquisa (otimizam as expressões por meio de reduções ou expansões). Além disso, as buscas não podem ser reproduzidas devido à constante entrada e saída de conteúdos (indexação) na Internet.

Mais do que as questões relacionadas aos critérios de recuperação de informações dos motores de busca, Schäfer e Bildhauer (2013) acreditam que a opção pelo uso de *corpora* provenientes de métodos automáticos de coleta requer precaução extra do pesquisador no que diz respeito a aspectos, tais como: remoção do *boilerplate* (quais partes do documento foram removidas) e do ruído linguístico dos documentos (quais os tipos de ruídos existentes e qual a precisão da remoção deles); introdução de ruído linguístico (quais ruídos foram introduzidos após o processamento dos documentos); remoção de arquivos duplicados (*deduplication*) (quais documentos foram removidos e quais foram os

[9] Consoante Rubi (2009), a revocação "pode ser mensurada por meio da relação entre o número de documentos relevantes sobre determinado tema, recuperados pelo sistema de busca, e o número total de documentos sobre o tema, existentes nos registros do mesmo sistema" (RUBI, 2009, p. 85). A precisão "pode ser mensurada por meio da relação entre os documentos relevantes recuperados e número total de documentos recuperados" (RUBI, 2009, p. 85-86).

[10] Por exemplo, os conteúdos patrocinados.

critérios de remoção) e a forma pela qual a amostragem dos dados foi criada. Além das precauções de Schäfer e Bildhauer (2013), na literatura da LC, identificamos outros problemas que podem surgir numa situação em que se opta por automatizar a coleta de um *corpus*:

1) **Problema da replicabilidade:** está relacionado à mutabilidade dos dados na Internet. Para Mcenery e Hardie (2011), os estudos com *corpora* coletados de forma automática na Internet são difíceis de serem replicados com o passar do tempo em virtude de haver constante mudança de dados na rede;
2) **Problema dos falso-positivos:** os falso-positivos são *tokens* e *types* que não possuem relação com qualquer elemento significativo de uma língua alvo de pesquisa, provenientes de erros de tokenização provocados pelo ruído linguístico de um *corpus*. Conforme Schäfer e Bildhauer (2013), os *web corpora* tendem a conter um alto nível de ruído linguístico;
3) **Problema da amostragem:** refere-se à incontrolabilidade e arbitrariedade da escolha dos dados dos *web corpora* no universo heterogêneo (RENOUF, 2007) dos dados disponíveis na Internet. De acordo com Schäfer e Bildhauer (2013), a coleta automática de documentos, geralmente, não segue um esquema amostral preestabelecido. Para esses autores, na melhor das hipóteses, *web corpora* são uma amostra randômica de dados da Internet, cuja composição exata é desconhecida e precisará ser estabelecida após a sua compilação. Para Mcenery e Hardie (2011), um dos aspectos do desconhecimento do conteúdo de *web corpora* é a dificuldade em determinar o gênero textual dos documentos coletados sem tê-los lido;
4) **Problema legal:** consiste no *download* e uso de textos de *sites* da Internet e na sua distribuição como parte de um *corpus* sem o consentimento dos autores (MCENERY; HARDIE, 2011, p. 58). Segundo Mcenery e Hardie (2011), as leis de direito autoral aplicam-se aos textos coletados automaticamente da Internet do mesmo modo que se aplicam aos materiais impressos e, por isso, podem gerar as mesmas implicações legais que outras formas de construção de *corpora*;

5) **Problema da violação da integridade dos textos:** Schäfer e Bildhauer (2013) argumentam que o processamento automático de coleta dos textos introduz erros nos dados originais deles que podem reduzir a qualidade dos dados. Para exemplificar, os autores mencionam a remoção automática de dados duplicados que, se for feita no interior dos textos, no nível dos parágrafos, pode implicar a inclusão de materiais incompletos em um *corpus*;
6) **Problema das consultas em massa *(batch or bulk requests*):** a obtenção dos dados dos *web corpora* depende do envio das sementes (ALUÍSIO; ALMEIDA, 2006, p. 168) que serão utilizadas como parâmetro de consulta pelos motores de busca. Schäfer e Bildhauer (2013) explicam que, no intuito de evitar abusos, os motores de busca apresentam restrições para o processamento gratuito de grandes volumes de consultas automáticas. Portanto, a construção de *web corpora* de grandes proporções por meio da coleta automática de dados demandará do pesquisador o pagamento pelo serviço de busca que ultrapassa os limites de consultas dos motores até que o volume de dados necessário para os *corpora* seja atingido.

A compilação automática de *corpora* pode ser "adequada para uma grande variedade de propósitos" (MCENERY; HARDIE, 2011, p. 8)[11] e os *web corpora* são extremamente valiosos para as pesquisas que demandam a análise de grandes volumes de informação (BERGH; ZANCHETTA, 2008, p. 320) e em que "o valor do volume dos dados se sobrepõe à qualidade proporcionada pela sua limpeza" (SCHÄFER; BILDHAUER, 2013, p. 126),[12] ainda que apresentem problemas e possam ter sua utilidade considerada limitada por grupos de linguistas (RUNDELL; KILGARRIFF, 2011, p. 262). Para Rundell e Kilgarriff (2011), os *web corpora* adequam-se, por exemplo, às pesquisas lexicográficas para a criação de dicionários gerais em que "os benefícios da abundância de dados superam os problemas dos *web corpora*" (RUNDELL; KILGARRIFF, 2011, p. 262).[13]

---

[11] Original: "suitable for a wide variety of purposes".
[12] Original: "values the amount of available data more highly than the cleanliness of a corpus".
[13] Original: "the benefits of abundant data outweigh most of the perceived disadvantages of web corpora".

3 Metodologia

Como explicado na introdução deste artigo, o objetivo principal da pesquisa nele descrita é determinar os efeitos da incorporação do *ToGatherUp* no esforço necessário para a construção manual de *corpora*. A forma encontrada pelos autores para atingirem esse propósito foi a realização de um experimento de comparação entre os esforços necessários para a construção de duas versões idênticas do CoCLI, sendo que o projeto de elaboração de uma delas contou com a incorporação do *ToGatherUp* e o outro não. Para que a confrontação fosse possível, em um primeiro momento, eles estabeleceram um critério objetivo e um método para a medição do esforço das atividades de cada um dos projetos de construção de *corpora*. Na sequência, à medida que executaram a construção dos *corpora*, tabularam os esforços necessários para a realização de cada uma das atividades dos projetos. Por fim, realizaram o experimento por meio de um teste estatístico para a comparação dos dados tabulados. Nos tópicos desta seção, explanamos cada um desses passos.

3.1 Como mensurar o esforço?

Apesar de o esforço ser um tema recorrente entre os autores da LC, na revisão da literatura da área, os autores não identificaram trabalhos que tenham se debruçado sobre a sua investigação. Por essa razão, tendo em vista o resultado dessa averiguação e o objetivo da pesquisa, eles precisaram formular métricas e um método de mensuração do esforço dos projetos de construção de *corpora*. A criação da métrica feita por eles baseou-se no conceito de medição proposto por Fenton e Bieman (2014), no livro "*Software Metrics: A Rigorous and Practical Approach*". Conforme esses dois autores, a "medição é o processo pelo qual números ou símbolos são associados aos atributos[14] de uma entidade[15] do mundo real, de modo que seja possível descrevê-los de acordo com um conjunto

[14] Os atributos são as características ou propriedades das entidades.

[15] As entidades são representações de objetos e eventos do mundo real. Por exemplo: uma pessoa, um lugar, um objeto, uma ideia, um produto, um processo ou uma atividade. Do mesmo modo que uma pessoa (entidade) pode ser descrita a partir de suas características (por exemplo: altura, sexo e idade), as atividades podem ser descritas a partir de seus atributos (por exemplo: duração, *inputs* e *outputs*).

de regras[16] bem definidas"[17] (FENTON; BIEMAN, 2014, p. 5) e compará-los com atributos semelhantes de outras entidades.

A partir do conceito de métrica citado, os pesquisadores assumiram as atividades dos projetos de construção de *corpora* como entidades e estabeleceram que o *input*,[18] o *output*[19] e o tempo de duração seriam os seus atributos. Deste modo, eles passaram a dispor de elementos para mensurar o esforço das atividades e foram capazes de compor uma métrica,[20] o Esforço da Atividade (EA), que quantifica, em segundos, o esforço despendido para a completude de uma atividade de um projeto de construção de *corpus*. O Esforço da Atividade (EA) estabelece que o esforço de uma atividade é igual ao quociente entre o intervalo de tempo decorrido entre o início e o fim da atividade (a duração da atividade) e o resultado da atividade, ou seja, a sua completude. Nessa relação, o tempo do pesquisador é o *input*[21] que equivale ao tempo de duração da atividade e o resultado da atividade é o *output* que corresponde ao número 1 (um)[22] – forma que os autores estabeleceram para quantificar e denotar a completude da atividade.[23] A Figura 7 ilustra os 3 atributos de uma atividade no escopo do EA.

---

[16] As regras ditam como a medição deve ser realizada.

[17] Original: "Measurement is the process by which numbers or symbols are assigned to attributes of entities in the real world in such a way so as to describe them according to clearly defined rules".

[18] Os *inputs* são as entradas necessárias para a realização de uma atividade. No caso, realizamos um recorte nas entradas que considerou apenas o tempo despendido pelo criador do *corpus* na execução da atividade.

[19] Os *outputs* são os produtos ou entregas (resultados) de uma atividade.

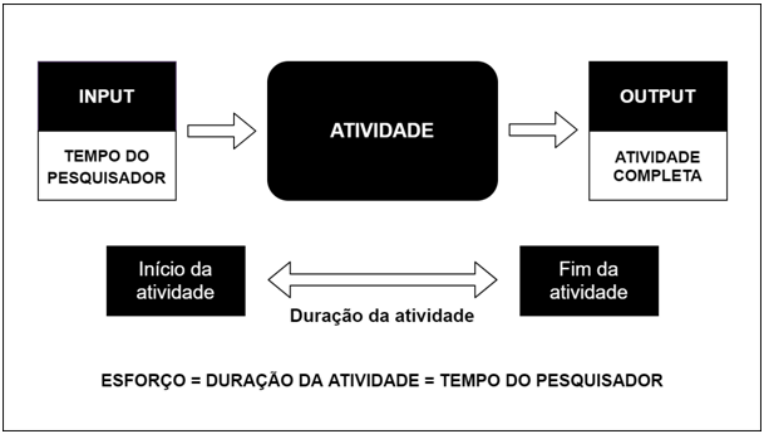
[20] Métricas são unidades de medidas criadas a partir de medições.

[21] Apesar de utilizarem somente o tempo do pesquisador como *input* da atividade, os autores estão cientes da existência de outros *inputs* necessários para a realização de uma tarefa, como o conhecimento do pesquisador. A decisão pelo uso do tempo do pesquisador justifica-se pelo fato de o tempo ser, geralmente, reportado como o recurso primário para a execução de uma atividade. Ademais, o tempo do pesquisador apresenta-se como um *input* quantificável e de fácil mensuração em relação aos *inputs* mais abstratos, como o conhecimento.

[22] De acordo com o raciocínio aplicado, a não completude da atividade corresponderia ao número 0 (zero).

[23] A completude de uma atividade pode ser compreendida como a finalização de 100% de suas tarefas.

FIGURA 7 – Atributos do modelo de regras do EA



Fonte: Oliveira (2019, p. 77.)

Ao criarem o EA, os autores passaram a dispor da variável básica para o cálculo do esforço investido em um projeto de construção de *corpora*. A forma de calcular esse esforço foi a criação de uma outra métrica, o Esforço Total do Projeto (ETP). O ETP é uma métrica que quantifica, em segundos, o esforço despendido para a completude[24] de um projeto de construção de um *corpus* e corresponde à soma de todos os EAs das atividades do projeto. Assim, o ETP pode ser expresso da seguinte maneira: ETP = EA 1 + EA 2 + EA 3 + EA 4 + EA *n*. Em suma, a comparação entre os esforços empregados na construção de cada versão do CoCLI deu-se pela comparação entre os ETP de cada um dos projetos. Para encontrar o ETP de cada projeto, em um primeiro momento, foi calculado o EA de cada atividade dos projetos. O método para o cálculo do ETP referente à construção da versão do CoCLI que não passou pela intervenção do *ToGatherUp* consistiu nos passos a seguir:

1. Identificação das atividades realizadas para a construção do projeto de acordo com a sistemática de organização de projetos proposta pelos autores da pesquisa. As atividades identificadas foram: a) localização dos dados; b) permissão de uso dos dados; c) captura dos dados; d) conversão dos dados; e) limpeza e normalização dos

[24] A completude de um projeto pode ser compreendida como a finalização de 100% das suas atividades.

dados; f) salvamento de arquivos; g) enriquecimento dos dados; h) nomeação dos arquivos;

2. Cálculo do EA das atividades identificadas. Para melhor compreensão, os autores atribuíram uma sigla para cada EA calculado. Desse modo, obtiveram a seguinte lista: a) Esforço da Atividade da localização dos dados (EALD); b) Esforço da Atividade da obtenção de permissão de uso dos dados (EAOPD); c) Esforço da Atividade de captura dos dados (EACD); d) Esforço da Atividade de conversão dos dados (EACVD); e) Esforço da Atividade de limpeza e normalização dos dados (EALND); f) Esforço da Atividade de salvamento de arquivos (EASA); g) Esforço da Atividade de enriquecimento dos dados (EAED); h) Esforço da Atividade de nomeação dos arquivos (EANA).
3. Aplicação do modelo de cálculo do ETP, expresso por: ETP1[25] = EALD + EAOPD + EACD + EACVD + EALND + EASA + EAED + EANA.

No que tange ao método para o cálculo do ETP concernente à elaboração da versão do CoCLI que contou com a incorporação do *ToGatherUp*, os passos foram:

1. Identificação das atividades realizadas para a construção do projeto de acordo com a sistemática de organização de projetos proposta pelos autores da pesquisa. As atividades identificadas[26] foram: a) localização dos dados; b) permissão de uso dos dados; c) captura dos dados; d) conversão dos dados; e) limpeza e normalização dos dados; f) cadastramento de textos;[27]
2. Cálculo do EA das atividades identificadas. De forma análoga ao passo dois do método citado anteriormente, os autores atribuíram siglas para cada EA calculado. Logo, obtiveram a seguinte lista:

[25] O ETP1 diz respeito ao projeto não intervencionado pelo *ToGatherUp*.

[26] As atividades a, b, c e d são comuns aos dois projetos. As atividades de salvamento, nomeação de arquivos e enriquecimento dos dados foram automatizadas pelos recursos do *ToGatherUp* e, por isso, não geraram seus respectivos EAs. Portanto, não as incluímos no cálculo do projeto intervencionado pelo *ToGatherUp*.

[27] O cadastramento de texto é uma atividade específica da construção de *corpora* no *ToGatherUp*.

a) EALD; b) EAOPD; c) EACD; d) EACVD; e) EALND; f) Esforço da Atividade de cadastramento de textos (EACT).

3. Aplicação do modelo de cálculo do ETP, expresso por: ETP2[28] = EALD + EAOPD + EACD + EACVD + EALND + EACT.

Em ambos os projetos, os pesquisadores fizeram a medição da duração das atividades, necessária para a obtenção do EA de cada uma das atividades, com o uso de um cronômetro disponível na interface do *ToGatherUp*. Para a obtenção da quantidade de segundos relativa à duração de uma atividade, o cronômetro foi acionado assim que a atividade foi iniciada e paralisado logo após a conclusão dela. A informação[29] fornecida pelo cronômetro (Instrumento 1) foi tabulada em uma planilha do *Google* (Instrumento 2), que serviu para a extração do conjunto de dados (*dataset*) analisado no experimento da pesquisa.

Além do EA e do ETP, os pesquisadores criaram a métrica Esforço Total de Coleta do Texto (ETCT) para determinar o esforço despendido para a inclusão de uma única unidade de texto em um *corpus.* O ETCT pode ser expresso, de forma semelhante ao ETP, pela fórmula ETCT = EA 1 + EA 2 + EA 3 + EA 4 + EA *n.* Porém, o contexto de aplicação da expressão é limitado somente aos EAs de uma única unidade textual.

3.2 O *ToGatherUp*

O *ToGatherUp*[30] é uma ferramenta *on-line* (http://www.ileel.ufu.br/togatherup) desenvolvida pelos autores da pesquisa aqui retratada que oferece suporte a projetos de construção manual de *corpora*. As principais funcionalidades da ferramenta são a inserção automática de cabeçalho de metadados nos arquivos do *corpus*, a nomeação do

[28] O ETP2 alude ao projeto intervencionado pelo *ToGatherUp*.

[29] O tempo decorrido entre o início e o fim da atividade. Ou seja, a duração da atividade.

[30] O nome *ToGatherUp* surgiu da associação entre o ato de construir um corpus e o verbo frasal *gather up*, da língua inglesa, que, de acordo com o *Macmillan Dictionary* significa "pegar coisas de lugares diferentes e colocá-las juntas", no original "to pick up things from several different places and put them together" (MACMILLAN DICTIONARY, 2018). Para reforçar a associação, no *design* da logomarca da ferramenta, os autores incluíram o símbolo 輯, um ideograma da língua japonesa que, conforme Jisho (http://jisho.org), um dicionário japonês *on-line*, pode ser traduzido para as seguintes palavras da língua inglesa: a) *gather*; b) *collect;* c) *compile*.

arquivo de acordo com uma convenção preestabelecida pelo criador do *corpus* e o armazenamento do arquivo no diretório correspondente ao seu posicionamento na estrutura hierárquica do projeto. Além dessas funcionalidades, o *ToGatherUp* exibe ao pesquisador uma interface em que é possível visualizar estatísticas sobre a quantidade de textos e palavras coletadas, conferindo a ele maior controle em relação ao andamento de um projeto. O *ToGatherUp* possui os seguintes recursos:

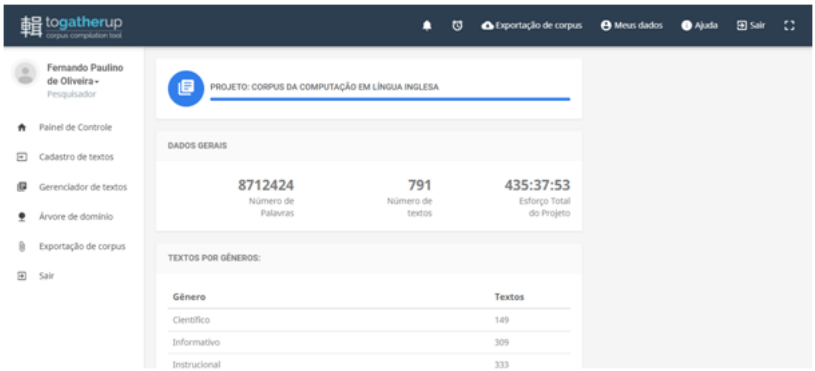
1) Painel de Controle (*Data Overview*): permite a visualização da quantidade total de palavras e de textos de um *corpus*, do ETP, das quantidades de palavras e textos para cada um dos gêneros, tipos textuais, meios de distribuição, áreas e subáreas de um *corpus* e facilita o acompanhamento visual da evolução da coleta de textos de um *corpus*;
2) Cadastro de Textos (*Data Entry*): apresenta um formulário com os campos para a entrada dos dados do texto. Os campos são: Subárea, Título, Língua, Fonte, Gênero textual, Tipos textuais, Meio de Distribuição e ETCT;
3) Gerenciador de Textos (*Data Manager*): interface que exibe uma lista com os textos de um *corpus*, em forma de tabela, e possibilita a localização (pesquisa) de um texto (ou textos) a partir dos metadados dele;
4) Árvore de Domínio (*Domain Tree*): interface para a visualização da organização hierárquica adotada no projeto de um *corpus*;
5) Exportação de *Corpus* (*Data Exporter*): funcionalidade que exporta os arquivos de um *corpus* em diretórios organizados de acordo com a hierarquia do projeto.

Na sequência, abordamos cada um dos referidos recursos do *ToGatherUp*, ilustrando-os com exemplos extraídos do projeto de construção do CoCLI em que ocorreu a incorporação do *ToGatherUp*.

3.2.1 Painel de Controle

O Painel de Controle é a interface principal do *ToGatherUp* e exibe as informações gerais do projeto de construção de um *corpus*. O objetivo dele é oferecer ao pesquisador uma visão geral da evolução da coleta de textos do *corpus*. A Figura 8 ilustra parcialmente o Painel de Controle do *ToGatherUp* com informações do projeto do CoCLI.

FIGURA 8 – Painel de Controle com informações do CoCLI



Fonte: *ToGatherUp*.

O Painel de Controle contém outros cinco painéis: a) Dados gerais, que apresenta o número total de palavras, a quantidade total de textos e o ETP de um *corpus*; b) Textos por Gêneros, que apresenta as quantidades totais de textos para cada gênero textual que compõe um *corpus*; c) Textos por Tipos Textuais, que apresenta o número total de textos para cada tipo textual de um *corpus*; d) Textos por Meios de Distribuição, que apresenta, de maneira discriminada, os meios de comunicação em que os textos de um *corpus* foram obtidos durante sua coleta, quantificando-os; e) Textos por Áreas e Subáreas, que fornece a visão da quantidade de textos e de palavras para cada item da hierarquia adotada num projeto de construção de *corpus*.

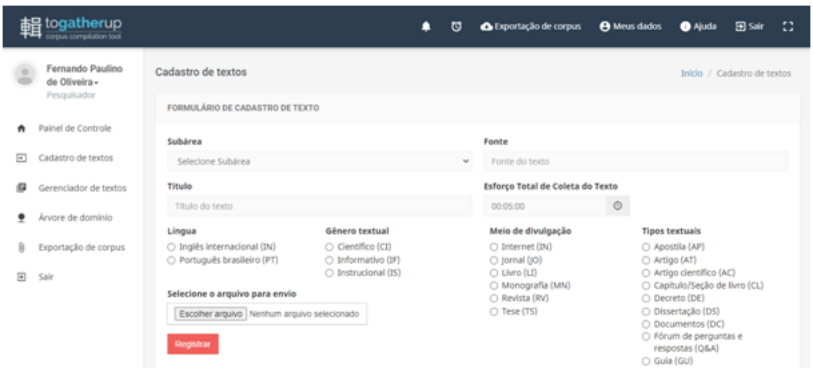
3.2.2 Cadastro de Textos

No *ToGatherUp*, a inclusão de um texto em um *corpus* é realizada através do recurso denominado Cadastro de Textos, presente na Figura 9. O Cadastro de Textos é uma interface *web* que apresenta um formulário, a ser preenchido pelo pesquisador de forma manual, composto pelos seguintes campos:[31] (a) Subárea; (b) Título; (c) Língua; (d) Fonte; (e) Gênero Textual; (f) Tipos Textuais; (g) Meio de Distribuição; (h) ETCT.

[31] Os campos citados foram estabelecidos para o projeto do CoCLI. Os campos do Cadastro de Textos podem ser definidos pelo pesquisador no momento da configuração do projeto no *ToGatherUp*. A data de publicação do texto e a sua autoria são exemplos de informações que podem ser incluídas durante a configuração do projeto.

Além desses campos, o formulário apresenta, ainda, a opção para que o pesquisador possa anexar o arquivo do texto.

FIGURA 9 – Formulário de Cadastro de Textos do *ToGatherUp*



Fonte: *ToGatherUp*.

Ao incluirmos um texto no *corpus* por meio do Cadastro de Textos, o *ToGatherUp* desencadeia, de forma automática, as atividades: a) Atividade 1: Registro dos metadados do texto no banco de dados; b) Atividade 2: Nomeação[32] do arquivo do texto; c) Atividade 3: Inserção de cabeçalho no arquivo do texto; d) Atividade 4: Armazenamento do arquivo do texto. Na sequência, descrevemos cada uma dessas atividades e como elas foram configuradas no projeto de construção do CoCLI.

a) Atividade 1: Registro dos metadados do texto no banco de dados

Ao armazenar os textos de um *corpus*, o pesquisador precisa estabelecer padrões descritivos que otimizem o acesso a eles, a recuperação e o reuso deles. Para atender essa necessidade, o *ToGatherUp* faz uso de metadados para a catalogação dos textos dos *corpora*. A utilização de metadados surgiu no âmbito das Ciências da Informação

[32] Na realidade, o que ocorre é uma renomeação, porque, para que seja possível a sua submissão no *ToGatherUp*, o arquivo precisa ter sido previamente salvo pelo pesquisador. O *ToGatherUp* desconsidera qualquer que seja o nome dado a um arquivo submetido a ele e procede com a sua renomeação em conformidade com os metadados do texto e com a convenção de nomeação de arquivos do projeto.

como uma solução para a organização de dados. Para Alves (2010), os metadados podem ser definidos como:

> [...] atributos que representam uma entidade (objeto do mundo real) em um sistema de informação. Em outras palavras, são elementos descritivos ou atributos referenciais codificados que representam características próprias ou atribuídas às entidades; são ainda dados que descrevem outros dados em um sistema de informação, com o intuito de identificar de forma única uma entidade (recurso informacional) para posterior recuperação (ALVES, 2010, p. 47).

Para o projeto do CoCLI, cada campo do formulário de Cadastro de Textos correspondeu à um metadado estabelecido de acordo com os critérios de desenho. Desse modo, os metadados do CoCLI apresentam-se conforme o Quadro 1.

QUADRO 1 – Metadados do CoCLI

Metadados	Descrição
(a) Subárea	Informa a subárea do texto.
(b) Título	Informa o nome dado para o texto.
(c) Língua	Informa o idioma em que o texto foi escrito.
(d) Fonte	Informa a origem do texto.
(e) Gênero textual	Informa o gênero textual do texto.
(f) Tipos textuais	Informa o tipo textual do texto.
(g) Meio de distribuição	Informa o meio em que o texto foi divulgado.
(h) ETCT	Informa o esforço total referente à soma de todos os EAs realizados para a inclusão de uma unidade textual no *corpus*.[33]

Fonte: Oliveira (2019, p. 92.)

[33] A obtenção do ETCT depende do registro do EA de cada uma das atividades necessárias para a coleta do texto. É importante lembrar que o ToGatherUp não apresenta uma forma de registro para cada um dos EAs. O software tem somente um cronômetro que pode ser utilizado para a captura da duração de cada atividade, que pode ser registrada em um tipo de controle escolhido pelo pesquisador.

Além dos metadados do Quadro 1, o *ToGatherUp* registra, de forma automática, o um conjunto de metadados sem que ocorra a intervenção do pesquisador. O Quadro 2 apresenta esses metadados.

QUADRO 2 – Metadados gerados de forma automática pelo *ToGatherUp*

Metadados	**Descrição**
(a) Domínio	Informa o domínio do texto (área do conhecimento/ especialidade a qual pertence).[34]
(b) Número de palavras	Informa o número de palavras do texto.[35]
(c) Data da inclusão	Informa a data e a hora em que o texto foi incluído no *corpus*.[36]
(l) Identificador do arquivo (ID)	Informa o número de identificação do texto no banco de dados do *ToGatherUp*.[37]

Fonte: Oliveira (2019, p. 92.)

b) Atividade 2: Nomeação dos arquivos dos textos

O *ToGatherUp* faz a nomeação automática dos textos de um *corpus* durante a submissão deles pelo Cadastro de Textos de acordo com os metadados do texto e com uma convenção de nomeação de arquivos definida durante a configuração do projeto no sistema. Com base na convenção estabelecida para a nomeação dos textos do CoCLI, um dos textos desse *corpus* foi nomeado, por exemplo, desta forma: IN-CO-IF-AT-IN-25Sep2017-797.txt. O nome do arquivo é constituído por sete partes distintas, separadas por hífen, e finalizado com a extensão correspondente ao formato dele (.txt). Cada uma das partes é formada por uma abreviação que se associa a um metadado do texto:

a) a primeira (IN) informa a língua do texto. Para a língua inglesa, foi utilizada a abreviação IN;

[34] O domínio do texto é estabelecido durante as configurações do projeto no *ToGatherUp*. Por essa razão, o *ToGatherUp* é capaz de incluí-lo, automaticamente, como um metadado.
[35] O *ToGatherUp* possui um algoritmo que contabiliza a quantidade de palavras do texto.
[36] O *ToGatherUp* considera a data e a hora do servidor em que o sistema está instalado. Por isso, o pesquisador não precisa informar esses dados.
[37] O ID é gerado de forma incremental e automática pelo *ToGatherUp*.

b) a segunda (CO) diz respeito ao domínio (área do conhecimento/especialidade a que pertence o texto). Como o CoCLI é do domínio da Computação, foi utilizada CO para abreviá-lo;

c) a terceira (IF) refere-se ao gênero do texto. A abreviação CI foi utilizada para o gênero científico, a IF para o gênero informativo e a IS para o gênero instrucional;

d) a quarta (AT) alude ao tipo do texto. As abreviações referentes aos tipos textuais dos textos do CoCLI foram estabelecidas da seguinte maneira:
   - Apostila (AP);
   - Artigo (AT);
   - Artigo científico (AC);
   - Capítulo/Seção de livro (CL);
   - Decreto (DE);
   - Dissertação (DS);
   - Documentos (DC);
   - Fórum de perguntas e respostas (Q&A);
   - Guia (GU);
   - Livro (LV);
   - Manual (MA);
   - Monografia (MN);
   - Norma técnica (NR);
   - Nota técnica (NT);
   - Notícia (NO);
   - Portaria (PA);
   - Relatório (RL);
   - Reportagem (RP);
   - Tese (TS);
   - Transcrição (TR);
   - Tutorial (TT).

e) a quinta parte (IN) é relativa ao meio de divulgação do texto. Como todos os textos do CoCLI são provenientes da Internet, foi utilizada a abreviação IN para representá-la;

f) a sexta parte (25Sep2017) informa a data de coleta do texto;

g) a sétima parte (797) indica o identificador (ID) do texto no banco de dados do *ToGatherUp*. Cada texto recebe um ID único ao ser registrado no banco de dados do sistema, o que evita a possibilidade de que textos com metadados idênticos recebam um mesmo nome.

c) Atividade 3: Inserção de cabeçalho nos arquivos de texto

Os metadados dos textos do CoCLI foram usados pelo *ToGatherUp* para a criação e inserção automática de cabeçalho nos arquivos dos textos. Para que isso fosse possível, nas configurações da ferramenta, foi necessário estabelecer a estrutura do cabeçalho a ser utilizada que, na pesquisa, conteve apenas a origem do texto e a sua data de inclusão no *corpus*. Com base nisso, o *ToGatherUp* procedeu com a inserção do cabeçalho nos textos, alimentando-os com os metadados fornecidos no Cadastro de Textos da ferramenta.

d) Atividade 4: Armazenamento do arquivo do texto

O armazenamento de arquivos através de métodos tradicionais comuns no cotidiano das organizações e no gerenciamento de informações pessoais é natural. No entanto, Dourish (2003, p. 4) aponta que estudos realizados por Barreau e Nardi (1995) e por Kaptelinin (1996) revelam que essa prática é problemática, pois dificulta a reorganização das informações quando elas assumem funções diferentes das originais ou quando elas não se adequam a somente um dos *loci* de armazenamento.

Considerando essa problemática, o *ToGatherUp* foi desenvolvido de modo que ele fosse capaz tanto de armazenar os textos do CoCLI de acordo com a Árvore de Domínio da Computação (uma estrutura hierárquica fixa) como de reorganizá-los seguindo outras configurações hierárquicas. A capacidade do *ToGatherUp* de reorganizar o armazenamento dos textos deve-se à incorporação, em seu desenvolvimento, de um modelo conceitual de gerenciamento de arquivos chamado *Placeless Documents.* O *Placeless Documents* foi criado por Paul Dourish (2003), pesquisador do *Xerox Palo Alto Research Center*, localizado em Palo Alto, na Califórnia, nos Estados Unidos, e propõe a organização de documentos a partir das suas propriedades, conforme as diferentes necessidades de seus usuários.

Nesse modelo, a associação das propriedades dos documentos (informações sobre os próprios documentos), chamadas de *active properties*, aos documentos permite que eles sejam organizados de acordo com essas propriedades ao invés de obedecerem a uma estrutura hierárquica predeterminada. Ao oferecer essa nova forma de organização baseada em propriedades, o *Placeless Documents* possibilita o agrupamento, de diferentes maneiras, de um conjunto de documentos,

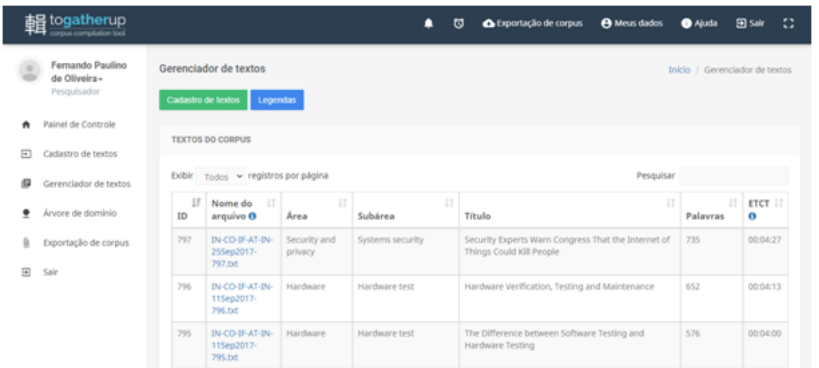
o que soluciona o problema da reorganização de arquivos de acordo com suas funções. A flexibilidade proporcionada pelo *Placeless Documents* foi a principal razão para a sua incorporação ao *ToGatherUp*. No entanto, o modelo implementado no *ToGatherUp* baseou-se na associação dos metadados dos textos do CoCLI ao invés das propriedades dos seus arquivos.

Ao submetermos um texto por meio do formulário do Cadastro de Textos do *ToGatherUp*, seus metadados são registrados no banco de dados do sistema e seu arquivo é armazenado em um diretório comum do servidor *web* em que o sistema está instalado. Por seguir o modelo *Placeless Documents*, o local de armazenamento dos arquivos dentro da infraestrutura do *ToGatherUp* é irrelevante, uma vez que será a necessidade do pesquisador que irá determinar seu posicionamento na estrutura de diretórios, que é gerada no momento da sua exportação para o processamento em outras ferramentas computacionais.

3.2.3 O Gerenciador de Textos

Além das informações quantitativas disponíveis no Painel de Controle, o *ToGatherUp* apresenta uma interface, nomeada como Gerenciador de Textos, que permite ao pesquisador a visualização dos textos de um *corpus*, em forma de tabela, e a pesquisa por um ou mais textos do *corpus* com base em suas informações. A Figura 10 mostra a interface do Gerenciador de Textos.

FIGURA 10 – Gerenciador de Textos do *ToGatherUp*



Fonte: *ToGatherUp*.

A tabela do Gerenciador de Textos possui as colunas: a) ID; b) Nome do arquivo; c) Área; d) Subárea; e) Título; f) Palavras (número de palavras); g) ETCT. O clique sobre o título de cada coluna faz com que suas informações sejam visualizadas em ordem crescente ou decrescente, no caso dos dados numéricos, ou em ordem alfabética, no caso dos dados alfabéticos ou alfanuméricos.

3.2.4 A Exportação de *Corpus*

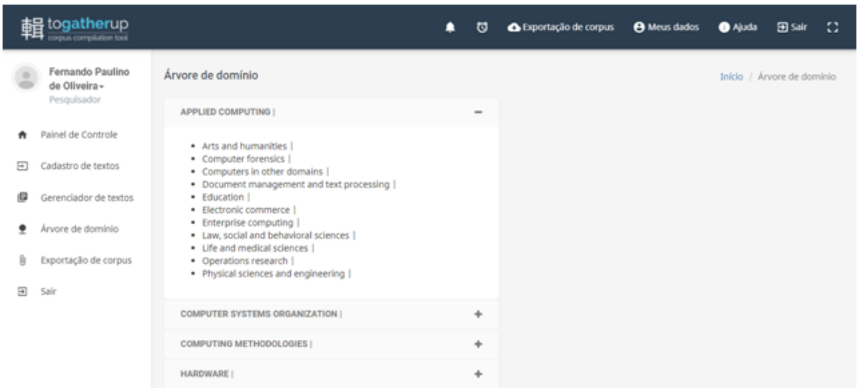
Ao término do projeto de construção de um *corpus*, os seus dados precisam ser disponibilizados para o processamento em ferramentas computacionais. Pensando nisso, o *ToGatherUp* possui o recurso Exportação de *Corpus*, que operacionaliza a exportação dos arquivos do *corpus* de modo que eles possam ser manipulados em outras ferramentas computacionais. Para realizar a exportação, o pesquisador deve utilizar a opção Exportação de *Corpus*, disponível na barra superior e no menu principal do sistema, a qualquer momento que julgar necessário. Ao acioná-la, o *ToGatherUp* cria, de forma automática, um arquivo compactado contendo os textos do *corpus* organizados em diretórios e subdiretórios, conforme a estrutura estabelecida para o projeto, nomeados de forma padronizada e com os seus cabeçalhos já inseridos no início do conteúdo de cada arquivo.

No caso da exportação dos dados do CoCLI, o *ToGatherUp* criou, automaticamente, um arquivo compactado com a extensão .zip, contendo os textos do *corpus* alocados em diretórios correspondentes aos seus campos semânticos na Árvore de Domínios da Computação. No entanto, é importante salientarmos que a flexibilidade oferecida pelo modelo *Placeless documents* e pelo uso dos metadados permite que a exportação seja feita de acordo com outros esquemas. Para exemplificarmos, poderíamos gerar, a partir do conjunto de textos do CoCLI, um *subcorpus* composto somente por textos do gênero científico, caso as configurações de exportação do *ToGatherUp* fossem definidas para esse novo esquema.

3.2.5 A Árvore de Domínio

A Árvore de Domínio é a interface do *ToGatherUp* que exibe a organização hierárquica adotada no projeto de construção de um *corpus*. A Figura 11 exibe a Árvore de Domínio da Computação, com suas áreas e subáreas, adotada no projeto de construção do CoCLI.

FIGURA 11 – Árvore de Domínio do CoCLI



Fonte: *ToGatherUp*.

3.3 A construção do CoCLI

Conforme explicamos anteriormente, a pesquisa aqui retratada comparou os esforços investidos na construção de duas versões idênticas do *Corpus* da Computação da Língua Inglesa (CoCLI), sendo que o projeto de uma delas contou com a incorporação do *ToGatherUp* e o outro não. Ou seja, os projetos foram executados por meio de métodos distintos. Com o intuito de facilitar a compreensão do texto, utilizamos a expressão "Método 1" para referenciar o método que não envolveu a incorporação do *ToGatherUp* e "Método 2" para o que adotou a ferramenta. Apesar da distinção, os dois métodos apresentam um conjunto de atividades em comum e um conjunto de atividades próprias de cada um deles. A seguir, descrevemos a parte comum entre os métodos e, na sequência, tratamos da parte em que eles se distinguem um do outro.

3.3.1 Atividades comuns dos métodos 1 e 2

A parte comum entre os métodos 1 e 2 compreende atividades das fases inicial e de execução dos projetos de construção de *corpora*. A primeira atividade realizada foi a definição do desenho do *corpus*. Após a seleção e definição dos critérios, o desenho do CoCLI apresentou a configuração do Quadro 3.

QUADRO 3 – Desenho do CoCLI

Critério	Definição
Objetivo	Recuperar informações, extrair termos, definir termos e identificar exemplos de uso de termos.
Domínio:[38]	Textos restritos às áreas e subáreas da Computação.
Tipo	Especializado (composto por textos das áreas e subáreas da Computação).
Tempo	Sincrônico (contempla textos publicados no período de 2000 a 2018).
Língua	Monolíngue (apenas textos escritos na língua inglesa).
Gênero e tipo textual	Textos científicos (artigos científicos, capítulos/seções de livro, teses, dissertações, monografias e livros), informativos (artigos, notícias, relatórios e reportagens) e instrucionais ou normativos (apostilas, perguntas e respostas de fóruns, guias, manuais, decretos, normas técnicas, notas técnicas, portarias, tutoriais e documentos).
Tamanho	Cada campo nocional da CSS deverá contar com, no mínimo, 100 mil palavras.[39]
Modalidade	Escrita.
Público-alvo	Pesquisadores, aprendizes e profissionais da Computação.
Estado natural dos textos	Formato eletrônico e sem a necessidade de reconhecimento de seus caracteres.[40]

Fonte: Oliveira (2019, p. 111)

Após o estabelecimento do *design* do CoCLI, foram definidos os recursos financeiros, tecnológicos, materiais e humanos que seriam despendidos para a execução dos projetos e o cronograma das suas realizações. Como todo o trabalho da pesquisa foi realizado pelos seus autores e a hospedagem do *ToGatherUp* foi feita, de forma gratuita, no

[38] Assunto do *corpus*.

[39] Os autores da pesquisa não identificaram, na literatura da LC, um número padrão estabelecido para um *corpus* ou para as ramificações de uma Árvore de Domínio. Por essa razão, estabeleceram o número de 100 mil palavras como padrão para a pesquisa, partindo do pressuposto de que esse valor é suficiente para a recuperação de informações em uma pesquisa terminológica

[40] Essa condição dos textos facilita a captura deles.

servidor *web* do ILEEL[41] da UFU,[42] como parte dos projetos do GPELC,[43] sob o domínio www.togatherup.ileel.ufu.br, não foi necessário investir recursos financeiros para a sua realização. Com essas definições, foi encerrada a fase inicial dos projetos de construção do CoCLI e passou-se à fase de execução deles, na qual iniciou-se o processo de obtenção dos dados dos *corpora*.

A primeira atividade desse processo foi a localização de textos que pudessem ser incluídos nos *corpora*. A escolha dos textos foi feita com base nas referências dos currículos dos cursos da área da Computação de centros acadêmicos de referência. Os textos foram obtidos na Internet e a permissão para o seu uso não foi solicitada devido à dificuldade de obtenção de autorização para a grande quantidade de materiais necessária para os *corpora*. Além desse motivo, a obtenção da permissão de uso foi desconsiderada pela falta de intenção de publicização do CoCLI ao término de sua construção.

Logo após, iniciou-se o processo de preparação dos textos para incluí-los nos *corpora.* Os textos dos *corpora* foram obtidos em suas fontes originais nos formatos PDF, DOC e HTML, convertidos para o formato TXT, limpos por meio da remoção dos elementos citados no Quadro 4 e normalizados por meio dos procedimentos descritos no Quadro 5.

QUADRO 4 – Procedimentos de limpeza de *corpus*

Procedimentos
(a) Remoção de cabeçalhos e rodapés de páginas.
(b) Remoção de elementos gráficos (figuras, imagens e gráficos).
(c) Remoção de imagens.
(d) Remoção de notas de rodapé e fim.[44]
(e) Remoção de números de página.
(f) Remoção de referências bibliográficas.

[41] Instituto de Letras e Linguística.

[42] Universidade Federal de Uberlândia.

[43] Grupo de Pesquisa e Estudos em Linguística de *Corpus.*

[44] Optamos por excluir esses elementos dos textos por julgarmos o restante das informações das produções escritas suficiente para os objetivos da pesquisa

(g) Remoção de listas (sumários, tabelas, figuras, abreviações e gráficos).
(h) Remoção de tabelas e quadros.
(h) Remoção de títulos e subtítulos.
(i) Remoção de legendas de tabelas, figuras e quadros.

Fonte: Oliveira (2019, p. 114.)

QUADRO 5 – Procedimentos de normalização textual

Procedimentos
(a) Remoção de hifens no final de linha.
(b) Remoção de quebras de linhas/parágrafos/páginas/seções.
(c) Remoção de espaços em branco duplicados.
(d) Remoção de marcas de parágrafos/recuos.
(e) Remoção de linhas em branco.
(f) Padronização de hifens, apóstrofos, traços e aspas.

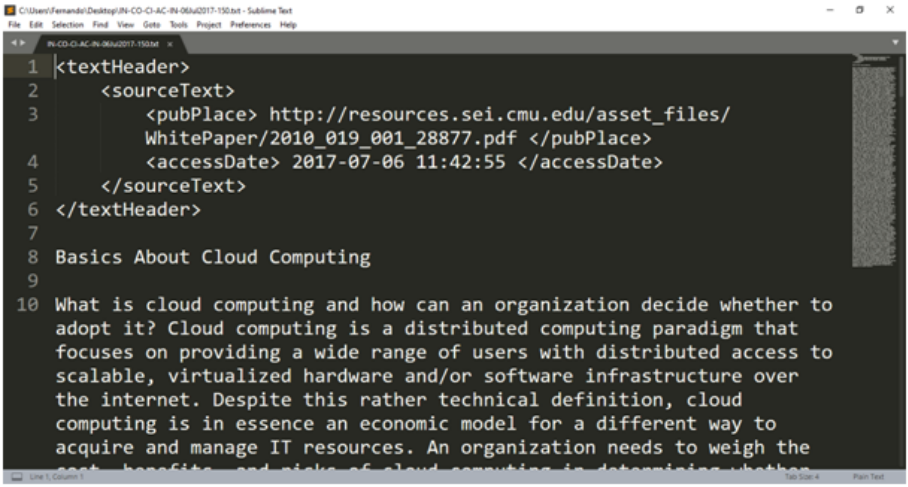
Fonte: Oliveira (2019, p. 114.)

Após a conclusão da conversão, limpeza e normalização dos textos, iniciou-se a execução da última atividade do processo de preparação dos *corpora* – o enriquecimento dos dados.

3.3.2 Atividades distintas dos métodos 1 e 2

As atividades de enriquecimento e relativas ao armazenamento dos dados compõem o conjunto de atividades que tiveram a forma de execução completamente alterada com a incorporação do *ToGatherUp*. O enriquecimento dos dados dos *corpora* da pesquisa consistiu na inserção de cabeçalhos construídos a partir de metadados dos textos. A Figura 12 ilustra um exemplo de cabeçalho inserido pelo *ToGatherUp* em um dos textos do CoCLI.

FIGURA 12 – Cabeçalho do texto *Basics about Cloud Computing*



Fonte: *ToGatherUp*.

No Método 1, o cabeçalho da Figura 12 foi construído e o inserido no arquivo do texto de forma manual. Já no Método 2, o *ToGatherUp* foi programado para construir e inserir, automaticamente, o cabeçalho no texto, de acordo com a estrutura XML definida em sua programação e os metadados do texto. A inclusão dos cabeçalhos encerrou o processo de preparação dos textos, que foi sucedido pelas atividades de armazenamento dos dados dos *corpora*. Para a nomeação dos arquivos dos textos, foi utilizada a convenção de nomeação de arquivos apresentada no subtópico Cadastro de Textos. No Método 1, os textos dos *corpora* foram nomeados manualmente e, no Método 2, todo o trabalho foi executado de maneira automática pelo *ToGatherUp* durante o registro do texto no Cadastro de Textos da ferramenta. Para que isso fosse possível, o *ToGatherUp* foi programado para nomear os arquivos de acordo com as regras da convenção de nomeação de arquivos adotada no projeto e com os metadados dos textos.

A última atividade dos projetos de construção do CoCLI foi o salvamento (arquivamento) dos textos. No Método 1, os arquivos dos *corpora* foram salvos manualmente nos diretórios, criados com o *Windows Explorer* do *Windows*, correspondentes às áreas e subáreas presentes na Árvore de Domínio da Computação. No Método 2, o *ToGatherUp* armazenou automaticamente os arquivos em consonância com os princípios

e a funcionalidade de armazenamento da ferramenta apresentados no item d) Atividade 4: Armazenamento do arquivo do texto do subtópico 3.2.2 Cadastro de Textos. No contexto do *ToGatherUp*, o local dos arquivos é indiferente, uma vez que a ferramenta é capaz de organizar os arquivos de acordo com a consulta estabelecida pelo usuário do sistema. No caso do CoCLI, o *ToGatherUp* foi programado para exportar os arquivos conforme a estrutura da Árvore de Domínio da Computação.

3.4 O experimento

O experimento realizado na pesquisa consistiu na realização de um teste estatístico que comparou os ETP de cada um dos projetos de construção do CoCLI. O objetivo do experimento foi testar a hipótese de que a incorporação do *ToGatherUp* em projetos de construção manual de *corpora* poupa o tempo e minimiza o esforço do pesquisador dispensados à execução das atividades de elaboração de *corpora*, de modo semelhante ao que ocorre com as atividades de análise de *corpora* mediadas pelo uso de computadores (criação automática de listas de palavras e linhas de concordância, evidenciação de padrões linguísticos e etiquetagem de *corpora*). Para a realização do experimento foi necessário realizar a tabulação manual dos EAs, fornecidos pelo cronômetro do *ToGatherUp* (Instrumento 1), em uma planilha do *Google* (Instrumento 2), para cada uma das atividades de cada um dos projetos de construção do CoCLI. Desses conjuntos de dados (*dataset*) foram extraídas amostras aleatórias referentes aos mesmos 50 textos de cada *corpus*. Os dados das amostras foram submetidos a um teste estatístico que permitiu determinar o efeito da incorporação do *ToGatherUp* na construção manual das duas versões do CoCLI.

De acordo com Rumsey (2010), testar uma hipótese é uma tentativa de se "confirmar ou negar uma declaração sobre uma população[45] a partir dos dados de sua amostra"[46] (RUMSEY, 2010, p. 87).[47] Para

[45] Para Correia (2003), população é "uma coleção completa de todos os elementos a serem estudados" (CORREIA, 2003, p. 9).

[46] Consoante Correia (2003), uma subcoleção de elementos extraídos de uma população" (CORREIA, 2003, p. 9).amostra é "uma subcoleção de elementos extraídos de uma população" (CORREIA, 2003, p. 9).

[47] Original: "trying to confirm or deny a claim about a population using data from a sample".

a autora, quando um teste de hipóteses[48] envolve a comparação entre parâmetros numéricos, o objeto de interesse é a diferença entre as médias[49] (*means*) desses parâmetros. Como a análise envolveu a comparação entre o ETP dos diferentes projetos de construção do CoCLI (dois parâmetros numéricos), foi utilizado um teste de hipóteses conhecido como *T-Test* que, segundo Dodge (2008), é apropriado para testar hipóteses a partir da comparação entre as médias de duas populações em que os elementos de uma delas possuem uma relação com os elementos da outra.

Para a realização do T-Test, os dados tabulados foram importados no *software Statistics Statistical Package for the Social Sciences* (SSPS),[50] uma ferramenta de análise estatísticas, desenvolvida pela IBM, amplamente usada em pesquisas acadêmicas que envolvem a realização de testes estatísticos. Na sequência, foi utilizada uma função do SSPS para criar uma amostra aleatória[51] de cinquenta textos do CoCLI. Por fim, o SSPS comparou o ETCT resultante da aplicação do Método 1 (que abreviamos como ETCT – Método 1) e o ETCT resultante da aplicação do Método 2 (que passamos a chamar de ETCT – Método 2) dos 50 textos selecionados. Os dados referentes ao ETCT – Método 1 constituíram o Grupo de Controle[52] (*control group*) do experimento e os dados relativos ao ETCT – Método 2 formaram o Grupo Experimental (*treatment group*). O tratamento que diferenciou o Grupo de Controle do Grupo Experimental foi a manipulação dos EAs automatizados pelo *ToGatherUp* no Método 2.

---

[48] Segundo Correia (2003), um teste de hipóteses é "técnica para se fazer inferência estatística. Ou seja, a partir de um teste de hipóteses realizado com os dados amostrais, pode-se fazer inferências sobre a população" (CORREIA, 2003, p. 100).

[49] Segundo Correia (2003), um teste de hipóteses é "técnica para se fazer inferência estatística. Ou seja, a partir de um teste de hipóteses realizado com os dados amostrais, pode-se fazer inferências sobre a população" (CORREIA, 2003, p. 100).

[50] O SSPS foi escolhido por realizar os cálculos estatísticos de forma automática. Disponível em: https://www.ibm.com/br-pt/products/spss-statistics. Acesso em: 23 fev. 2019.

[51] Os registros que compuseram o conjunto de dados analisados foram criados de forma automática e aleatória pelo SSPS.

[52] De acordo com Rumsey (2010), as amostras que são expostas a condições normais (não recebem tratamento ou recebem um tratamento falso, também chamado de placebo) denominam-se Grupo de Controle. Já as amostras sujeitas a tratamento que afeta seus atributos são chamadas de Grupo Experimental.

A análise dos resultados fornecidos pelo SSPS considerou os conceitos estatísticos de hipótese nula[53] (*null hypothesis*) e hipótese alternativa (*alternate hypothesis*). Segundo Charles Brase e Corrine Brase (2011, p. 411), a hipótese nula ou "hipótese estatística"[54] é a declaração que está sob teste e, geralmente, associa-se a resultados como "não houve efeito", "não houve diferença" ou "nada foi alterado" entre a média calculada para o Grupo de Controle e a média calculada para o Grupo Experimental. A hipótese alternativa[55] é definida pelos autores como qualquer declaração diferente da hipótese nula. De acordo com os conceitos de hipótese nula e alternativa, a nossa hipótese da pesquisa pode ser feita da seguinte maneira: a hipótese deve ser rejeitada caso o resultado do *T-Test* revele que o ETCT do método que utiliza o *ToGatherUp* é igual ou maior do que o ETCT do método que não utiliza a ferramenta. Se a hipótese nula for rejeitada, ou seja, se o *T-Test* mostrar que o ETCT do método que utiliza o *ToGatherUp* é menor do que o ETCT do método que não utiliza a ferramenta, a hipótese alternativa deve ser aceita e a hipótese confirmada.

O resultado de um teste de hipótese é estatisticamente significativo quando a probabilidade de que ele tenha ocorrido por acaso seja muito improvável. Por essa razão, os autores da pesquisa preocuparam-se em determinar o nível de significância que foi considerado no teste. Para Rumsey (2010), o nível de significância de um teste de hipótese, também conhecido como *alpha level* (α), é dado pelo *p-value* (*probability value*) que, geralmente, é definido em 0.05[56] ou 0.01. Caso o *p-value* é maior ou igual a α, a hipótese nula deve ser aceita e, caso o *p-value* é menor que α, a hipótese nula deve ser rejeitada. Em outras palavras, o resultado de um teste de hipótese é estatisticamente significativo quando, a partir do seu

[53] Na estatística, a hipótese nula é representada por H0 e a hipótese alternativa, por H1.

[54] Para Correia (2003), a hipótese estatística "trata-se de [i.e. trata de] uma suposição quanto ao valor de um parâmetro populacional, ou quanto à natureza da distribuição de probabilidade de uma variável populacional" (CORREIA, 2003, p. 100).

[55] Autores como Rumsey (2010) também usam a expressão "hipótese de pesquisa" para referenciar a hipótese alternativa.

[56] De acordo com Rumsey (2010), um *p-value* de 0.05 e um *p-value* de 0.01 indicam, respectivamente, que em 95% e 99% das vezes os resultados da amostra poderão se repetir caso o experimento seja realizado novamente com outras amostras aleatórias da mesma população sob as mesmas condições. Para Rumsey (2010), outros valores podem ser assumidos para o *p-value* e essa determinação depende de cada pesquisador.

p-value, é possível rejeitar a hipótese nula devido à improbabilidade de que ela ocorra. A rejeição da hipótese nula, consequentemente, leva-nos a acreditar que a hipótese alternativa pode ser verdadeira. Sendo assim, os autores definiram o *p-value* a ser considerado no teste em 0.05 por julgarem esse nível de significância bastante aceitável para o propósito da pesquisa.

4 Resultados

A execução do *T-Test* no SSPS gerou o resultado apresentado na Tabela 1.

TABELA 1 – Resultado do *T-Test*

		Paired Differences					*t*	*df*	*Sig. (2-tailed)*
	Mean	*Std. Deviation*	*Std. Error Mean*	*95% Confidence Interval of the Difference*					
				Lower	*Upper*				
Pair 1	ETCT 1[57] & ETCT 2[58]	131,140	4,333	613	129,909	132,371	214,003	49	0,000

Fonte: Oliveira (2019, p. 141.)

O ETCT da amostra construída com Método 1, indicado na coluna *Mean* da Tabela 1, foi, em média, 131 segundos maior do que o ETCT da amostra elaborada com o Método 2. Esse valor pode ser considerado estatisticamente significativo, uma vez que o *p-value* do teste foi igual a 0,000, conforme indica a coluna *Sig. (2-tailed) da* Tabela 4, um valor bem inferior ao *p-value* (0.05) estabelecido para a garantia da significância estatística do *T-Test*. Portanto, com base no resultado do *T-Test*, a hipótese nula da pesquisa (Hipótese nula (H_0): ETCT – Método 2 = ou > ETCT – Método 1) pode ser rejeitada pelos pesquisadores e eles puderam afirmar, por inferência, que os resultados encontrados sugerem que a incorporação do *ToGatherUp* reduz o ETP de construção manual de *corpora*.

[57] Referente ao Método 1.
[58] Referente ao Método 2.

6 Considerações finais

A pesquisa retratada neste artigo é o resultado de um trabalho sistemático para a determinação do efeito da incorporação do *ToGatherUp* em projetos de construção manual de *corpora* e, até onde pudemos verificar por meio da revisão bibliográfica da LC, consiste em um dos primeiros trabalhos a propor uma forma de mensurar o esforço necessário para a realização de projetos de elaboração manual de *corpora* e a propor uma sistematização do trabalho de criação manual de *corpora*, respeitando princípios e métodos da LC e da área de Gerenciamento de Projetos.

O uso do *ToGatherUp* que, no momento da redação deste artigo, está passando por ajustes para que possa ser disponibilizado em 2021 e utilizado, gratuitamente, em outras pesquisas é outro ponto de destaque da pesquisa. Acreditamos que a disponibilização da ferramenta irá contribuir para o preenchimento da lacuna[59] existente na LC referente à carência de ferramentas voltadas para o suporte das atividades de construção manual de *corpora* compostos por grande volume de dados.

Além dessas contribuições, a pesquisa traz uma importante discussão sobre possíveis complicações do uso de *web corpora* nas pesquisas em que existe a preocupação quanto à precisão de análises, visto que as ferramentas de coleta automática de textos, no estágio atual da tecnologia, não conseguem lidar com os problemas apontados na fundamentação teórica deste artigo. Essa discussão ganha mais relevância ao considerarmos o fato identificado na pesquisa de que o percentual do EALND, em ambos os métodos de construção do CoCLI, foi maior do que todos os demais esforços somados juntos, atingindo 83,29% no Método 1 e 90,02% no Método 2, corroborando a ideia de Dasu e Johnson (2003) de que a limpeza e a normalização podem ocupar cerca de 80% do tempo compreendido entre a obtenção de um texto e sua análise. Se o maior esforço de um projeto de construção de *corpora* está nas atividades de limpeza e normalização e as ferramentas de coleta automática de textos negligenciam essas atividades, as análises feitas a partir de *corpora* coletados automaticamente correm o risco de serem postas em xeque.

[59] A referida lacuna foi identificada por meio de um levantamento realizado durante a pesquisa em que foram analisadas dez ferramentas da LC apontadas para a criação de *corpora* pelo projeto *Corpus Analysis* (KLEIBER; BERBERICH, 2018), desenvolvido por Ingo Kleiber e Kristin Berberich, da Universidade de Heidelberg, na Alemanha.

Referências

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, São Leopoldo, v. 4, n. 3, p. 156-178, 2006. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em: 2 abr. 2019.

ALVES, R. C. V. *Metadados como elementos do processo de catalogação.* 2010. 132f. Tese (Doutorado em Ciência da Informação) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, 2010. Disponível em: https://repositorio.unesp.br/handle/11449/103361. Acesso em: 2 abr. 2019.

ANTHONY, L. *EncodeAnt*. Version 1.2.0. [Computer Software]. Tokyo: Waseda University, 2016. Disponível em: http://www.laurenceanthony.net. Acesso em: 2 abr. 2019.

ATKINS, S.; CLEAR, J.; OSTLER, N. Corpus design criteria. *Literary and Linguistic Computing,* Oxford, v. 7, n. 1, p. 1-16, 1992. DOI: https://doi.org/10.1093/llc/7.1.1. Disponível em: https://academic.oup.com/dsh/article-abstract/7/1/1/1028498?redirectedFrom=fulltext. Acesso em: 17 abr. 2019.

BAKER, P. Corpus Methods in Linguistics. *In*: LITOSSELITI, L. (ed.). *Research Methods in Linguistics.* New York: Continnum International Publishing Group, 2010. p. 93-113.

BARONI, M.; BERNARDINI, S. *BootCaT*. Version 1.08. [Computer Software]. Trento/Forlì: Universities of Bologna, 2004. Disponível em: http://bootcat.dipintra.it. Acesso em: 2 abr. 2019.

BARONI, M. *et al*. WebBootCaT: A Web Tool for Instant Corpora. *In*: EURALEX INTERNATIONAL CONGRESS, 12th., 2006, Torino. *Proceedings* […]. Torino: Edizioni dell'Orso s.r.l., 2006. p. 123-131. Disponível em: https://euralex.org/publications/webbootcat-a-web-tool-for-instant-corpora/. Acesso em: 2 abr. 2019.

BARREAU, D.; NARDI, B. Finding and Reminding: File Organization from the Desktop. *ACM SIGCHI Bulletin,* New York, v. 27, n. 3, p. 39-43, 1995. DOI: https://doi.org/10.1145/221296.221307. Disponível em: https://dl.acm.org/citation.cfm?id=221307. Acesso em: 17 abr. 2019.

BERBER SARDINHA, T. A influência do tamanho do corpus de referência da obtenção de palavras-chave usando o Programa Computacional Wordsmith Tools. *The ESPecialist*, São Paulo, v. 26, n. 2, p. 188, 2005. Disponível em: https://revistas.pucsp.br/esp/article/view/9290. Acesso em: 27 nov. 2020.

BERBER SARDINHA, T. *Linguística de Corpus*. São Paulo: Manole, 2004.

BERGH, G.; ZANCHETTA, E. Web linguistics. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics:* An International Handbook. Berlin: Mouton de Gruyter, 2008. p. 309-327.

BIANCHI, F. *Culture, Corpora and Semantics*: Methodological Issues in Using Elicited and Corpus Data for Cultural Comparison. Lecce: ESE Salento University Publishing, 2012. Disponível em: http://siba-ese.unisalento.it/index.php/culturecorpora/article/viewFile/12427/11066. Acesso em: 10 jan. 2019.

BIBER, D. Representativeness in Corpus Design. *Literary and Linguistic Computing,* Oxford, v. 8, n. 4, p. 223-257, 1993. DOI: https://doi.org/10.1093/llc/8.4.243. Disponível em: http://otipl.philol.msu.ru/media/biber930.pdf. Acesso em: 2 abr. 2019.

BLECHA, J. *Building Specialized Corpora.* 2012. 159f. Thesis (Master in English Language and Literature) – Faculty of Arts, Department of English and American Studies, Masaryk University, Brno, República Tcheca, 2012. Disponível em: https://is.muni.cz/th/aki90/179991_Building_Specialized_Corpora.pdf. Acesso em: 2 abr. 2019.

BRASE, C. H.; BRASE, C. P. *Understandable Statistics*: Concepts and Methods, 10. ed. Boston: Cengage Learning, 2011.

CORREIA, M. S. B. B. *Probabilidade e estatística*. 2. ed. Belo Horizonte: PUC Minas Virtual, 2003. Disponível em: http://estpoli.pbworks.com/f/livro_probabilidade_estatistica_2a_ed.pdf. Acesso em: 25 fev. 2019.

DASU, T.; JOHNSON, T. *Exploratory Data Mining and Data Cleaning*. Hoboken: John Wiley & Sons, 2003. DOI: https://doi.org/10.1002/0471448354.

DODGE, Y. *The Concise Encyclopedia of Statistics*. New York: Springer-Verlag, 2008.

DOURISH, P. The Appropriation of Interactive Technologies: Some Lessons from Placeless Documents. *Computer Supported Cooperative Work (CSCW),* Dordrecht, v. 12, n. 4, p. 465-490, 2003. DOI: https://doi.org/10.1023/A:1026149119426. Disponível em: https://link.springer.com/article/10.1023/A:1026149119426. Acesso em: 17 abr. 2019.

EDWARD, R. P. Computational Tools and Methods for Corpus Compilation and Analysis. *In*: BIBER, D; REPPEN, R. (ed.). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015. p. 32-49.

ESCARTÍN, C. P. Design and compilation of a specialized Spanish-German parallel corpus. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE (LREC), 2012, Istanbul. *Proceedings* […]. Istanbul: European Language Resources Association (ELRA), 2012. p. 2199-2206. Disponível em: http://www.lrec-conf.org/proceedings/lrec2012/pdf/577_Paper.pdf. Acesso em: 2 abr. 2019.

FENTON, N.; BIEMAN, J. *Software Metrics*: A Rigorous and Practical Approach. 3. ed. Boca Raton: CRC Press, 2014. DOI: https://doi.org/10.1201/b17461.

FRANKENBERG-GARCIA, A. Prefácio. *In*: SHEPHERD, T. M. G.; BERBER SARDINHA, T.; PINTO, M. V. (org.). *Caminhos da linguística de corpus*. São Paulo: Mercado de Letras, 2012. p. 11-14.

FROMM, G. O uso de *corpora* na análise linguística. *Revista Factus,* São Paulo, v. 1, n. 1, p. 69-76, 2003. Disponível em: http://www.ileel.ufu.br/guifromm/upload/ousodecorporanaproducaolinguistica.pdf. Acesso em: 17 abr. 2019.

FROMM, G. *VoTec*: a construção de vocabulários eletrônicos para aprendizes de tradução. 2007. 214 f. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2007. Disponível em: http://www.teses.usp.br/teses/disponiveis/8/8147/tde-08072008-150855/pt-br.php. Acesso em: 2 abr. 2019.

GARRETSON, G. Desiderata for Linguistic Software Design. *Internatinal Journal of English Studies (IJES),* Murcia, v. 8, n. 1, 67-94, 2008. Disponível em: http://revistas.um.es/ijes/article/view/49101. Acesso em: 2 abr. 2019.

GARSIDE, R.; SMITH, N. A Hybrid Grammatical Tagger: CLAWS 4. *In*: GARSIDE, R.; LEECH, G.; MCENERY, T. (eds.). *Corpus annotation*: Linguistic Information from Computer Text Corpora. London: Routledge; Taylor & Francis, 1997. p. 102-121. DOI: https://doi.org/10.4324/9781315841366

GOOGLE. *Refinar pesquisas na Web*, 2019. Disponível em: https://support.google.com/websearch/answer/2466433?hl=pt-BR. Acesso em: 1 abr. 2019.

GRIES, S. T. What is Corpus Linguistics? *Language and Linguistics Compass*, Hoboken, v. 3, n. 5, p. 1225-1241, 2009. DOI: https://doi.org/10.1111/j.1749-818X.2009.00149.x. Disponível em: https://onlinelibrary.wiley.com/doi/full/10.1111/j.1749-818X.2009.00149.x. Acesso em: 2 abr. 2019.

KAPTELININ, V. Creating Computer-Based Work Environments: An Empirical Study of Macintosh Users. *In*: ACM SIGCPR/SIGMIS CONFERENCE ON COMPUTER PERSONNEL RESEARC, 1996, Denver. *Proceedings* […]. Denver: ACM, 1996. p. 360-366. DOI: https://doi.org/10.1145/238857.238921. Disponível em: https://dl.acm.org/citation.cfm?id=238921. Acesso em: 17 abr. 2019.

KEHOE, A.; GEE, M. New Corpora from the Web: Making Web Text More 'Text-Like'. *Studies in Variation, Contacts and Change in English,* Helsinki, v. 2, [s.p.], 2007. Disponível em: http://www.helsinki.fi/varieng/series/volumes/02/kehoe_gee/. Acesso em: 18 jan. 2019.

KENNEDY, G. *An Introduction to Corpus Linguistics*. New York: Longman, 1998.

KLEIBER, I.; BERBERICH, K. *Corpus Analysis*. Heidelberg: Heidelberg University, 2018. Disponível em: https://corpus-analysis.com/. Acesso em: 25 jan. 2019.

KÜBLER, N.; ASTON, G. Using Corpora in Translation. *In*: O'KEEFFE, A.; MCCARTHY, M. J. (org.). *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 2010. p. 501-515. DOI: https://doi.org/10.4324/9780203856949-36

LEECH, G. Adding Linguistic Annotation. *In*: WYNNE, M. (ed.). *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford:

Oxbow Books, 2005. p. 17-29. Disponível em: http://ota.ox.ac.uk/documents/creating/dlc/. Acesso em: 2 abr. 2019.

MACMILLAN DICTIONARY. *Gather Up*. 2018. Disponível em: https://www.macmillandictionary.com/dictionary/british/gather-up. Acesso em: 21 jun. 2018.

MACMULLEN, W. J. Requirements Definition and Design Criteria for Test Corpora in Information Science. *SILS Technical Report 2003-03*. School of Information and Library Science: University of North Carolina at Chapel Hill. p. 3-21, 2003. Disponível em: https://sils.unc.edu/sites/default/files/general/research/TR-2003-03.pdf. Acesso em: 10 jan. 2019.

MARTINET, A. *Elementos de linguística geral.* 8. ed. Lisboa: Martins Fontes, 1978.

MCENERY, T.; HARDIE, A. *Corpus Linguistics*: Method, Theory and Practice. Cambridge: Cambridge University Press, 2011. DOI: https://doi.org/10.1017/CBO9780511981395.

MCENERY, T.; XIAO, R.; TONO, Y. *Corpus-Based Language Studies*: An Advanced Resource Book. London; New York: Routledge, 2006. Disponível em: https://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xCBLS/chapters/A10.pdf. Acesso em: 10 jan. 2019.

MEYER, C. F. *English Corpus Linguistics*: An Introduction. Cambridge: Cambridge University Press, 2004.

MINSHALL, D. E. *A Computer Science Word List.* 2013. 98f. Dissertation (Master of Arts - MA TEFL) – University of Swansea, Swansea, UK, 2013. Disponível em: https://www.baleap.org/wp-content/uploads/2016/03/Daniel-Minshall.pdf. Acesso em: 10 jan. 2019.

NELSON, M. Building a Written Corpus: What Are the Basics? *In*: O'KEEFFE, A.; MCCARTHY, M. J. (org.). *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 2010. p. 53-65. DOI: https://doi.org/10.4324/9780203856949-5

NEUMANN, S.; HANSEN-SCHIRRA, S. Corpus Methodology and Design. *In*: HANSEN-SCHIRRA, S.; NEUMANN, S.; STEINER, E. (org.). *Cross-Linguistic Corpora for the Study of Translations:* Insights from the Language Pair English-German. Berlin: De Gruyter Mouton, 2012. p. 21-34. DOI: https://doi.org/10.1515/9783110260328.

OLIVEIRA, F. P. *ToGatherUp:* um protótipo de ferramenta para a construção de *corpora* a produção de vocabulários bilíngues direcionada por corpus. 2019. 219f. Dissertação (Mestrado em Estudos Linguísticos) – Instituto de Letras e Linguística, Universidade Federal de Uberlândia, 2019. Disponível em: https://repositorio.ufu.br/bitstream/123456789/25433/1/ToGatherUpProtótipoFerramenta. Acesso em: 2 abr. 2019.

PROJECT MANAGEMENT INSTITUTE. *Um Guia do Conhecimento em Gerenciamento de Projetos (Guia PMBOK)*. 5. ed. Newtown Square: Project Management Institute, 2013.

RENOUF, A. Corpus Development 25 Years on: From Super-Corpus to Cybercorpus. *Language and Computers: Studies in Practical Linguistics*, [*S.l.*], v. 62, n. 1, p. 27-49, 2007. DOI: https://doi.org/10.1163/9789401204347_004.

RUBI, M. P. Os princípios da política de indexação na análise de assunto para catalogação: especificidade, exaustividade, revocação e precisão na perspectiva dos catalogadores e usuários. *In*: FUJITA, M. S. L. *et al*. (org.). *A indexação de livros:* a percepção de catalogadores e usuários de bibliotecas universitárias: um estudo de observação do contexto sociocognitivo com protocolos verbais. São Paulo: Cultura Acadêmica, 2009. p. 81-93.

RUMSEY, D. *Statistics Essentials for Dummies*. Hoboken: John Wiley & Sons, 2010.

RUNDELL, M.; KILGARRIFF, A. Automating the Creation of Dictionaries: Where Will It All End? *In*: MEUNIER, F. *et al*. (ed.). *A Taste for Corpora:* A Tribute to Professor Sylviane Granger. Amsterdam: Benjamins, 2011. p. 257-281. DOI: https://doi.org/10.1075/scl.45.15run.

SANTOS, A. *Contributions for Building a Corpora-Flow System.* 2011. 100f. Dissertação (Master in Informatics Engineering) – Escola de Engenharia, Universidade do Minho, Guimarães, PT, 2011. Disponível em: https://repositorium.sdum.uminho.pt/bitstream/1822/28122/1/eeum_di_dissertacao_pg15973.pdf. Acesso em: 17 abr. 2019.

SCHÄFER, R.; BILDHAUER, F. *Web Corpus Construction*. Toronto: University of Toronto, 2013. DOI: https://doi.org/10.2200/S00508ED1V01Y201305HLT022.

SEDLAR, E. *Database-Managed File System*. US Pat. US20050091287A1. Redwood Shores, CA: Oracle International Corporation, 2005.

SEMINO, E.; SHORT, M. *Corpus Stylistics*: Speech, Writing and Thought Presentation in a Corpus of English Writing. London: Routledge, 2004. DOI: https://doi.org/10.4324/9780203494073.

SIMSKE, S. J. *Systems and Methods for Processing Text-Based Electronic Documents*. U.S. Patent n. 7,106, 905. [*S.l.*]: Hewlett-Packard Development Company, 2006.

SINCLAIR, J. McH. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J. McH. Corpus and Text – Basic Principles. *In*: WYNNE, M. (ed.). *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford: Oxbow Books, 2005. [s.p.]. Disponível em: https://ota.ox.ac.uk/documents/creating/dlc/chapter1.htm. Acesso em: 2 abr. 2019.

TAGNIN, S. E. O. Glossário de linguística de *corpus*. *In*: VIANA, V.; TAGNIN, S. E. O. (org.). *Corpora no ensino de línguas estrangeiras.* São Paulo: HUB Editorial, 2010. p. 349-353.

TAGNIN, S. E. O. *Corpora na tradução*. São Paulo: Hub Editorial, 2015.

VOORMANN, H.; GUT, U. Agile Corpus Creation. *Corpus Linguistics and Linguistic Theory,* Berlin, v. 4, n. 2, p. 235-251, 2008. DOI: https://doi.org/10.1515/CLLT.2008.010.

WIDDOWSON, H.G. *Linguistics*. Oxford: Oxford University Press, 1996.

ZANETTIN, F. *Translation-driven corpora*: Corpus resources for descriptive and applied translation studies. London: Routledge, 2014. DOI: https://doi.org/10.4324/9781315759661.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 959-998, 2021

Inteligibilidade e convencionalidade em textos de divulgação da área médica em português brasileiro

Readability and conventionality in expository texts in Brazilian Portuguese

Yuli Souza Carvalho
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul / Brasil
yuli_@live.ca
https://orcid.org/0000-0002-7914-8459

Rozane Rodrigues Rebechi
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul / Brasil
rozane.rebechi@ufrgs.br
https://orcid.org/0000-0002-1878-7548

Resumo: O objetivo desta pesquisa é cotejar dados indicativos de inteligibilidade e convencionalidade em textos de divulgação da área médica em português para verificar sua adequação ao público brasileiro. Para tanto, apoiamo-nos nos pressupostos da Linguística de *Corpus* para a compilação e o processamento de um *corpus* paralelo, formado por textos escritos originalmente em inglês e suas traduções para o português, e um *corpus* comparável, composto pelos textos traduzidos em português e por textos originalmente escritos nesse idioma. A metodologia do estudo combina análises quantitativas – para levantamento de inteligibilidade, chavicidade e colocação – e qualitativas – para análise das palavras em contexto. Em relação à inteligibilidade, as ferramentas apontaram que os textos escritos em português são 'difíceis' para o leitor médio brasileiro, com grau de instrução inferior ao Ensino Médio. Já os textos traduzidos foram considerados 'razoavelmente difíceis', de acordo com esse mesmo critério de avaliação, que classificou os originais em inglês como 'razoavelmente fáceis', considerando-se seu público alvo, ou seja, o leitor médio estadunidense. A análise qualitativa apontou que os textos traduzidos apresentam quebras de convencionalidade, demonstrando preferência por equivalentes *prima facie*, nem sempre condizentes com os

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.959-998

padrões observados nos textos de mesmo gênero escritos originalmente em português. Apesar de a ferramenta de acessibilidade textual indicar que tanto os textos escritos originalmente em português quanto aqueles traduzidos não se mostram totalmente adequados para o leitor-alvo brasileiro de textos de divulgação médica, acreditamos que a quebra da convencionalidade, identificada nos textos traduzidos, pode dificultar ainda mais a compreensão do leitor médio de resultados de pesquisas científicas da área da saúde.

Palavras-chave: textos de divulgação; tradução; convencionalidade; inteligibilidade.

Abstract: The aim of this research is to collate data from intelligibility and conventionality in health-related expository texts in Portuguese to investigate their appropriateness to Brazilians. To this end, we rely on Corpus Linguistics for the compilation and processing of a parallel corpus, comprising texts originally written in English and their translations into Portuguese, and a comparable corpus, composed of texts translated into Portuguese and texts originally written in that language. Our methodology combines quantitative analysis – to assess readability, keyness, and collocation – and qualitative analysis – to investigate words in context. Regarding readability, the tools pointed out that texts written in Portuguese are 'difficult' for the average Brazilian reader, with a level of education lower than High School. The translated texts were considered 'fairly difficult', according to this same evaluation criterion, which classified the originals in English as 'fairly easy', considering its target audience, that is, the average American reader. The qualitative analysis pointed out that the translated texts may compromise conventionality, revealing a preference for *prima facie* equivalents, not always consistent with the patterns observed in original Brazilian Portuguese counterparts. Although the accessibility evaluation tool indicates that both the texts originally written in Portuguese and those translated into Portuguese do not prove to be entirely suitable for the Brazilian target reader of medical expository texts, we believe that, by breaking conventionality, the translated texts may hinder even more the average reader's comprehension of results of scientific research.

Keywords: expository texts; translation; conventionality; readability.

Recebido em 09 de outubro de 2020

Aceito em 23 de novembro de 2020

1 Introdução

Os textos de divulgação são de suma importância, já que visam ao compartilhamento, com o público geral, de resultados de pesquisas desenvolvidas por especialistas em diversas áreas do conhecimento.

Na área médica, eles desempenham papel ainda mais importante, pois instruem a população em relação a questões de saúde. O papel preponderante dos textos de divulgação fica ainda mais evidente em momentos como o que estamos vivendo, quando o mundo todo enfrenta uma questão de saúde pública sem precedentes: a pandemia da COVID-19. Empenhados em pesquisar sobre o controle, a transmissão e o tratamento da doença, ao mesmo tempo em que buscam desenvolver uma vacina que consiga combater o vírus transmissor, médicos e cientistas têm publicado uma infinidade de artigos acadêmicos que relatam suas descobertas.

Em geral, as publicações acadêmicas são escritas para a comunidade discursiva que possui conhecimento prévio do domínio especializado no âmbito em que a pesquisa se insere, mas não se mostram adequadas para o público leigo, que espera receber informações menos técnicas para aplicá-las no dia a dia. Sendo assim, é necessário que os dados apresentados nos textos de divulgação sejam fornecidos de forma clara e acessível. Afinal, assim como elevadores e rampas visam tornar as construções acessíveis ao público geral, os textos de divulgação devem ter características que possibilitem sua acessibilidade para leitores de diferentes faixas etárias, classes sociais e níveis de escolaridade. Textos muito complexos para grande parcela do público podem acabar não atingindo o objetivo final, qual seja, o de instruir a população geral acerca de temas especializados.

Atualmente, grande parte das publicações científicas é compartilhada em língua inglesa, a *lingua franca* da academia (JENKINS, 2009), e traduzida para os inúmeros vernáculos falados no mundo. De acordo com Rosselli (2016), cerca de 96% de todos os artigos indexados em 2015 na base de dados especializada em artigos biomédicos PubMed foram publicados em inglês. Assim, a tradução tem papel central no compartilhamento de informações. Ao considerarmos que os textos de divulgação devem atingir uma grande parcela da população, é necessário também investigar se o processo de tradução desses materiais resulta em informações acessíveis ao leitor, especialmente no que diz respeito a aspectos como inteligibilidade e convencionalidade.

Muito já se discutiu sobre 'fidelidade' da tradução, conceito superado – especialmente no que tange aos textos especializados – e que deu lugar à 'lealdade' (NORD, 2006), conceito que coloca em foco a função do texto traduzido para o público-alvo a que se destina. Com

isso em mente, podemos afirmar que um texto de divulgação, escrito em inglês para determinado público-alvo, só vai 'funcionar' bem para o leitor da tradução se estiver adequado às características leitoras desse público.

Assim, o objetivo da pesquisa[1] é investigar, a partir de dados levantados por ferramentas de análise textual, como se dá a relação entre inteligibilidade e convencionalidade em textos de divulgação da área médica, a fim de verificar sua adequação para o público brasileiro, quer tenham esses textos sido escritos originalmente em português, quer sejam traduções. Para tanto, foram compilados dois *corpora* de estudo, um *corpus* paralelo e um *corpus* comparável. O *corpus* paralelo é composto por textos de divulgação escritos originalmente em língua inglesa e suas respectivas traduções para o português, extraídos do portal MedlinePlus (U.S. NATIONAL LIBRARY OF MEDICINE, 2020); e o comparável, com essas traduções para o português e textos originalmente escritos em português brasileiro do mesmo gênero, publicados pelo Ministério da Saúde (BRASIL, 2018). Primeiramente, esses *corpora* foram analisados pelas ferramentas Coh-Metrix (GRAESSER *et al*., 2017) e Coh-Metrix-Port (NILC, 2020), para verificação do Índice Flesch – que estima a inteligibilidade de um texto. Palavras-chave e colocações foram identificadas por meio do AntConc (ANTHONY, 2019), enquanto o AntPConc (ANTHONY, 2017) foi usado para alinhar trechos originais e respectivas traduções, a fim de que se investigassem as palavras em contexto.

Em resumo, nossa investigação visa responder a duas perguntas: (i) de acordo com as ferramentas de inteligibilidade, os textos escritos em português e os traduzidos para esse idioma se mostram acessíveis para o público-alvo? e (ii) tomando como premissa que a convencionalidade característica dos gêneros textuais facilita a compreensão, os textos traduzidos revelam padrões observados em textos de divulgação escritos originalmente em português brasileiro?

Este artigo está dividido em cinco seções. Após esta introdução, será apresentada a fundamentação teórica, onde serão abordados conceitos de Linguística de *Corpus*, tradução, acessibilidade textual e gênero, pertinentes a este estudo. A terceira seção delineará a metodologia da pesquisa, apresentando os *corpora* de estudo e as ferramentas utilizadas

[1] Este artigo apresenta resultados de pesquisa de Mestrado desenvolvida no âmbito do Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul.

para o levantamento dos dados. A quarta seção apresentará e discutirá os resultados obtidos. Por fim, apresentaremos as considerações finais do estudo.

2 Abordagem teórica

Para o desenvolvimento desta pesquisa, apoiamo-nos nos pressupostos da inteligibilidade textual, da tradução e da Linguística de *Corpus* para analisarmos textos de divulgação da área médica em português. Abaixo relatamos os conceitos de cada área julgados relevantes para o estudo.

2.1 Acessibilidade textual

Assim como as rampas de acesso para cadeirantes ou os elevadores, que podem ser observados em prédios públicos, e o piso tátil para pessoas cegas, que é visto pelas ruas, a acessibilidade atua, no âmbito textual, como um facilitador para o entendimento da mensagem pelo leitor (FINATTO, 2020). Dessa forma, a acessibilidade textual pode ser entendida como uma condição desejada de qualidade de texto, evitando que ele imponha barreiras linguísticas ao público leitor, para que este tenha condições de compreender as informações compartilhadas. Para os fins deste trabalho, apresentamos a inteligibilidade e a convencionalidade como aspectos cruciais para analisar a acessibilidade dos textos.

A fim de estimar o quão acessíveis são os textos, foram desenvolvidas fórmulas de inteligibilidade, que demonstram, matematicamente, a adequação de um texto para determinado público.

O Índice Flesch é o mais aplicado para pesquisas que se baseiam em estimativas de inteligibilidade, que, é importante ressaltar, não deve ser confundida com legibilidade. A primeira está relacionada ao que é inteligível, ou seja, aquilo que é de fácil compreensão, enquanto a segunda se refere ao que é legível, ou seja, que está claro e nítido (DUBAY, 2004). A legibilidade tem relação com design e *layout* – como a diagramação, o tipo e o tamanho de fonte etc. –, relacionando-se à acessibilidade visual. Ambas estão ligadas a formas de compreensão, respectivamente, à compreensão de forma mental e à compreensão de forma visual. Neste artigo, trataremos apenas de inteligibilidade, já que nos ocupamos apenas dos aspectos linguísticos do texto.

Desenvolvido originalmente para o inglês, o Índice Flesch leva em consideração duas variáveis: o comprimento médio das frases, ou seja, o número de palavras do texto dividido pelo número total de sentenças; e a média de sílabas por palavras, resultado do número total de sílabas dividido pelo número de palavras do texto. Esse índice foi adaptado para a língua portuguesa por pesquisadores do Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo (MARTINS *et al*., 1996), que tinham como objetivo adequá-lo à realidade das palavras e sílabas do português brasileiro, já que o comprimento das palavras e das frases em língua portuguesa difere bastante daquele em língua inglesa.

A fim de exemplificarmos a importância da inteligibilidade textual, mencionamos um estudo realizado no ano de 2003, nos Estados Unidos, e relatado por DuBay (2004), que apontou que acidentes de trânsito vinham causando cerca de 46% das mortes acidentais de crianças, apesar de se esperar que o uso correto do bebê conforto, da cadeirinha ou do assento para as crianças reduzisse os riscos de fatalidades em 71%. Na verdade, a maioria desses assentos vinha sendo instalada incorretamente. A partir dessa conclusão, foi conduzido um estudo para analisar a inteligibilidade dos manuais de instalação desses assentos. Descobriu-se, então, que os 107 manuais analisados foram escritos em uma linguagem considerada difícil para 80% dos adultos no país, privados, portanto, da total compreensão das instruções descritas no material (DUBAY, 2004). Diante do exposto, podemos concluir que, apesar de terem domínio da língua em que os textos estavam escritos – inglês –, os leitores não tinham total acesso à mensagem que esses textos pretendiam transmitir. Daí a importância de se produzirem textos inteligíveis, especialmente aqueles que tenham como objetivo instruir o leitor correta e adequadamente.

Para tirar o maior proveito da comunicação, o leitor espera reconhecer nos textos padrões já observados em outros de mesmo gênero, ou seja, os textos devem revelar convenções características. Segundo Tagnin (2013), a convencionalidade ocorre em três níveis linguísticos, a saber, (i) sintático – relacionado à forma como os elementos se combinam gramaticalmente; (ii) semântico – com relação ao significado dos elementos que constituem as frases, sentenças, expressões idiomáticas etc.; e (iii) pragmático – quanto à interação entre os participantes do ato comunicativo. É importante ressaltar que as convenções linguísticas são consagradas pelo uso, nem sempre 'respeitando' uma lógica conceitual – por exemplo, em português convencionou-se dizer 'achados e perdidos',

apesar de esses acontecimentos ocorrerem em ordem inversa, ordem, esta, respeitada na expressão em língua inglesa – '*lost and found*'. Portanto, a convencionalidade está relacionada ao domínio da fluência de determinada língua (TAGNIN, 2013), já que os falantes dispõem de unidades linguísticas previamente armazenadas na memória.

Assim, ao produzir um enunciado, repetem-se sintagmas já utilizados, causando a cristalização de padrões linguísticos. Ao lermos textos de divulgação, esperamos identificar os padrões típicos desse gênero, a fim de absorvermos conteúdos novos de forma clara e precisa. Portanto, com base no que foi exposto, acreditamos que, caso se constate quebra da convencionalidade nos textos traduzidos, quando comparados àqueles originalmente escritos em determinado idioma, poderia haver prejuízos à acessibilidade textual, aspecto imprescindível para a divulgação de informações, especialmente aquelas de utilidade pública.

2.2 Os estudos de tradução e a Linguística de *Corpus*

A abordagem funcionalista da tradução, proposta por Nord (2006) com base em textos especializados, coloca em foco o propósito do texto traduzido, sempre com vistas a atingir seu público-alvo. Em outras palavras, o texto traduzido deve 'funcionar' apropriadamente para aqueles que o recebem, assim como se espera que o texto de partida tenha funcionado para seu público leitor. Dessa forma, além de fazer escolhas terminológicas adequadas às diferentes áreas de especialidade, o tradutor deve avaliar as capacidades de compreensão e cooperação de sua audiência, antecipando os possíveis efeitos que determinadas escolhas textuais poderão ter sobre o leitor. Nesse sentido, *corpora* textuais não só auxiliam os tradutores a evidenciar características próprias dos textos de partida, como também os ajudam a encontrar soluções na língua de chegada, já que evidenciam os padrões indicativos de convencionalidade (STEWART, 2000).

Ao se aliarem à Linguística de *Corpus*, os estudos de tradução possibilitaram apontamentos importantes sobre textos traduzidos em comparação com textos originais. No que diz respeito às tendências tradutórias, Baker (1993) cita o exemplo da explicitação, que fornece ao leitor da tradução informações que no original estavam subentendidas. Nessa direção, a tradução apresenta também movimentos de desambiguação e facilitação, como, por exemplo, o uso de formas

pronominais mais precisas, possibilitando que o leitor identifique o referente sem muito esforço.

Frankenberg-Garcia (2006) apresenta resultados similares para a língua portuguesa. A partir das análises feitas por meio do COMPARA,[2] a pesquisadora atesta que "as traduções tendem a ser mais longas do que os textos-fonte, tanto na direção inglês-português como na direção português-inglês" (FRANKENBERG-GARCIA, 2006, p. 147),[3] corroborando o que foi apresentado em Baker (1993).

Além das tendências tradutórias, Baker (1993) menciona estudos sobre a chamada "terceira língua"[4] na tradução, que consiste no resultado do confronto entre língua-fonte e língua-alvo, imprimindo ao texto traduzido características que o distanciam tanto do texto-fonte quanto de textos originalmente produzidos na língua-alvo. Portanto, a forma como o texto-fonte influencia o texto-alvo pode resultar em perda de padrões característicos dos gêneros textuais, comprometendo a acessibilidade. Aqui, entendemos essa "terceira língua" como a quebra da convencionalidade, ou seja, dos padrões observados nos textos escritos originalmente nas respectivas línguas e culturas, a depender das características dos gêneros em questão.

2.3 Texto de divulgação

Bhatia (1993, p. 49) define gênero como "um evento comunicativo reconhecível, caracterizado por um conjunto de propósito(s) comunicativo(s) identificado(s) e mutuamente compreendidos pelos membros da comunidade profissional ou acadêmica em que geralmente ocorre."[5] Textos de divulgação, assim como outros gêneros textuais, possuem características específicas, sendo a mais marcante delas o fato de estarem situados entre o falar científico e o falar comum, apresentando vocabulário desses dois registros. Nesse sentido,

---

[2] *Corpus* paralelo bidirecional de português e inglês. Disponível em: https://www.linguateca.pt/COMPARA/Bem-vindos.html. Acesso em: 25 ago. 2020.

[3] No original: "translations tended to be longer than source texts in both the English-Portuguese and the Portuguese-English directions."

[4] No original: "third code".

[5] No original: "a recognizable communicative event characterized by a set of communicative purpose(s) identified and mutually understood by the members of the professional or academic community in which it regularly occurs."

> seria redutor pensar a divulgação científica apenas como uma redução ou adaptação de textos científicos, elaborados para a leitura de pares dotados de uma mesma competência profissional. Ao contrário, a divulgação científica em seu amplo universo, ainda carente de descrições, afigura-se como uma categoria textual autêntica, com regras próprias de produção de significação e de recursos que visam a uma comunicação eficiente. (KRIEGER, 2009, p. 9-10).

Massarani e Moreira (2005) distinguem três categorias na comunicação científica: i) os discursos científicos primários, escritos por pesquisadores para pesquisadores; ii) os discursos didáticos, como os manuais científicos para ensino; e iii) os discursos de divulgação científica. Enquanto o destinatário do texto científico é um par com conhecimento especializado sobre o tema, o texto de divulgação é voltado para pessoas leigas, sem, necessariamente, conhecimento prévio construído sobre o que será abordado no texto (SANTIAGO, 2007). Andreetto (2013, p. 8) corrobora essa visão:

> Apesar de os artigos acadêmicos também terem o objetivo de informar, a organização dessa informação segue padrões diferentes daqueles na área da divulgação, que embora também tenham como objetivo transmitir conhecimento, o fazem respeitando a convencionalidade esperada por seu público alvo, que normalmente é leigo em uma dada área de especialidade.

A partir de uma pesquisa empírica com profissionais brasileiros da área médica que necessitavam se preparar para um exame de proficiência linguística em inglês contendo textos de divulgação desse domínio especializado, Andreetto (2013) concluiu que esses especialistas tinham mais dificuldade para compreender textos de divulgação do que textos acadêmicos, já que estavam mais habituados aos padrões destes. Portanto, observamos que a adequada compreensão da informação não depende exclusivamente do nível de escolaridade do público leitor, mas prevê a familiaridade com os discursos a que esse público é exposto, ou seja, aos padrões reconhecidos por determinada comunidade discursiva (cf. BHATIA, 1993).

Dessa forma, ao escrever um texto de divulgação, o especialista deve transmitir a mensagem de maneira diferente do que faria se estivesse compartilhando informações com outros especialistas no assunto. Além da terminologia adotada, as estruturas linguísticas características desse

gênero textual também devem ser reconhecidas pelo leitor. Ou, segundo Zamboni (2001 *apud* ANDREETTO, 2013), para essa tarefa, deve-se transformar o discurso científico em discurso do "cotidiano".

3 Metodologia

Conforme mencionado na Introdução deste artigo, para este estudo foram realizadas análises quantitativas, por meio dos índices de inteligibilidade e chavicidade, e também análises qualitativas, auxiliadas por ferramentas de Linguística de *Corpus*, pois, para esgotar o objetivo da investigação proposta, a pesquisa não poderia se manter apenas no âmbito estatístico, mas sim tomá-lo como base para uma investigação mais aprofundada. Afinal, de acordo com Biderman (1967), os "primeiros senões facilmente apreensíveis são constituídos pelos dois aspectos irredutíveis da realidade linguística: o elemento qualitativo e o quantitativo", reiterando a importância de se analisarem os textos por esses dois vieses. Nesta seção, explicaremos a construção dos *corpora* de estudo, bem como as ferramentas utilizadas nas análises desses *corpora*.

3.1 Os *corpora* de estudo

A fim de compararmos as características de textos de divulgação da área médica escritos originalmente em português com aqueles traduzidos, e verificarmos se há diferenças significativas no que tange à convencionalidade do gênero, compilamos um *corpus* paralelo – originais em inglês e traduções para o português – e um *corpus* comparável, composto pelos mesmos textos traduzidos, e textos escritos originalmente em português.

O *corpus* paralelo foi compilado a partir de textos de divulgação do site MedlinePlus, portal da U.S. National Library of Medicine (2020), que publica informações sobre saúde para pacientes, seus familiares e amigos em diversas línguas, abrangendo sintomas e tratamentos, compondo um material que, de acordo com o site, é "de fácil leitura".[6] Esse *corpus* é composto pelos 66 textos em inglês e suas traduções para o português disponíveis.

O *subcorpus* em português apresenta estruturas bastante similares ao da sua contraparte de traduções, pois foi compilado com base em critérios compartilhados específicos desse gênero. Por exemplo, a grande

[6] No original: "Easy-to-Read Materials".

maioria dos textos tem subtítulos dividindo os diferentes tópicos; há, também, listas de itens introduzidas por elementos gráficos, que tornam as informações visualmente mais claras (DUBAY, 2004). Esse *subcorpus* contempla os 191 textos da Biblioteca Virtual em Saúde, mantida pelo Ministério da Saúde.

Ressaltamos que nosso objetivo não era balancear os *corpora* em número de textos ou *tokens*, mas, sim, coletar a totalidade de textos disponíveis em ambos os portais. A Tabela 1 resume os números relativos aos *corpora* de estudo.

TABELA 1 – Números de *types* e *tokens* dos *subcorpora* da pesquisa

Corpus	MedlinePlus (EN)	MedlinePlus (PT)	Ministério da Saúde (PT)
Textos	66	66	191
Tokens	34.765	39.476	84.085
Types	3.088	4.554	9.666
TTR*	8,88%	11,53%	11,49%

**Type-Token Ratio*

Fonte: Elaborada pelas autoras.

É interessante notar que, comparando-se a extensão média dos textos que compõem os *corpora*, os textos originalmente escritos em inglês são mais longos (526 *tokens*, em média) do que aqueles escritos em português (440 *tokens*), corroborando as conclusões de Rebechi (2017) e Fuchs (2018) sobre as diferenças nos tamanhos dos textos de mesmo gênero escritos em português brasileiro e em inglês estadunidense. Já as traduções para o português ficaram mais longas (598 *tokens*) do que os textos de partida, contrariando os estudos citados, mas confirmando os achados de Baker (1993) e Frankenberg-Garcia (2006) sobre a tendência de explicitação do texto traduzido.

O cálculo de *type-token ratio* indica a porcentagem da riqueza lexical do corpus, sendo demonstrado pelo cálculo *types*÷*tokens*×100. De acordo com Berber Sardinha (2004, p. 94), "Quanto maior o seu valor, mais palavras diferentes o texto conterá. Em contraposição, um valor baixo indicará um número alto de repetições, o que pode indicar um texto menos rico ou variado, do ponto de vista de seu vocabulário". Por estarmos lidando com línguas diferentes, o índice calculado para o inglês não pode ser comparado com o do português. O que podemos

constatar, a partir desse levantamento, é que os índices de riqueza lexical dos textos de divulgação em português são muito próximos, com uma diferença de apenas 0,04% a mais para as traduções.

3.2 As ferramentas de análise

Para a análise de inteligibilidade, as ferramentas utilizadas foram o Coh-Metrix (GRAESSER *et al*., 2017) e o Coh-Metrix-Port (NILC, 2020), para os textos em inglês e em português, respectivamente. O índice utilizado para a análise foi o Índice Flesch, que considera duas métricas para estimar o nível de dificuldade do texto: o comprimento médio das frases e a média de sílabas por palavras. Como resultado do cálculo, obtém-se um índice que pode ir de 0 a 100, sendo que quanto mais próximo de 0, mais difícil seria o texto; e quanto mais perto de 100, mais fácil.

Vale ressaltar que, de acordo com Graesser *et al*. (2004), as fórmulas matemáticas de inteligibilidade, como o Índice Flesch, ignoram componentes linguísticos e discursivos que influenciam a compreensão textual. Dessa forma, apesar de os parâmetros de tamanho de sentenças e de palavras serem indicativos de legibilidade, eles não conseguem revelar com precisão e por si só a complexidade de um texto. Por exemplo, aspectos como número de palavras diferentes, frequência de uso das palavras e sua regularidade ou irregularidade seriam interessantes de serem estudados em um cálculo de inteligibilidade. Por esse motivo, se viu a necessidade de aliar análises de inteligibilidade a análises manuais, partindo-se, também, de dados quantitativos revelados por ferramentas computacionais.

A análise de padrões linguísticos foi feita por meio do AntConc (ANTHONY, 2019), um *software* gratuito, que pode ser facilmente baixado e utilizado off-line. Com base em cálculos estatísticos, ele possibilita que sejam feitos levantamentos de palavras características do *corpus* de estudo – palavras-chave –, de padrões linguísticos, como colocações e *clusters*, entre outras aplicações. A partir desses levantamentos, o *software* permite que sejam feitas análises quantitativas aliadas ao olhar atento do pesquisador.

O *software* AntPConc (ANTHONY, 2017) foi utilizado na investigação do *corpus* paralelo, depois que os dois *subcorpora* – originais em inglês e traduções em português – foram devidamente alinhados, por meio do LF Aligner (FARKAS, 2018).

4 Análises e discussão

Nesta seção, explicitaremos as análises quantitativas e qualitativas realizadas a partir dos *corpora* de estudo, e apresentaremos e discutiremos os resultados dos levantamentos, que nos guiaram na reflexão sobre a adequação dos textos de divulgação da área médica para o público-alvo, ou seja, a população de não especialistas.

4.1 Levantamento de inteligibilidade

Iniciamos a análise quantitativa dos *corpora* de estudo a partir do levantamento de inteligibilidade, possibilitado pelo cálculo automático do Índice Flesch. Para esse levantamento, primeiramente foi aplicado o Índice Flesch à íntegra dos textos que compõem o *corpus* paralelo, ou seja, dos 66 textos em língua inglesa do MedlinePlus e suas respectivas traduções em língua portuguesa. Já para o levantamento do Índice Flesch dos textos escritos originalmente em língua portuguesa, foi utilizada uma amostra de 66 dos 191 textos, a fim de equiparar ao número de textos do *corpus* paralelo.

Ressaltamos que a ferramenta adaptada para a língua portuguesa (SCARTON *et al*., 2009) possui algumas limitações em comparação com a da língua inglesa. Por exemplo, o Coh-Metrix-Port processa textos com até mil palavras, ao passo que o Coh-Metrix limita a análise a 15 mil caracteres. Apesar da diferença entre os parâmetros utilizados, observamos que o Coh-Metrix aceita textos mais longos do que o Coh-Metrix-Port. Por isso, todos os textos em inglês foram processados pelo Coh-Metrix, enquanto o Coh-Metrix-Port processou somente 59 textos do MedlinePlus (PT), já que os outros sete ultrapassavam o limite de palavras processáveis pela ferramenta. Quanto aos textos do Ministério da Saúde, todos foram processados, já que não extrapolaram o limite.

Como mencionado anteriormente, os resultados do cálculo do Índice Flesch vão de 0 a 100, variando entre ‘muito difícil’ e ‘muito fácil’. Na Tabela 2, a seguir, é possível observar a escala de valores, além de uma estimativa do nível de escolaridade que compreende cada um desses graus de inteligibilidade.

TABELA 2 – Interpretação do Índice Flesch

Valor do Índice	**Descrição de Inteligibilidade**	**Escolaridade Estimada (EUA)**	**Escolaridade Estimada (BR)**
0 a 29	Muito difícil	*College graduate*	Universitários*
30 a 49	Difícil	*13th to 16th grade*	EM ou universitários
50 a 59	Razoavelmente difícil	*10th to 12th grade*	EM
60 a 69	Padrão	*8 th to 9 th grade*	Até 8ª série do EF
70 a 79	Razoavelmente fácil	*7 th grade*	Até 8ª série do EF
80 a 89	Fácil	*6 th grade*	Até 8ª série do EF
90 a 100	Muito fácil	*5 th grade*	Até 4ª série do EF

* Apenas para áreas acadêmicas específicas.
EM Ensino Médio
EF Ensino Fundamental
Fonte: Elaborada com base em Flesch (1949 *apud* DUBAY, 2004) e Martins *et al*. (1996).

Vale ressaltar que a tabela apresenta uma comparação somente em relação ao nível de escolaridade entre a população brasileira e a estadunidense, sem considerar fatores como nível de proficiência em leitura ou classe social dos indivíduos.

Com base nos valores de Índice Flesch levantados para cada texto que compõe os *corpora*, foram feitos os cálculos estatísticos de média, mediana, variância e desvio padrão de cada *subcorpus*, utilizando-se a ferramenta de equações do programa Excel. A média e a mediana são medidas de tendência central: a média consiste na soma de todos os elementos em análise (os índices) divididos pelo número total de elementos (número de textos processados); e a mediana é o valor central da série. Já a variância e o desvio padrão são medidas de dispersão, servindo para apontar a variabilidade dos dados em torno da média. A variância mostra o quão distantes os valores estão da média, sendo calculada a partir da soma dos quadrados da diferença entre cada valor e a média, dividida pelo número total de elementos. O desvio padrão é a medida do grau de dispersão em relação à média. Para calculá-lo, basta extrair a raiz quadrada da variância (MORATO, 2011).

As medidas de dispersão dos textos em inglês do MedlinePlus (EN) foram calculadas aplicando-se a equação de população, pois todos os textos foram avaliados pelo Coh-Metrix. Já dos *subcorpora* em

português, foram calculadas as medidas de dispersão de uma amostra, pois não foram avaliados todos os textos que os compõem, conforme explicamos anteriormente. Os dados estatísticos podem ser observados na Tabela 3, a seguir.

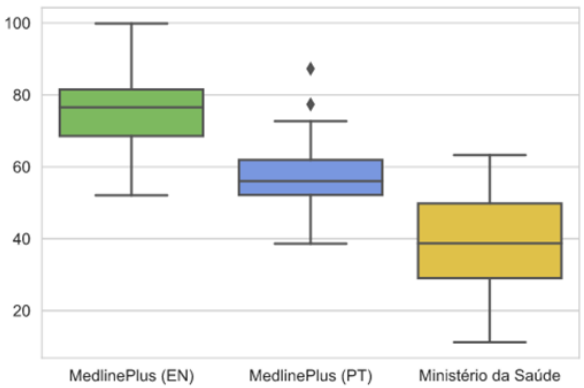
TABELA 3 – Resultado de cálculos estatísticos descritivos

	MedlinePlus (EN)	**MedlinePlus (PT)**	**Ministério da Saúde**
Média	74,845	57,659	39,115
Mediana	76,516	55,99	38,699
Variância	102,501	72,0444	170,942
Desvio padrão	10,124	8,488	13,074

Fonte: Elaborada pelas autoras.

A partir desses dados estatísticos levantados, foram construídos gráficos, a fim de facilitar a visualização e comparação entre os levantamentos dos diferentes *corpora*. O Gráfico 1, do tipo *boxplot*, apresenta visualmente os dados de mediana, que corresponde à linha horizontal no centro dos blocos, e de desvio padrão, que corresponde aos blocos coloridos. As linhas horizontais fora dos blocos representam o valor máximo e o valor mínimo observados nos textos de cada *subcorpus*. Por fim, os pequenos losangos representam valores discrepantes entre os textos pertencentes a cada amostra (*subcorpus*).

GRÁFICO 1 – *Boxplot* dos dados do Índice Flesch



Fonte: Elaborado pelas autoras.

A partir dos dados apresentados, é possível observar que as médias calculadas para o Índice Flesch apontam os textos do MedlinePlus (EN) com índices de inteligibilidade mais altos (média de 74,845), os do *subcorpus* traduzidos para o português, intermediários (57,659), e o *subcorpus* dos textos do Ministério da Saúde, originalmente escritos em português, com índices de inteligibilidade mais baixos (39,115). Além disso, os valores de desvio padrão mostram que os índices do *subcorpus* do Ministério da Saúde são os que contam com maior dispersão – de 13,074 –, seguidos pelos do MedlinePlus (EN) – de 10,124. Por último, vêm os do MedlinePlus (PT), com menor dispersão – de 8,488. O valor mínimo dos índices do MedlinePlus (EN) é de 52,05, o central (mediana) é de 76,516, e o máximo é de 99,821. No MedlinePlus (PT), o valor mínimo é de 38,654, o central é de 55,99, e o máximo, 72,367, com 77,367 e 87,276 como valores discrepantes. Para o *subcorpus* do Ministério da Saúde, observamos valor mínimo de 11,234, valor central de 38,699, e máximo de 63,272.

O acesso à informação é um direito assegurado na lei (BRASIL, 2015). Contudo, para que a população possa usufruir desse direito, devem-se garantir condições não só de acesso, mas também de compreensão desses materiais. Além disso, sendo a tradução imprescindível para a disseminação do conhecimento científico, que, muitas vezes, é transmitido em língua inglesa, ao traduzir textos com a finalidade de atingir o público geral, o tradutor deve levar em conta informações sobre o leitor a quem as traduções se destinam (NORD, 2006). Portanto, para analisar esses dados, é necessário utilizar como base informações sobre os leitores médios estadunidenses e brasileiros.

Segundo dados de 2017, 90% da população estadunidense possui, no mínimo, *High School* completo (U.S. CENSUS BUREAU, 2017). Já no Brasil, de acordo com o último censo, a parcela da população que tem grau equivalente de instrução, ou seja, Ensino Médio, não chega aos 50%. A discrepância fica ainda mais evidente se considerarmos que, enquanto 46,2% da população brasileira possui Ensino Médio ou graus mais altos de instrução, há uma considerável parcela de 33,7% que não completou ao menos o Ensino Fundamental (IBGE, 2019). Assim, ao produzir conteúdo escrito para o público geral, os órgãos competentes devem considerar esses dados, pois dentre os 53,8% da população brasileira que possui graus de instrução inferiores ao Ensino Médio, a maioria se concentra no Ensino Fundamental incompleto. Portanto, para

se mostrarem adequados ao leitor médio, os textos brasileiros deveriam se encaixar no nível 'fácil', ou seja, terem índices de inteligibilidade entre 70 e 100. Já os textos estadunidenses, para serem acessíveis à maior parcela da população, poderiam apresentar inteligibilidade 'difícil', com índices entre 30 e 60, mas nossos levantamentos mostraram que são 'razoavelmente fáceis', podendo atingir, inclusive, a pequena parcela da população com níveis inferiores de escolaridade. Assim, com base nos levantamentos apresentados e nos dados de escolaridade, podemos concluir que a população que mais necessitaria que as informações fossem fornecidas de maneira acessível seria, justamente, a população que a receberia com maior complexidade, ou seja, a brasileira.

A partir desses dados estatísticos de inteligibilidade textual, procederemos à análise de padrões linguísticos, utilizando os *softwares* AntConc e AntPConc, a fim de identificarmos o que recorre em textos da área e verificarmos se tais padrões são mantidos nos textos traduzidos. Em outras palavras, pretendemos avaliar se os textos traduzidos mantêm a convencionalidade característica do gênero textual em estudo, que será analisada com base nos textos escritos originalmente em português. Nesse sentido, enfatizamos a necessidade de não apenas olhar para métricas superficiais de inteligibilidade, mas também para os padrões que são imprescindíveis para a caracterização dos gêneros e, portanto, para a sua compreensão pela comunidade discursiva a que se destina.

4.2 Levantamento de padrões linguísticos

A análise qualitativa manual dos textos de divulgação em língua portuguesa partiu das palavras-chave, ou seja, das palavras estatisticamente mais frequentes no *corpus* de estudo do que no *corpus* de referência. A decisão de olhar somente para palavras-chave das traduções se deve ao nosso interesse em buscar por padrões observados nos textos originalmente escritos em português e verificar se eles se mantêm nos textos traduzidos ou se estes indicariam quebras de convencionalidade, dificultando o entendimento do texto pelo público não especialista. Para esse contraste, compilamos textos também da área médica, contudo de um gênero distinto – artigos científicos –, a fim de que fossem destacadas as palavras estatisticamente relevantes relacionadas ao gênero dos textos, não necessariamente à área de especialidade.

Para a montagem do *corpus* de referência em português, foram selecionados 117 textos, compilados a partir da palavra de busca 'medicina' no Portal de Periódicos Capes, totalizando 450 mil *tokens*. O número de palavras do *corpus* de referência baseou-se nos tamanhos críticos apontados por Berber Sardinha (2004), sendo cinco vezes o tamanho do maior *subcorpus* de estudo em português.

Segundo Gabrielatos (2018), a chavicidade de um item deve ser estabelecida utilizando-se uma combinação de duas métricas que se complementam: tamanho de efeito e significância. A primeira é usada para determinar a diferença entre as frequências de determinada palavra nos dois *corpora* – quanto maior a diferença entre o *corpus* investigado e o *corpus* de referência, maior será a razão logarítmica –, enquanto a segunda aponta o tamanho dessa diferença. Nesta pesquisa, utilizamos, respectivamente, o *odds ratio* e o *log likelihood*, disponíveis no AntConc.

O *odds ratio* será levado em consideração para determinar o ponto de corte para as palavras serem ou não consideradas chave, estabelecido em 10. O *log-likelihood* (*Keyness*, no AntConc) foi estabelecido em $p < 0{,}0001$, indicando uma margem de erro de até 0,01% para as palavras-chave terem sido levantadas erroneamente. De acordo com Brezina (2018), para pesquisas na área de Ciências Humanas, é aceitável utilizar até $p < 0{,}05$. O *log-likelihood* foi o índice adotado para a organização das palavras-chave, apresentadas em ordem decrescente de chavicidade. De acordo com Pojanapunya e Todd (2018), o teste *log-likelihood* enfatiza palavras relativamente comuns, que servem ao propósito de pesquisas orientadas pelo gênero – interesse deste estudo –, ao passo que o *odds ratio* dá ênfase a palavras mais especializadas, que são mais adequadas para pesquisas terminológicas, por exemplo.

Por não se tratar de um *corpus* etiquetado morfossintaticamente e devido ao número de flexões possíveis na língua portuguesa (em substantivos, adjetivos, artigos e verbos), optamos por fazer a lematização manual das palavras-chave do *corpus* comparável. As palavras foram agrupadas sob aquela com o valor de *log-likelihood* mais alto. Por exemplo, os itens do *subcorpus* do MedlinePlus (PT) 'precisa' (com *log-likelihood* de 95,08) e 'precisará' (45,34) foram agrupadas sob 'precisar' (97,83).

Palavras homógrafas foram diferenciadas manualmente. Por exemplo, 'sente', que pode ser conjugação dos verbos 'sentir' e 'sentar', teve suas frequências mantidas separadamente. Após a distinção do número de ocorrências referentes a cada um dos verbos, foi recalculado

o *odds ratio* de cada uma das formas, utilizando-se uma calculadora de chavicidade e efeito,[7] a fim de determinar se esses itens seguiriam ou não sendo palavras-chave, de acordo com os critérios estipulados.

Após a lematização manual, a lista de palavras-chave exclusivas do MedlinePlus (PT) mostrou 167 itens; a lista exclusiva do Ministério da Saúde, 115. Uma amostra da lista de palavras-chave do *corpus* comparável, organizada por chavicidade, pode ser observada na Tabela 4.

TABELA 4 – Amostra de palavras-chave exclusivas do *corpus* comparável

	MedlinePlus (PT)		**Ministério da Saúde**	
Posição	**Frequência**	**Palavras-chave**	**Frequência**	**Palavras-chave**
1	393	**seu**	153 [+25]	alimentos; alimento
2	251	médico	135	evitar
3	154	dor	48	camisinha
4	134	cirurgia	38	lixo
5	86 [+35]	tomar; tome	45	dentes
6	64	ligue	58	mulher
7	93	poderá	49	pé
8	100	fazer	35 [+16]	picada; picadas
9	116	pessoas	38	roupas
10	92	depois	43 [+21]	acidentes; acidente
11	52 [+20]	incisão; incisões	39	coluna
12	44	catapora	38	objetos
13	83	**sinais**	40	provocar
14	56	reação	25	manchas
15	35	vaginal	37	passo
16	55	**use**	24	hpv
17	35	alérgica	27	nariz
18	62	semanas	25	chão
19	48 [+13; +12; +36; +11; +12]	ajuda; ajudá[-lo/-la]; ajudam; ajudar; ajudará; ajude	36	veias
20	36	converse	22 [+18]	joelhos; joelho

Fonte: elaborada com base em AntConc (ANTHONY, 2019).

[7] Disponível em: http://ucrel.lancs.ac.uk/llwizard.html. Acesso em: 15 maio 2020.

A partir da lista de palavras-chave, levantamos os itens que ocorreram estatisticamente com mais frequência nos textos traduzidos, sendo excluídos, contudo, aqueles que denominam doenças, como 'catapora' e 'glaucoma', sintomas de doenças, como 'tontura' e 'erupção', e partes do corpo, já que se referem a questões específicas, que podem não ter sido abordadas na mesma proporção nos dois *subcorpora*. Privilegiamos, portanto, as palavras-chave que, em tese, poderiam ocorrer em qualquer texto de divulgação da área médica, mas que não aparecem na lista de palavras-chave do *subcorpus* de textos originais em português.

Devido à limitação de espaço, neste artigo focaremos (i) o pronome adjetivo possessivo 'seu', palavra com maior chavicidade nos textos traduzidos; (ii) o substantivo plural 'sinais', na décima terceira posição, que é o primeiro substantivo exclusivo do *subcorpus* traduzido que se enquadra nos pré-requisitos da análise (o item 'pessoas' aparece na forma singular nas palavras-chave do *subcorpus* de originais em português); e (iii) o verbo no imperativo 'use', na décima sexta, já que, além de ser palavra-chave apenas dos textos traduzidos, é um entre os 28 verbos no modo imperativo dessa lista, enquanto a de palavras-chave dos textos originais compreende apenas dois – 'utilize' e 'retire'.

4.2.1 Análise de 'seu'

Como é possível observar na amostra de palavras-chave (TABELA 4), 'seu' é a primeira palavra estatisticamente relevante característica do *subcorpus* do MedlinePlus (PT), com 393 ocorrências (99,55 a cada 10.000). Para entender o seu papel nesse *corpus*, partimos para a análise dos colocados da palavra. Primeiro, buscamos os colocados em uma janela de 4 palavras à direita. A partir da lista de colocados, foi possível observar que as palavras que mais recorrem à direita de 'seu' são 'médico' (176 ocorrências), seguida de 'bebê' (48) e 'filho' (30). Já nos textos originais em português, há 193 ocorrências de 'seu', e, na mesma janela de até 4 palavras à direita, o substantivo 'médico' é colocado desse pronome adjetivo possessivo apenas 25 vezes. Nesse *subcorpus* (MS), o colocado mais frequente à esquerda de médico é 'o' (34 ocorrências). Esses índices parecem ainda mais discrepantes se considerarmos que o *subcorpus* do MS tem mais do que o dobro de palavras do MedlinePlus (PT).

A fim de confirmarmos se a alta recorrência da colocação 'seu médico' no MedlinePlus (PT) decorre da influência dos textos de partida, buscamos os colocados imediatamente à esquerda de '*doctor*' – ou seja,

janela de 1 item. Os resultados corroboraram a hipótese levantada, mostrando que nos textos o colocado '*your*' aparece com alta frequência com '*doctor*' – 204 ocorrências. A Tabela 5 sintetiza esses dados.

TABELA 5 – Colocados imediatamente à esquerda de 'médico' e '*doctor*' nos *corpora* da pesquisa

MedlinePlus (EN)			**MedlinePlus (PT)**			**Ministério da Saúde**		
Freq.	**Colocado**	**Item de busca**	**Freq.**	**Colocado**	**Item de busca**	**Freq.**	**Colocado**	**Item de busca**
204	*your*	doctor	168	seu	médico	27	o	médico
21	*the*		53	o		25	um	
8	*baby's*		8	ao		23	seu	
			6	um		17	pelo	
						7	ao	
						6	atendimento	
						6	do	

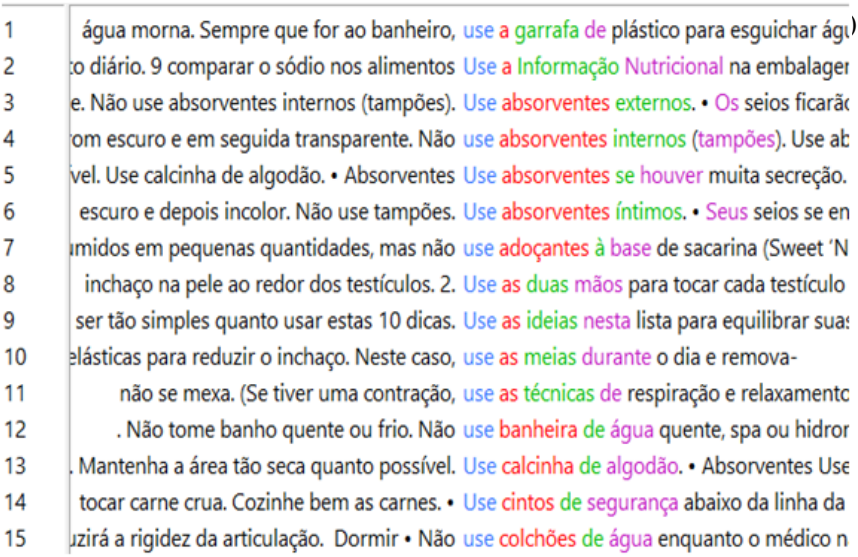
Fonte: elaborada com base em AntConc (ANTHONY, 2019).

A partir desse levantamento, foi possível observar que, apesar de 'médico' ser colocado de 'seu' no *subcorpus* do Ministério da Saúde, na verdade, 'o médico' e 'um médico' são mais utilizadas do que 'seu médico'. Já no *subcorpus* de textos traduzidos, a sequência mais utilizada é 'seu médico', enquanto 'o médico' ocorre apenas 53 vezes, quase $\frac{1}{3}$ das ocorrências daquela. Essas duas sequências mais utilizadas vão ao encontro do que ocorre no *subcorpus* dos textos originais, sendo '*your doctor*' a sequência mais utilizada, com 204 ocorrências, e '*the doctor*', a segunda mais utilizada, com 21 ocorrências. Também é possível observar que, no MedlinePlus (PT), 'o médico' ocorre mais que o dobro de vezes de '*the doctor*' no Medline (EN), sendo a escolha de tradução para '*your doctor*' em 43 ocorrências.

4.2.2 Análise de 'use'

A forma verbal 'use' tem 55 ocorrências (13,93 a cada 10.000 palavras) no *subcorpus* do MedlinePlus (PT), fazendo dela uma palavra-chave. A partir das linhas de concordância, foi possível observar construções de 'use' com diversos substantivos. Por exemplo, observamos

a utilização de sequências como ‘use absorventes’ (4 ocorrências), ‘use tampões’ (4 ocorrências) e ‘use preservativo(s)’ (2 ocorrências). Algumas linhas de concordância que exemplificam o tipo de orientações são apresentadas na Figura 1.



Fonte: AntConc (ANTHONY, 2019).

Como ‘use’ não aparece entre as palavras-chave do Ministério da Saúde, observamos a lista de palavras-chave a fim de identificar que outra possibilidade poderia ter sido a escolhida pelos autores. Identificamos, então, a palavra-chave ‘utilize’, que ocorre 17 vezes nesse *subcorpus* (MS) (2,02 a cada 10.000), enquanto ‘use’ ocorre 41 vezes (4,88 a cada 10.000) – mas pode ter ficado de fora das palavras-chave por ser também recorrente no *corpus* de referência. De fato, diversos dicionários, como, por exemplo, o Dicionário Online de Português[8] e o Caldas Aulete Digital,[9] apresentam ‘utilizar’ como sinônimo de ‘usar’. Supomos, assim, que a maior recorrência de ‘use’ nos textos traduzidos pudesse ser decorrente da forma cognata ‘*use*’ dos textos em inglês. A fim

[8] Disponível em: https://www.dicio.com.br/usar/. Acesso em: 26 set. 2020.
[9] Disponível em: http://www.aulete.com.br/usar. Acesso em: 26 set. 2020.

de confirmar ou refutar tal hipótese, procedemos com o alinhamento do *corpus* paralelo, por meio da ferramenta LF Aligner (FARKAS, 2018), e o investigamos usando o *software* AntPConc (ANTHONY, 2017). A Figura 2 apresenta uma amostra do alinhamento de '*use*', partindo dos textos em inglês.

FIGURA 2 – Linhas de concordância de '*use*' no *corpus* paralelo do MedlinePlus (PT-EN)

Line	KWIC
1	Use MyPlate to build your healthy eating style ar
2	Use a clean part of the washcloth and wash the
3	Use a nipple for your baby's age.
4	• Use a piece of paper and a pen to mark
5	Use a pillow or folded blanket across your abdon
6	Use a pillow or folded blanket over your incision
7	• Use a sitz bath to relieve discomfort.
8	ess than 20 feet from any window, door, or vent. Use an extension cord that is more than 20 feet l
9	• Use an inflatable, donut-shaped, ring when sittin

Line	Reference
1	Fazer escolhas alimentares para uma vida saudável pode ser tão simples quanto usar estas 10 dicas.
2	Use uma parte limpa da toalha de mão para lavar o outro olho.
3	Use um bico compatível com a idade do bebê.
4	• Use uma folha de papel e uma caneta e anote os movimentos.
5	Use um travesseiro ou manta dobrada sobre o abdômen ou peito para proteger as incisões quando tossir.
6	Use um travesseiro ou manta dobrada sobre a incisão como apoio enquanto estiver respirando profundamente ou tossindo.
7	• Use um banho de assento para aliviar o desconforto.
8	Não instale geradores, lavadoras de alta pressão ou motores à gasolina a uma distância menor do que 6 metros de qualquer janela, porta ou saída de ar.
9	• Use um anel inflável como uma pequena bóia quando estiver sentada.

Fonte: AntPConc (ANTHONY, 2017).

Antes de procedermos para a análise das linhas em paralelo, é importante ressaltar que a forma '*use*', em inglês, pode remeter ao substantivo e ao verbo em forma de infinitivo, além das conjugações em primeira e segunda pessoas do singular e das três pessoas do plural. A análise das 90 ocorrências de *use*, em inglês, revelou que a palavra foi traduzida como (i) diferentes formas do verbo 'usar', 'utilizar' e 'consumir' (62, 3 e 1 ocorrências, respectivamente), (ii) os substantivos 'uso' e 'consumo' (7 e 2 ocorrências, respectivamente), (iii) por paráfrases – por exemplo, a frase "*Each time you use the toilet*" foi traduzida por

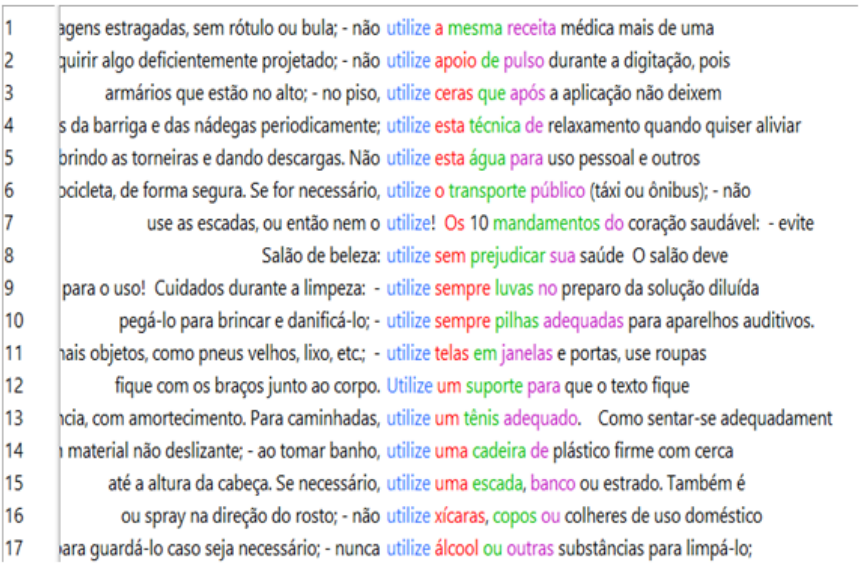
"Todas as vezes que for ao banheiro" – (7 ocorrências). Para as restantes oito ocorrências de '*use*', não foram identificados equivalentes.

A busca em direção contrária revelou que as 55 ocorrências de 'use' nas traduções são provenientes de '*use*' (41 ocorrências), '*wear*' (6 ocorrências), '*take*' (1 ocorrência), '*spend*' (1 ocorrência), '*douche*' – traduzido por 'use ducha vaginal' – (1 ocorrência), e cinco das suas ocorrências não foram traduzidas. Já a forma 'utilize' foi a escolha tradutória para apenas uma frase – "*Never use a generator inside your home*…" ["Nunca utilize este gerador dentro de casa…"]. Portanto, podemos concluir que 'use' é a tradução mais recorrente do cognato '*use*', enquanto 'utilize' e outras possibilidades são preteridas nas traduções.

Diferentemente do que foi observado no *subcorpus* traduzido, há recorrência tanto de 'use' quanto de 'utilize' nos textos originalmente escritos em português. A palavra 'use' ocorre 41 vezes (4,88 a cada 10.000) e 'utilize', 17 (2,02 a cada 10.000). A fim de verificar se existe diferença de uso entre esses dois verbos no *subcorpus* de textos originais em português, observamos seus colocados na janela de 4 itens, mais especificamente os substantivos à direita dos verbos. Esse levantamento mostrou que as ocorrências de 'use' no *subcorpus* do Ministério da Saúde estão bastante ligadas a objetos que são portados ('chapéu', 'cintos', 'roupas', 'sapatos' e 'óculos escuros'), medicamentos e, como observado anteriormente, a forma verbal 'use' aparece próxima à palavra 'camisinha'.

Já a palavra 'utilize', no *subcorpus* do Ministério da Saúde, ocorre em contextos semelhantes aos de 'use' no *subcorpus* do MedlinePlus (PT). Aparecem formulações como 'não utilize', 'nunca utilize' e 'utilize sempre', seguidos de alguns substantivos, porém sem recorrências, conforme as linhas de concordância a seguir (FIGURA 3).

FIGURA 3 – Linhas de concordância de 'utilize' no *subcorpus* do MedlinePlus (PT)



1	agens estragadas, sem rótulo ou bula; - não	utilize	a mesma receita médica mais de uma
2	quirir algo deficientemente projetado; - não	utilize	apoio de pulso durante a digitação, pois
3	armários que estão no alto; - no piso,	utilize	ceras que após a aplicação não deixem
4	s da barriga e das nádegas periodicamente;	utilize	esta técnica de relaxamento quando quiser aliviar
5	brindo as torneiras e dando descargas. Não	utilize	esta água para uso pessoal e outros
6	bicicleta, de forma segura. Se for necessário,	utilize	o transporte público (táxi ou ônibus); - não
7	use as escadas, ou então nem o	utilize!	Os 10 mandamentos do coração saudável: - evite
8	Salão de beleza:	utilize	sem prejudicar sua saúde O salão deve
9	para o uso! Cuidados durante a limpeza: -	utilize	sempre luvas no preparo da solução diluída
10	pegá-lo para brincar e danificá-lo; -	utilize	sempre pilhas adequadas para aparelhos auditivos.
11	ais objetos, como pneus velhos, lixo, etc.; -	utilize	telas em janelas e portas, use roupas
12	fique com os braços junto ao corpo.	Utilize	um suporte para que o texto fique
13	ncia, com amortecimento. Para caminhadas,	utilize	um tênis adequado. Como sentar-se adequadament
14	material não deslizante; - ao tomar banho,	utilize	uma cadeira de plástico firme com cerca
15	até a altura da cabeça. Se necessário,	utilize	uma escada, banco ou estrado. Também é
16	ou spray na direção do rosto; - não	utilize	xícaras, copos ou colheres de uso doméstico
17	ara guardá-lo caso seja necessário; - nunca	utilize	álcool ou outras substâncias para limpá-lo;

Fonte: AntConc (ANTHONY, 2019).

Com exceção das linhas 3 e 6, nas quais 'utilize' está associado a itens de vestuário, as outras ocorrências se combinam a objetos que servirão para ajudar a cumprir determinada ação. Por exemplo, na linha 1, recomenda-se utilizar uma cadeira de plástico como auxílio para tomar banho. Contudo, devido ao tamanho reduzido dos *corpora* compilados para este estudo, julgamos inapropriado traçarmos generalizações sobre os contextos de ocorrência de 'use' e 'utilize'. Por isso, recorremos a um *corpus* de grandes proporções, mas de língua geral, o Corpus do Português (DAVIES, 2015). Esse *corpus*, que está disponível gratuitamente on-line, permite que seja feita a comparação entre os colocados de duas palavras por meio da ferramenta *Compare*.

Apesar de se tratar de um *corpus* geral de língua portuguesa, que abrange textos de diferentes países falantes da língua, o *subcorpus* Web/Dialects permite que a busca seja feita apenas em textos em português brasileiro, com cerca de 655 milhões de palavras. Para traçar um paralelo com os levantamentos do AntConc, optamos por levantar somente os substantivos que aparecem à direita dos verbos. A janela escolhida foi, também, de até 4 palavras.

A Tabela 6 apresenta os 25 primeiros substantivos que ocorrem à direita dos verbos ‘use’ e ‘utilize’, organizados pelo valor do índice (*score*) de relação entre o colocado e a palavra de busca.

TABELA 6 – Colocados de ‘use’ e ‘utilize’ no *subcorpus* Web/Dialects

Colocado de ‘use’	**Freq.**	**Score**	**Colocado de ‘utilize’**	**Freq.**	**Score**
curador	357	216,5	selo	101	333,1
bálsamo	211	128	créditos	136	224,3
camisinha	206	124,9	fins	575	99,8
comentários	88	53,4	login	106	87,4
autoridade	67	40,6	polegares	26	85,8
box	119	36,1	palhas	11	72,6
mente	44	26,7	pagamento	20	66
blusas	40	24,3	responsabilidade	114	62,7
maquiagem	66	20	fórum	26	42,9
vestido	33	20	x	141	42,3
mouse	237	18	email	123	14
perfume	29	17,6	sistemas	12	13,2
criatividade	342	17,3	formulário	139	10,2
kit	114	17,3	formas	33	8,4
instrumentos	104	15,8	rolagem	24	7,9
chicote	25	15,2	senha	118	7,6
calça	48	14,6	aparelhos	29	7,4
jeans	24	14,6	botão	321	6,9
saias	24	14,6	dispositivo	10	6,6
branco	23	13,9	seção	13	6,1
talento	22	13,3	forma	149	6,1
vestidos	22	13,3	letras	54	5,9
chapéu	21	12,7	código	117	5,8
salto	21	12,7	tag	12	5,7
seleção	21	12,7	aço	13	5,4

Fonte: Corpus do Português (DAVIES, 2015).

Corroborando os resultados observados no *subcorpus* de textos originais em português, no *subcorpus* de língua geral ‘use’ se associa

majoritariamente a objetos que são portados, especialmente peças de vestuário – 'blusas', 'maquiagem', 'vestido', 'instrumentos', 'calça', 'jeans', 'saias', 'vestidos', 'chapéu' e 'salto'. Já a forma verbal 'utilize' é empregada com referência a palavras como 'créditos', 'pagamento', 'responsabilidade', 'sistemas', 'aparelhos', 'dispositivo', 'forma', 'letras', dentre outras. Esses resultados indicam que nem sempre o equivalente 'use' é a opção mais convencional para recuperar '*use*' nos textos traduzidos, mas que outras possibilidades – 'utilize', além de paráfrases – deveriam ser consideradas a fim de se manter a convencionalidade observada em textos desse gênero escritos originalmente em português.

4.2.3 Análise de 'sinais'

Diante da alta chavicidade de 'sinais' nos textos traduzidos, diferentemente do que ocorre nos textos originalmente escritos em português, selecionamos mais essa palavra para fazermos uma análise quantitativa, a fim de identificar o porquê da discrepância.

A partir do alinhamento dos textos em inglês e suas traduções, buscamos identificar a(s) palavra(s) que tivesse(m) originado a tradução 'sinais'. Observamos, então, que das 83 ocorrências de 'sinais' no *subcorpus* de traduções, apenas 11 não partiram de '*signs*', mas foram resultado de explicitações, escolhas tradutórias de '*changes*' e '*cues*', ou simplesmente não tiveram o excerto de partida identificado.

Por meio desse levantamento, verificamos que, das 121 ocorrências (34,81 a cada 10.000) de '*sign(s)*' como substantivo – observamos, também, duas ocorrências do verbo '*sign*' [assinar] –, 83 foram traduzidas por 'sinal'/'sinais', 31 por 'sintomas', e as restantes foram traduzidas por 'efeitos' (1 ocorrência), ou simplesmente não foram traduzidas (7 ocorrências).

Verificamos, posteriormente, que a palavra 'sintomas' ocorre nos textos traduzidos 52 vezes no total (13,17 a cada 10.000) – sendo 3 vezes na forma singular 'sintoma'. Por observar que havia ainda 21 ocorrências da palavra com origens desconhecidas, alinhamos originais e traduções a partir da palavra de busca 'sintoma(s)'. Das 21 ocorrências, 4 são traduções do cognato '*symptoms*'. Outras ocorrências têm origem em frases nas quais não se utiliza a palavra '*signs*' nem '*symptoms*', ou seja, Ø (10 ocorrências) – podendo ser observada na Figura 4. Por fim, há 7 ocorrências para as quais não foram localizadas correspondentes em inglês.

FIGURA 4 – Linhas de concordância de ‘sintomas’ no *corpus* paralelo do MedlinePlus (PT-EN)

Line	KWIC
38	spirar, batimentos cardíacos acelerados, tontura e fraqueza. Esses sintomas normalmente cc
39	spirar, batimentos cardíacos acelerados, tontura e fraqueza. Esses sintomas normalmente cc
40	• sintomas parecidos com c
41	es que sejam causadas por um vírus não afetado pela vacina ou • sintomas parecidos com a
42	e um pronto-socorro ou entre em contato com seu médico se os sintomas piorarem ou se \
43	Contate seu médico caso estes sintomas piorem ou não c
44	Sarampo O vírus do sarampo causa sintomas que podem indu

Line	Reference
38	
39	
40	
41	
42	• Return to the Emergency Department or call your doctor if your signs get worse or you have a fever of more than 100.5 degrees F or 38 degrees C.
43	Call your doctor if this gets worse or does not go away in a few weeks.
44	MEASLES (M) can cause fever, cough, runny nose, and red, watery eyes, commonly followed by a rash that covers the whole body.

Fonte: AntPConc (ANTHONY, 2017).

Vale ressaltar que foram encontradas apenas 4 ocorrências (1,15 a cada 10.000) de ‘*symptoms(s)*’ nos textos originais em inglês, ao passo que seu cognato em português, ‘sintoma(s)’, ocorre 288 vezes (72,96 a cada 10.000) no Ministério da Saúde, o que parece indicar que os equivalentes *prima facie sign(s)* → sinal/sinais e *symptom(s)* → sintomas não são necessariamente utilizados em contextos semelhantes. A fim de tentar mapear se ocorre alguma distinção entre os contextos de ‘sinais’ e de ‘sintomas’ entre traduções e originais em português, foi feita a pesquisa por colocados. O levantamento de colocados de ‘sinais’ foi feito respeitando-se a janela de até 4 itens à direita e à esquerda, estabelecendo a frequência mínima de 6 ocorrências.

A partir das linhas de concordância mostradas pelo AntPConc e AntConc, é possível observar que os usos de ‘sinais’ e de ‘sintomas’ parecem estar indiscriminados no *subcorpus* traduzido. Ou seja, não é possível distinguir os contextos em que os substantivos aparecem. Assim como aparece próximo a nomes de doenças (como ‘catarata’, ‘asma’ e ‘ataque cardíaco’), ela também aparece perto de indícios de problemas de saúde (‘estresse’, ‘fome’ e ‘ruptura ou descolamento de retina’).

No *subcorpus* do Ministério da Saúde, enquanto 'sintoma(s)' ocorre 288 vezes (34,25 a cada 10.000), 'sinal/sinais' ocorre 54 vezes (6,42 a cada 10.000). Vale enfatizar que a palavra 'sintomas' aparece com frequência como um subtítulo dos textos, fazendo com que sua frequência aumente significativamente.

Em razão do número elevado de ocorrências, investigamos o motivo pelo qual a palavra não consta na lista de palavras-chave desse *corpus*. A partir de sua frequência, os índices da palavra foram determinados a partir da calculadora de chavicidade e efeito. O *log-likelihood* da palavra é de 354,16 e o *odds ratio* é de 5,8163, que não alcança o ponto de corte para figurar na lista de palavras-chave. No *corpus* de referência de artigos científicos, a palavra 'sintomas' tem 280 ocorrências (6,19 a cada 10.000). Portanto, o *odds ratio* da palavra acabou sendo neutralizado, fazendo com que ela não fosse considerada chave.

Há ocorrências da palavra 'sinais' que se dão na sequência 'sinais e sintomas' (16 vezes). Essa sequência é, muitas vezes, utilizada como um subtítulo para organizar as partes do texto, como mencionado anteriormente sobre a palavra 'sintomas'.

As linhas de concordância de 'sinais' no *subcorpus* do Ministério da Saúde parecem indicar aspectos em comum em seus contextos de uso (FIGURA 5).

FIGURA 5 – Linhas de concordância de 'sinais' no *subcorpus* do Ministério da Saúde

9	. Sintomas e sinais de alerta: Muitos sintomas são comuns aos
10	desidratação. Sinais de desidratação: - olhos fundos; - ausência
11	- observar os sinais de desidratação. Sinais de desidratação: - o
12	menstrual e sinais de desnutrição. Diagnóstico: A doença só p
13	iverem dado sinais de erupção, é necessário procurar o dentista
14	idamente os sinais de gravidade da doença, a tratar adequadam
40	istem outros sinais que indiquem que a fraqueza é ou
41	Os primeiros sinais são: fraqueza, transpiração, palidez,
42	bilização dos sinais vitais. Lembre-se: Não abra mão da

Fonte: AntConc (ANTHONY, 2019).

Podemos observar que, por exemplo, em 'sinais de desidratação', a primeira característica é 'olhos fundos'. Em outro momento, é possível

ver a frase ‘Os primeiros sinais são: fraqueza, transpiração, palidez [...]’. Aqui, podemos observar que se usa ‘sinais’ para aspectos visíveis (‘fraqueza’, ‘transpiração’ e ‘palidez’).

Outro aspecto é que, em geral, a palavra ‘sinais’ não aparece associada diretamente às doenças, como no *subcorpus* do MedlinePlus (‘sinais de asma’; ‘sinais de doença arterial coronariana’; ‘sinais de derrame’; ‘sinais de glaucoma’; dentre outros). Pode-se observar que a palavra ‘sinais’ é mais associada a manifestações de problemas de saúde, por exemplo, ‘sinais de desidratação’, ‘sinais de desnutrição’, ‘sinais de erupção’, dentre outros.

Em um primeiro momento, apenas pelas linhas de concordância, é possível observar que, de fato, há uma distinção entre ‘sinais’ e ‘sintomas’. A fim de comprovar a existência dessa distinção, optamos por fazer este mesmo levantamento em um *corpus* de língua geral.

Utilizamos, novamente, os textos brasileiros do Corpus do Português, cujo papel, nesse caso, é de extrema relevância para comprovar os contextos convencionais de uso dessas palavras que são familiares ao público geral.

Assim como o levantamento de colocados feito nos *subcorpora* de estudo, mantivemos como padrão a janela de até 4 palavras. Esse levantamento foi feito apenas para colocados à direita das palavras ‘sinais’ e ‘sintomas’. Os primeiros 25 colocados de ‘sinais’ e de ‘sintomas’ que ocorrem nessa janela estão listados na Tabela 7. Pode-se observar suas frequências de co-ocorrência e o resultado estatístico da associação entre as palavras.

TABELA 7 – Colocados de ‘sinais’ e ‘sintomas’ no *subcorpus* Web/Dialects

Colocado de ‘sinais’	**Freq.**	**Score**	**Colocado de ‘sintomas’**	**Freq.**	**Score**
prodígios	491	983,9	TPM	159	317,4
tempos	354	709,4	psicóticos	157	313,4
maravilhas	277	555,1	hemolítico	112	223,6
vitais	243	486,9	urêmico	110	219,6
trânsito	233	466,9	sujeitos	93	185,6
elétricos	181	362,7	poliúria	84	167,7
libras	152	304,6	relatadas	76	151,7
rádio	149	298,6	artrite	150	149,7

distintivos	133	266,5	TDAH	75	149,7
pontuação	133	266,5	mosquito	62	123,8
emitidos	109	218,4	tratamentos	57	113,8
aparições	98	196,4	fibromialgia	51	101,8
gráficos	86	172,3	neuróticos	51	101,8
céu	81	162,3	queixas	87	86,8
seguirão	81	162,3	taquicardia	43	85,8
digitais	80	160,3	menopausa	167	83,3
sol	77	154,3	gastrite	39	77,8
vinda	74	148,3	cabeça	72	71,9
luminosos	147	147,3	ascensão	34	67,9
enviados	70	140,3	histéricos	34	67,9
sonoros	70	140,3	sinto	66	65,9
arrombamento	68	136,3	psiquiátricos	33	65,9
milagres	131	131,3	duram	32	63,9
espécie	112	112,2	alérgicos	31	61,9
terra	56	112,2	pré-menstrual	30	59,9

Fonte: Corpus do Português (DAVIES, 2015).

Por meio do levantamento, foi possível confirmar que há distinção entre as manifestações de ‘sinais’ e de ‘sintomas’ no *corpus* de língua geral. A palavra ‘sinais’ aparece associada a ‘trânsito’, ‘aparições’, ‘gráficos’ e ‘arrombamento’. Isso dá indícios de que a maneira como os ‘sinais’ se manifestam é de forma visual. Ou seja, são traços que você pode observar e enxergar. Por exemplo, ‘sinais de arrombamento’ em uma casa podem ser portas quebradas, com vestígios de terem sido forçadas por alguém, bagunça nos cômodos, demonstrando que alguém esteve por ali procurando algo. Assim, de acordo com os colocados observados, ‘sinais’ são traços que podem ser identificados por uma terceira pessoa observadora.

Além disso, podem ser observados colocados de ‘sinais’ mais relacionados à noção física de um conjunto que carrega informações ou dados. Aparecem colocados como ‘elétricos’, ‘rádio’, ‘emitidos’, ‘digitais’, ‘luminosos’, ‘enviados’ e ‘sonoros’, que indicam haver uma influência de contextos mais especializados.

Já associadas à palavra ‘sintomas’ podemos observar ocorrências como ‘psicótico’, ‘poliúria’, ‘neuróticos’, ‘taquicardia’, ‘histéricos’,

'psiquiátricos', 'alérgicos' e 'pré-menstrual'. Ou seja, essas são manifestações que as pessoas sentem mais do que visualizam. Assim, em oposição às manifestações de sinais, que podem ser observadas por uma terceira pessoa, as manifestações de sintomas parecem ser mais facilmente identificadas pela própria pessoa. Até porque, no geral, antes de procurar atendimento que confirme a doença, é necessário que o paciente reconheça os sintomas, para então partir para a análise de uma terceira pessoa (médico), para quem serão relatadas as manifestações. Além dessas manifestações, há algumas doenças e distúrbios que estão na lista de colocados, como '[síndrome] hemolítico urêmica', 'artrite', 'TDAH' (Transtorno do Déficit de Atenção com Hiperatividade), 'fibromialgia' e 'gastrite'.

No *subcorpus* traduzido do MedlinePlus, a palavra 'sinais' parece ser utilizada de forma indiscriminada, aparecendo diversas vezes como um falso sinônimo de 'sintomas' e como equivalente de '*signs*'. Já a palavra 'sintomas', por se tratar de uma palavra motivada pelo uso de '*symptoms*' ou por Ø, está sendo empregada seguindo um padrão semelhante aos aspectos aqui apresentados.

4.3 Discussão

Retomando as médias de Índice Flesch, para o *corpus* comparável, obtiveram-se índices de 57,659 para o *subcorpus* do MedlinePlus (PT) e 39,115 para o do Ministério da Saúde; para a língua inglesa, a média observada foi de 74,845 para o *subcorpus* do MedlinePlus (EN) – lembrando que índices mais próximos de 100 apontam para maior grau de facilidade, enquanto mais próximos de 0 demonstram maior grau de dificuldade. Com base nos níveis de escolaridade das populações estadunidense e brasileira, o índice de inteligibilidade dos textos em português não estão adequados para o seu público geral, ao passo que até mesmo o norte-americano com pouca escolaridade seria capaz de compreender os textos em inglês. O intervalo mais adequado dos índices para os textos em português seria de classificação 'fácil', entre 70 e 100. Já os textos em inglês, para serem acessíveis à maior parcela da população estadunidense, poderiam apresentar inteligibilidade 'difícil', com índices entre 30 e 60.

Enfatizamos que o levantamento quantitativo de inteligibilidade parte de noções superficiais do texto, quais sejam, o comprimento médio de palavras e de sentenças. Por esse motivo, o Índice Flesch

por si só não pode determinar com precisão o nível de dificuldade de um texto. Pode-se depreender, a partir dos resultados referentes a esse índice, que os textos originalmente escritos em português contam com estruturas, de forma geral – levando em consideração especificamente as palavras e as frases –, mais longas do que os originais em inglês e suas traduções. Entretanto, a investigação sobre a convencionalidade das traduções, por ser uma análise de padrões linguísticos desenvolvida em comparação com textos escritos originalmente na língua portuguesa, teve o papel imprescindível de indicar com mais precisão aspectos que podem dificultar a compreensão de um texto por leitores médios. Por isso, buscamos associar o levantamento puramente estatístico à análise qualitativa, já que consideramos indispensável que haja um olhar mais aprofundado sobre as questões linguísticas do texto para determinar se este conta ou não com barreiras que influenciam sua acessibilidade. Até o momento, não há ferramentas que permitam que a investigação sobre os padrões linguísticos seja desenvolvida com precisão a partir de análises exclusivamente quantitativas.

Nesse sentido, as ocorrências de 'use', no MedlinePlus (PT), e de 'utilize', no Ministério da Saúde, ocorrem em contextos similares, em que é dada alguma orientação para o leitor do que fazer quando se deparar com determinadas situações. Descobrimos que há, de fato, distinção entre aplicações de 'use' e de 'utilize', sendo que o primeiro se refere, principalmente, a objetos (com frequência, itens de vestuário), e o segundo é empregado com noções mais abstratas (como 'créditos', 'pagamento', 'sistemas' etc.). Essa descoberta foi feita a partir das linhas de concordância do *subcorpus* de textos originalmente escritos em português e, posteriormente, foi validada por meio do Corpus do Português (DAVIES, 2015).

A partir dos colocados de 'seu', foi possível observar nos textos traduzidos grande incidência da palavra 'médico'. Isso chamou a atenção porque, nos textos do Ministério da Saúde, os principais colocados imediatamente à esquerda de 'médico' são os artigos 'o' e 'um'. Vale ressaltar que das 251 ocorrências (63,58 a cada 10.000) de 'médico' no MedlinePlus (PT), 168 estão precedidas de 'seu'; enquanto no Ministério da Saúde, das 130 ocorrências (15,46 a cada 10.000) da palavra, apenas 23 ocorrem ao lado de 'seu'. Isso demonstra que ocorre uma interferência do texto em inglês sobre o texto traduzido, visto que no inglês, das 247

ocorrências (71,05 a cada 10.000) de '*doctor*', 204 estão antecedidas por '*your*'.

Outra clara interferência dos textos-fonte é o uso de 'sinais' e 'sintomas' nos textos traduzidos. Enquanto há distinção entre os usos de 'sinais' e 'sintomas' no *subcorpus* dos textos escritos originalmente em português, nas traduções há indícios de que a motivação se dê puramente pelo uso de palavras cognatas no texto-fonte. Prova disso é que a grande maioria das ocorrências de '*signs*' foram traduzidas como 'sinais' (83 de 117 ocorrências), enquanto todas as ocorrências de '*symptoms*' foram traduzidas para 'sintomas'. Adicionalmente, quando houve uso de Ø nos originais em inglês foram traduzidas como 'sintomas' (ou seja, não havia emprego nem da palavra '*symptoms*' nem de '*signs*'), aparece o emprego de 'sintomas' nas traduções, fazendo com que a palavra seja empregada em contextos mais similares entre si. Isso pode ser explicado pela influência do texto original sobre as decisões do tradutor, sendo que, quando há Ø, o tradutor consegue se desprender do padrão de uso da língua inglesa. Por exemplo, pode-se observar que 'sintomas' aparece associado a 'raiva', 'nervosismo', 'irritabilidade' e 'febre', manifestações essas que podem ser sentidas pela pessoa doente.

Pôde ser estabelecida a diferença entre os contextos de uso de 'sinais' e 'sintomas' a partir do *subcorpus* do Ministério da Saúde. Posteriormente, buscamos colocados também no Corpus do Português (DAVIES, 2015), para fazer a comprovação dos resultados em um corpo de textos que abrange diversas facetas da língua. Depreendemos que as ocorrências de 'sinais' estão mais ligadas a demonstrações que podem ser vistas, enquanto 'sintomas' se refere a sensações.

5 Considerações finais

No que diz respeito aos resultados quantitativos, foi possível traçar algumas conclusões em relação aos levantamentos de palavras-chave e n-gramas. A partir dos levantamentos de colocados das palavras-chave, concluímos que os textos traduzidos apresentavam, em diversos momentos, quebras de convencionalidade (TAGNIN, 2013), distanciando-se dos padrões utilizados nos textos escritos originalmente em português. Essas quebras de convencionalidade ocorrem devido ao uso de palavras cognatas do inglês e de traduções *prima facie*, fugindo dos padrões esperados para o português.

Vale ressaltar que quebras de convencionalidade, causadas pela influência do texto-fonte sobre o texto-alvo, podem gerar dificuldades no entendimento do texto pelo leitor médio. Como exemplo, cita-se o uso indiscriminado de 'sinais' e 'sintomas' no *subcorpus* do MedlinePlus (PT). Ao deparar com um texto destinado ao público geral tratando de doenças, é comum que parte dele foque em abordar os sintomas. Entretanto, no âmbito dos textos de divulgação da área médica, o uso de 'sinais' pode causar uma quebra de expectativa no leitor, podendo fazer com que ele não entenda a que tipo de manifestações o texto se refere, por exemplo.

Além disso, por estar com a mente e o olhar no texto original, o tradutor pode acabar não associando a ocorrência de uma palavra a opções de tradução que vão além do seu cognato, como se observa na tradução de '*use*' pelo cognato 'use'. Esse fenômeno configura o que Baker (1993) denomina "terceira língua" na tradução, que é quando o texto traduzido fica com características do texto-fonte, distanciando-se, assim, de padrões de convencionalidade da língua-alvo.

Por essas razões, consideramos que a convencionalidade, sendo analisada a partir de dados quantitativos levantados a partir de critérios delimitados, pode auxiliar na avaliação da acessibilidade de um texto com maior precisão do que somente a aplicação de métricas que partem de noções superficiais do texto – como é o caso do Índice Flesch, que se baseia somente em medidas de comprimento de frases e de palavras. Nesse sentido, salienta-se a importância do uso de uma metodologia que associe a análise quantitativa à análise qualitativa. A associação dessas investigações, possibilitada por ferramentas que fazem cálculos estatísticos somadas ao olhar do pesquisador, permite o enriquecimento dos resultados da pesquisa.

Em relação a limitações do estudo, a maior delas foi o tamanho dos *corpora*. Ao utilizar *corpora* paralelos em análises, é comum que seu tamanho não seja tão expressivo quanto o de textos escritos originalmente na língua. Isso ocorre porque o número de textos que se enquadram nos critérios de seleção acaba se reduzindo, pois necessita-se que estejam disponíveis tanto o texto-fonte quanto o texto-alvo. Devido a essa barreira, foi necessário recorrer diretamente à análise das linhas de concordância em momentos que o levantamento de colocados era inconclusivo. Também foi necessário recorrer a um *corpus* maior, de

língua geral, a fim de confirmar os padrões de colocação de determinadas palavras-chave.

Além disso, há outras limitações impostas ao estudo pelos *softwares* utilizados para analisar os *corpora*. O Coh-Metrix-Port estabelece um número máximo de palavras por texto para a análise, o que fez com que parte dos textos do *subcorpus* do MedlinePlus (PT) não fossem analisados pela ferramenta. Ademais, vale ressaltar que o *software* AntConc não realiza etiquetagem ou lematização de *corpus*; portanto, a fim de reduzir o número de palavras-chave dos *corpora*, foi necessário partir para a lematização manual.

Por fim, julgamos que a quebra de convencionalidade em traduções, conforme resultados apresentados na seção 4.2, compromete mais a compreensão do texto do que o uso de palavras ou frases consideradas longas. Portanto, acreditamos que a inteligibilidade está diretamente relacionada à convencionalidade, ou seja, à manutenção de padrões reconhecíveis pelo público leitor dos diferentes gêneros textuais.

Declaração das contribuições de cada autora

As autoras Yuli Souza Carvalho e Rozane Rodrigues Rebechi produziram colaborativamente este artigo. A pesquisa de Mestrado relatada no texto foi desenvolvida por Yuli Souza Carvalho, sob a orientação de Rozane Rodrigues Rebechi, no Programa de Pós-Graduação em Letras da Universidade Federal do Rio Grande do Sul. Yuli Souza Carvalho escreveu uma primeira versão do texto, participando da escrita de todas as seções, principalmente das seções 2, 3 e 4, revisando as versões seguintes do manuscrito, assim como formatando a versão final. Rozane Rodrigues Rebechi participou da escrita de todas as seções, principalmente do Resumo e do Abstract, da Introdução e das Considerações Finais, bem como da revisão de todas as seções do manuscrito.

Referências

ANDREETTO, M. D. *Por que os textos de divulgação são mais difíceis para aprendizes de leitura com necessidades específicas do que textos científicos?* Um estudo direcionado pelo corpus. 2013. 172f. Dissertação (Mestrado em Estudos Linguísticos e Literários em Inglês) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2013.

ANTHONY, L. *AntConc*. Versão 3.5.8. Tokyo: Waseda University, 2019.

ANTHONY, L. *AntPConc*. Versão 1.2.1. Tokyo: Waseda University, 2017.

BAKER, M. Corpus Linguistics and Translation Studies: Implications and Applications. *In*: BAKER, M.; FRANCIS, G.; TOGNINI-BONELLI, E. (eds.). *Text and Technology*: In Honour of John Sinclair. Philadelphia: John Benjamins, 1993. p. 233-250. DOI: https://doi.org/10.1075/z.64.15bak

BERBER SARDINHA, T. *Lingüística de Corpus*. São Paulo: Manole, 2004.

BHATIA, V. K. *Analysing Genre*: Language Use in Professional Settings. London; New York: Routledge, 1993.

BIDERMAN, M. T. C. Estatística linguística. *Alfa*, São Paulo, v. 11, p. 117-128, 1967.

BRASIL. *Lei nº 13.146, de 6 de julho de 2015*. Institui a Lei Brasileira de Inclusão da Pessoa com Deficiência (Estatuto da Pessoa com Deficiência). Brasília: Secretaria-Geral da Presidência da República, 2015.

BRASIL. Ministério da Saúde. *Biblioteca Virtual em Saúde*. Brasília: Ministério da Saúde, 2018.BREZINA, V. *Statistics in Corpus Linguistics*: A Practical Guide. Cambridge: Cambridge University Press, 2018. DOI: https://doi.org/10.1017/9781316410899

DAVIES, M. *Corpus do Português*. Provo: Brigham Young University, 2015. Disponível em: https://www.corpusdoportugues.org/. Acesso em: 12 jun. 2020.

DUBAY, W. H. *The Principles of Readability*. California: Impact Information, 2004.

FARKAS, A. *LF Aligner*. Versão 4.2. [s. l.]: Source Forge, 2018.

FINATTO, M. J. B. Acessibilidade textual e terminológica: promovendo a tradução intralinguística. *Revista Estudos Linguísticos*, São José do Rio Preto, v. 49, n. 1, p. 72-96, 2020. DOI: https://doi.org/10.21165/el.v49i1.2775

FLESCH, R. *The Art of Readable Writing*. Nova York: Harper, 1949.

FRANKENBERG-GARCIA, A. Using a Parallel Corpus in Translation Practice and Research. *In*: CONFERÊNCIA DE TRADUÇÃO PORTUGUESA, 1., 2006, Caparica, Portugal. *Actas da Contrapor*. Lisboa: [S.n.], 2006. p. 142-148.

FUCHS, S. N. *Orientações culturais e suas implicações para a tradução funcionalista*: um estudo na área do turismo à luz da Linguística de Corpus. 2018. 366f. Tese (Doutorado em Estudos da Tradução) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, SãoPaulo, 2018.

GABRIELATOS, C. Keyness Analysis: Nature, Metrics and Techniques. *In*: MARCHI, A.; TAYLOR, C. (ed.). *Corpus Approaches to Discourse*: A Critical Review. London: Routledge, 2018. p. 225-258. DOI: https://doi.org/10.4324/9781315179346-11

GRAESSER, A. C. *et al*. *Coh-Metrix*. Version 3.0. Tennessee: University of Memphis, 2017.

GRAESSER, A. C. *et al*. Coh-Metrix: Analysis of Text on Cohesion and Language. *Behavioral Research Methods*, [*S.l.*], v. 36, n. 2, p. 193-202, 2004. DOI: https://doi.org/10.3758/BF03195564

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA. IBGE. *Pesquisa Nacional por Amostra de Domicílios Contínua.* Educação 2018. Rio de Janeiro: IBGE, 2019.

JENKINS, J. English as a lingua franca: Interpretations and Attitudes. *World Englishes*, [*S.l.*], n. 28, v. 2, p. 200-207, 2009. DOI: https://doi.org/10.1111/j.1467-971X.2009.01582.x

KRIEGER, M. G. Divulgação científica e terminologia. *In*: SIMPÓSIO INTERNACIONAL DE ESTUDOS DE GÊNEROS TEXTUAIS, 5., 2009, Caxias do Sul. *Anais* [...]. Caxias do Sul: UCS, 2009. p. 1-11

MARTINS, T. B. F. *et al*. Readability Formulas Applied to Textbooks in Brazilian Portuguese. *Notas do ICMSC*, São Paulo, n. 28, p. 1-11, 1996.

MASSARANI, L.; MOREIRA, I. C. A retórica e a ciência: dos artigos originais à divulgação científica. *MultiCiência*, Campinas, n. 4, p. 1-12, 2005.

MORATO, R. G. *Conceitos básicos de Estatística Descritiva*. São Paulo: Universidade de São Paulo, 2011.

NORD, C. Loyalty and Fidelity in Specialized Translation. *Confluências*, [*S.l.*], n. 4, p. 29-42, 2006.

NÚCLEO INTERINSTITUCIONAL DE LINGUÍSTICA COMPUTACIONAL (NILC). *Coh-Metrix-Port*. Versão 3.0. São Paulo: Universidade de São Paulo, NILC, 2020.

POJANAPUNYA, P.; TODD, R. W. Log-Likelihood and Odds Ratio: Keyness Statistics for Different Purposes of Keyword Analysis. *Corpus Linguistics and Linguistic Theory*, [*S.l.*], v. 14, n. 1, p. 133-167, 2018. DOI: https://doi.org/10.1515/cllt-2015-0030

REBECHI, R. R. Fraseologias bilíngues português-inglês da culinária brasileira: estudo direcionado pelo corpus. *In*: RIBEIRO, E. S.; TABOSA, L. M. A.; SILVA, N. R. B. (org.). *Tradução em três vertentes*: teoria e prática, intersemiose e Linguística de Corpus. Mossoró: Queima-Bucha, 2017. p. 201-220.

ROSSELLI, D. The Language of Biomedical Sciences. *The Lancet*, Londres, v. 387, n. 10029, p. 1720-1721, 2016. DOI: https://doi.org/10.1016/S0140-6736(16)30259-8

SANTIAGO, M. S. *Redes de palavras-chave para artigos de divulgação científica da Medicina*: uma proposta à luz da Terminologia. 2007. 151f. Dissertação (Mestrado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, 2007.

SCARTON, C. E.; ALMEIDA, D. M.; ALUÍSIO, S. M. Análise da inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *In*: BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY, 7., 2009, São Carlos. *Proceedings* […]. São Carlos: WikiCFP, v. 1, 2009. p. 1-10.

STEWART, D. Conventionality, Creativity and Translated Text: The Implications of Electronic Corpora in Translation. *In*: OLOHAN, M. (org.). *Intercultural Faultlines*. Manchester, Northampton: St. Jerome Publishing, 2000. p. 73-91. DOI: https://doi.org/10.4324/9781315759951-6

TAGNIN, S. E. O. *O jeito que a gente diz*: combinações consagradas em inglês e português. Barueri: Disal, 2013.

U.S. CENSUS BUREAU. *Current Population Survey*. Annual Social and Economic Supplement. Suitland: Census Bureau, 2017.

U.S. NATIONAL LIBRARY OF MEDICINE. *MedlinePlus*. Bethesda, U.S.: Department of Health and Human Services, 2020.

ZAMBONI, L. M. S. *Cientistas, jornalistas e a divulgação científica*: subjetividade e heterogeneidade no discurso de divulgação científica. Campinas: Autores Associados, 2001.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 999-1032, 2021

CC BY

Propriedades linguísticas da redação do Enem: uma análise computacional

Linguistic properties of Enem essays: a computational analysis

Roberlei Alves Bertucci
Universidade Tecnológica Federal do Paraná (UTFPR), Curitiba, Paraná / Brasil
bertucci@utfpr.edu.br
https://orcid.org/0000-0003-4014-5610

Resumo: Este texto descreve algumas propriedades linguísticas recorrentes em textos nota 1000 do Enem. Parte-se do princípio de que o gênero *redação do Enem* tem características próprias, como a exposição aliada à argumentação, e que elementos como repertório e amplo uso de conectivos e modalizadores podem contribuir para a caracterização do texto. Nesta pesquisa, levando em conta o rigoroso processo de avaliação pelo qual passam, consideram-se as redações nota 1000 como exemplares prototípicos do gênero, ou seja, cumprem todos os requisitos exigidos pela banca. O *corpus* é constituído de 95 redações que alcançaram a nota máxima nos anos de 2014, 2018 e 2019, analisadas por meio do *software* Tropes, uma ferramenta computacional de análise lexical que verifica as recorrências de categorias e repertório. Os resultados mostraram que tais redações apresentam uma estrutura muito próxima àquela que se verifica na literatura sobre o tema, em especial o predomínio da estrutura impessoal (terceira pessoa), a amplitude do repertório (universo de referência), a variedade e alta frequência de conectivos e modalizadores, além da recorrência de verbos estativos, como *ser*. Com isso, conclui-se tanto que a ferramenta contribui para a descrição do gênero em questão quanto que os resultados fomentam um debate em torno da padronização da estrutura do texto do Enem.

Palavras-chave: gênero textual; ferramenta computacional; redação do Enem; propriedades linguísticas.

Abstract: This paper aims to describe some linguistic properties found in Enem essays graded to 1000 points, the maximal score. It considers that this genre has its own characteristics, such as the relation between exposition and argumentative types, a wide repertoire, as well as a large use of connectives and modal elements, all of them

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.999-1032

contribute to characterize this genre. In this research, by considering the strict evaluation process, it has been considered that these essays graded to 1000 points prototypically represent the genre, once that they meet the criteria required by the evaluation panel. The *corpus* consists of 95 essays that got 1000 points in 2014, 2018 and 2019 exams. It has been analyzed by means of Tropes, a computational tool that verifies the frequency of lexical items group them in categories and repertoire. The results show that such essays follow what the literature proposed by the genre, especially, the impersonality (third person only), the vast repertoire extension (universe of reference), the range of connectives and modal elements, besides the recurrence of stative verbs, such as *ser* ('to be'). Consequently, one concludes that both the tool contributes to the genre description, and the results put forward the debate around the standardization of the Enem essay structure.

Keywords: textual genre; computational tool; Enem essay; linguistic properties.

Recebido em 01 de outubro de 2020

Aceito em 25 de novembro de 2020

1 Introdução

Trabalhar com textos argumentativos requer do pesquisador a capacidade de refletir sobre estratégias linguísticas que amarram as ideias num todo capaz de defender uma tese. Em outras palavras, exige que ele seja capaz de descrever os elementos textuais que apontam para uma determinada conclusão argumentativa. Assim, a argumentação tem sido colocada em foco nos estudos em linguagem, esta encarada como meio de produção de interações sociais. Levando em conta esse caráter da linguagem, Koch (2011, p. 15) defende o trabalho com o tema por considerar que a argumentação caracteriza a ação linguística "dotada de intencionalidade". Além disso, para a autora, é pelo trabalho com a argumentação que se pode desenvolver "a capacidade de refletir, de maneira crítica, sobre o mundo", por meio da linguagem (KOCH, 2011, p. 15). De maneira mais completa, Koch (2011, p. 17, grifos da autora) assinala que

> a interação social por intermédio da língua caracteriza-se, fundamentalmente, pela argumentatividade. Como ser dotado de razão e vontade, o homem, constantemente, avalia, julga, critica, isto é, forma juízos de valor. Por outro lado, por meio do

> discurso – ação verbal dotada de intencionalidade – tenta influir sobre o comportamento do outro ou fazer com que compartilhe determinadas de suas [*sic*] opiniões. É por esta razão que se pode afirmar que o **ato de argumentar**, isto é, de orientar o discurso no sentido de determinadas conclusões, constitui o ato linguístico fundamental, pois a **todo e qualquer discurso subjaz uma ideologia**, na acepção mais ampla do termo.

Como se vê, Koch (2011) deixa claro seu posicionamento sobre o papel central da argumentação nas ações linguísticas cotidianas, já que os juízos que formamos, em geral, são compartilhados nessas interações, concretizadas nos discursos produzidos.

Nesse sentido, as redações do Exame Nacional do Ensino Médio (Enem) são uma fonte interessante de estudo de argumentação. Realizada por mais de 4 milhões de candidatos anualmente, a redação do Enem é um texto que merece uma atenção por parte de professores e pesquisadores, uma vez que o sucesso na prova pode contribuir para a continuidade dos estudos de muitos estudantes a partir da política educacional estabelecida (OLIVEIRA; CABRAL, 2017). Nessa prova, além da exposição, discussão e análise de um tema, há uma necessidade de articular informações de áreas diversas com coesão e coerência. Além disso, conforme se lê nos enunciados das propostas – e na Cartilha do Participante (BRASIL, 2019) –, o candidato precisa apontar uma proposta de intervenção na situação-problema indicada no tema. Por essa razão, selecionamos para este trabalho alguns exemplares de textos nota 1000 com a intenção de descrever as características linguísticas mais recorrentes nesse gênero.

Nesta pesquisa, consideramos os textos nota 1000 como prototípicos do gênero *redação do Enem*, porque, para a atribuição de tal nota, é necessária uma avaliação criteriosa da banca a partir do que exige o Exame. Sabe-se que a avaliação ocorre às cegas por no mínimo dois profissionais capacitados para a atividade (BRASIL, 2019), mas, no caso das redações que alcançaram a nota máxima, a atribuição é realizada por outros profissionais que precisam ratificar a avaliação. Essa prototipicidade tem levado alguns estudiosos a debaterem sobre um excesso de padronização em detrimento de questões autorais (LIMA, 2019; PAIVA, 2020). Ainda que indiretamente, acreditamos que o presente trabalho pode contribuir para o debate.

Apesar dessa discussão em torno da padronização, dados revelam que o número de textos que recebem a nota máxima vem caindo ao longo dos processos, como se vê no Quadro 1.

QUADRO 1 – Número de redações nota 1000 (últimos 7 anos do Enem)

Ano do Enem	Textos com nota 1000
2013	481
2014	250
2015	104
2016	77
2017	53
2018	55
2019	53

Fonte – Adaptado de Campos (2020).

Isso parece indicar que, apesar de a proposta de produção ser basicamente a mesma ao longo dos anos, os textos parecem estar se distanciando do padrão exigido.[1] A pergunta que primeiramente motivou o olhar para essas redações prototípicas foi: quais seriam as características linguísticas recorrentes da redação do Enem, a partir do que é exigido pelo Exame? Entendemos que a melhor resposta passaria por uma análise dos textos que obtiveram nota máxima no certame.

Várias pesquisas têm sido feitas sobre a redação do Enem, numa tentativa de se compreender melhor o gênero. Do ponto de vista argumentativo, muitos trabalhos relacionam, por exemplo, os aspectos da Nova Retórica com os textos do Enem, como se vê em Magalhães (2013), Barros e Albuquerque (2015), Azevedo (2015), Oliveira, F. (2016), Oliveira e Cabral (2017), Pinheiro e Cortez (2017), Bertucci (no prelo), entre outros.

[1] Apesar de não ser foco desta pesquisa, é igualmente válido considerar a subjetividade da banca de avaliação, conforme sugerem Cançado *et al*. (2020), ou o fato de a mesma banca ter se tornado mais rígida quanto à atribuição da nota máxima. Mesmo assim, esperava-se que mais candidatos tirassem a nota máxima ao longo do tempo, em razão de a preparação para o texto ser beneficiada pela recorrência do gênero no Exame.

Do ponto de vista da estrutura, Bertucci *et al*. (2020) descrevem a presença de recursos anafóricos nesses textos, mostrando como os encapsulamentos contribuem para a coesão e a direção argumentativa do texto. Um pouco mais próximo àquilo que realizamos aqui, Paiva (2020) abordou um conjunto de 32 redações nota 1000 entre os anos de 2014 e 2017. A autora atentou-se para marcas estruturais no texto, as quais são entendidas desta forma: "[...] o modo como o texto é composto e dividido, ou seja, toda a sua teia de escolhas, desencadeamentos e composição (...)". Mais especificamente, olhou

> para a presença ou não de título; quantidade de parágrafos; quantidade de períodos por parágrafos; sequência das ideias, evidenciando de que modo a tese, os argumentos e a proposta de intervenção foram organizados; presença ou não de discurso direto ou indireto; e a forma como os recursos linguísticos expressam esses discursos. (PAIVA, 2020, p. 54.)

Ainda que a autora tenha comentado sobre elementos linguísticos recorrentes, como o uso de verbos modais, o texto não enfoca aspectos quantitativos, como pretendemos. Assim, podemos dizer que as pesquisas mais próximas àquela que aqui desenvolvemos são de Pereira (2018) e Silva (2018). Com um *corpus* bem menor (apenas 16 redações), Pereira (2018) analisou elementos estruturais presentes em redações nota 1000 referentes ao ano de 2016, por meio do aplicativo Tropes. Paralelamente, Silva (2018) descreveu diferenças importantes em relação aos textos nota 1000 (verificados por Pereira, 2018) e aqueles considerados medianos, como o menor número de modalizadores ou conectivos. Sua análise, por meio do mesmo software, foi feita a partir de um *corpus* de textos de um curso preparatório.[2]

Nesse contexto, este trabalho tem como objetivo central descrever as características linguísticas recorrentes da redação do Enem. Para isso, selecionamos 95 textos de anos distintos, conforme sua disponibilidade: 20 (das 250) redações nota 1000 do processo de 2014, todas elas

[2] Seguindo a correta orientação de um parecerista anônimo, não faremos uma comparação em relação ao texto de Silva (2018), uma vez que o objetivo não é comparar nossa descrição com a análise da autora, cujo *corpus* é diverso daquele analisado na presente pesquisa (ela analisa textos considerados mediados, com notas abaixo de 700). Por isso, deixamos ao leitor interessado no tema a referência para leitura adicional.

divulgadas pela imprensa; 31 (dos 55) do Exame de 2018 e 44 (das 53) de 2019, divulgadas por Felpi (2019, 2020, respectivamente). Após essa coleta, utilizamos o Tropes, uma ferramenta computacional que faz uma caracterização linguística de textos, baseada especialmente nas ocorrências lexicais, já utilizada em outros trabalhos para análise de textos de gêneros distintos (ARAÚJO, 2017; BERTUCCI, 2020)

Entendemos que a utilização de ferramentas digitais que favoreçam a coleta e a análise de dados é essencial para a pesquisa empírica. É válido dizer que a Linguística de *Corpus* vem aproveitando bem o surgimento de novas tecnologias com o intuito de gravar, armazenar e analisar dados (RASO; MELO, 2012; SARDINHA, 2000). Nesse âmbito, pode-se dizer que houve avanços na descrição de gêneros textuais por meio de recursos computacionais, algo que tem contribuído muito para o entendimento das línguas, tal como se vê em Finatto (2017) – e nos artigos que fazem parte daquele número temático. Apesar disso, muito há ainda por ser feito, sendo importante destacar que o uso de ferramentas como o Tropes em ambientes de ensino e/ou pesquisa pode ajudar.

Por isso, o presente trabalho pode contribuir para pelo menos duas frentes distintas e complementares: a primeira, os estudos de *corpora*, na tentativa de preencher as lacunas desse tipo de trabalho em especial com gêneros escolares e com o uso de ferramentas digitais no Brasil; a segunda, o ensino, uma vez que, bem delineadas as propriedades do gênero em questão, os profissionais da área podem se beneficiar de métodos diversos de análise desse tipo de texto – ou ao menos, como indicamos antes, para um debate acerca da padronização do gênero.

2 Análise de gêneros e tecnologia

2.1 Linguagem e tecnologia

A discussão sobre a importância da tecnologia em nossas vidas vem sendo posta em diferentes perspectivas de análise. Isso certamente ocorre porque, como nos aponta Cupani (2016, p. 9) "a tecnologia nos afeta e desafia qualquer que seja a nossa atividade". E afeta porque a ela estão relacionadas diferentes práticas cotidianas, inclusive as linguísticas. Na mesma direção, o mesmo autor aponta que as técnicas são "procedimentos sujeito a regras" e, por isso, passíveis de serem aprendidas; são, ao mesmo tempo, "manifestações da capacidade humana

de fazer coisas" (CUPANI, 2016, p. 15). Consequentemente, os produtos (artefatos) decorrem do saber-fazer do homem, que modifica o seu meio também artisticamente, porque está carregado de criatividade para concepção do processo e para chegar ao resultado planejado. Dessa forma, tal procedimento está imbuído de regras relacionadas ao saber-fazer.

Nesse ponto, em que se concebe a tecnologia em relação direta com o planejamento, destaca-se o papel da linguagem. Vieira Pinto (2005) considera a linguagem um requisito essencial para planejamento por seu caráter simbólico: foi por isso que ela permitiu avançar no momento presente e estabelecer possibilidades de mudanças da realidade para além das necessidades imediatas. Para o autor, não fosse isso, o desenvolvimento de tecnologias associadas ao conhecimento, seria mínimo ou até mesmo impossível. Assim, a linguagem também preenche o requisito essencial para o desenvolvimento da própria tecnologia. Para Cassirer (1994, p. 47-48, grifos do autor) o caráter simbólico é a chave para se entender o próprio ser humano:

> Entre o sistema receptor e o sistema de reação, que se encontram em todas as espécies animais, encontramos no homem um terceiro elo, que podemos descrever como o *sistema simbólico*; esta nova aquisição transforma toda a vida humana. Em confronto com os outros animais, o homem não vive apenas numa realidade mais vasta; vive, por assim dizer, numa nova dimensão da realidade.

Como se lê acima, o sistema simbólico, ofereceu ao homem condições claras de diferenciação dos outros animais, mas, sobretudo, de superar dimensões mais próximas e desejar, planejar e até mesmo realizar outras, trazendo-as à dimensão real. A linguagem, portanto, exerceu um papel decisivo, por ser modeladora dessa capacidade.

Pode-se dizer que é graças a isso que as inovações modificam o ambiente humano. Barton e Lee (2015) consideram que toda mudança tecnológica causa alguma mudança na vida social; ou seja, à medida que novas técnicas, processos e produtos aparecem, a vida das pessoas se modifica. Como a linguagem é parte essencial no processo, não podemos desconsiderar a relação plena entre texto/linguagem e tecnologias: linguagem modificando práticas sociais (e tecnológicas) e tecnologias modificando/influenciando práticas linguísticas.

Com isso, é possível se falar sempre em uma valoração da tecnologia, seus processos e produtos. É por isso que Cupani (2016,

p. 12) argumenta que "aquilo que denominamos tecnologia se apresenta, pois, como uma realidade polifacetada: não apenas em forma de objetos e conjuntos de objetos, mas também como sistemas, como processos, como modo de proceder, como uma certa localidade". Em outras palavras, se o mundo está cada vez mais permeado de tecnologia e a linguagem é a grande mediadora da sociedade com essa realidade, lança-se o desafio aos estudiosos da área de identificarem objetos, processos e modos de proceder na coleta e análise de dados.

Nesse contexto, a escola é um espaço propício para a discussão da ciência e da tecnologia, pois tenta apresentar aos alunos de que forma o homem chegou aos conhecimentos que transformaram criativamente seu meio ao longo do tempo. Isso também se aplica aos gêneros textuais: espera-se que as aulas de línguas levem o aluno a compreender e a aplicar um "saber-fazer" específico para cada situação de interação linguística por meio de gêneros. Quanto ao objeto desta pesquisa, há inclusive uma série de aulas, vídeos e livros que apresentam supostas técnicas aos alunos para conquistar a (quase impossível!) nota 1000 (ALVES; BESSA, 2018).

Por isso, é inconcebível se pensar em uma discussão de linguagem e tecnologia sem se associar também a capacidade do homem de pensar sobre o próprio fazer da linguagem, ou seja, sem desenvolver elementos metalinguísticos específicos para ampliação das suas esferas de saber (AUROUX, 2014). Dessa forma, os gêneros podem ser compreendidos como parte fundamental dos processos de ensino-aprendizagem de língua, os instrumentos pelos quais as trocas linguísticas se realizam.

Acrescenta-se a essa reflexão que a possibilidade de analisar os gêneros por meio de ferramentas tecnológicas é uma das maneiras de contribuir para o entendimento das línguas naturais e dos processos relacionados à escrita. Isso nos leva a concluir que pesquisas de *corpus* são essenciais para se atingir esse objetivo.

Quando tratamos das pesquisas em *corpora*, podemos observar que o uso de tecnologias é um fato; afinal, ferramentas que contribuem para coleta, armazenamento e análise de *corpus* de grande extensão têm beneficiado as pesquisas, em especial com relação a textos escritos no Brasil (RASO; MELO, 2012). Ao mesmo tempo, o uso da internet tem facilitado também o acesso a gêneros que poderiam ficar restritos a seu ambiente de produção, como é o caso da redação do Enem. Assim, ações como de Felpi (2019, 2020), que compilou um bom número das redações nota 1000 do Enem de 2018 e 2019, contribuem para que

pesquisadores e estudantes analisem padrões de textos caracterizados de forma homogênea em uma situação, como é o caso da redação nota 1000.

O presente trabalho segue a tendência de análise com *corpora* de textos escritos e para estudos relacionados ao gênero, com a possibilidade de estudos relativos a *corpora* com propósitos específicos. Com isso, acreditamos que a utilização de textos produzidos em ambientes formais e monitorados (como o exame do Enem) pode fornecer elementos importantes para o estudo da língua, seja para a análise do discurso, o estudo de texto/gêneros, seja para a variação linguística, entre outros.

Nesse contexto, Finatto (2017) destaca a importância do apoio computacional quer para a formação de *corpora* de diferentes modos e fontes, quer para sua análise textual, em especial para a descrição de gêneros textuais/discursivos. Concordamos com a autora, pois tal apoio, especialmente para os linguistas, mostra que a tecnologia é um meio que ajuda a explicar diferentes fenômenos da língua natural. Por isso, este artigo quer colaborar para enfatizar o apoio dado de recursos de tecnologia na área da linguagem, em especial dos gêneros textuais: a partir da coleta de dados, vamos analisar o *corpus* por meio do Tropes. Nesse sentido, destaca-se que a contribuição dos estudos com *corpora* é oferecer aos pesquisadores a possibilidade de fazer análise de dados reais e, para tal, a aplicação de recursos de coleta, armazenamento e análise de dados é importante (SARDINHA, 2000).

Sobre a extensão de nosso *corpus*, nos baseamos na ideia de representatividade exposta em Sardinha (2000). Para o autor, o tamanho do *corpus* é menos importante que os dados, em si, que devem ser agrupados conforme critérios do pesquisador. Nesse aspecto, o que é de fato necessário é sua representatividade linguística em algum contexto. Logo, não se deve definir um tamanho mínimo para um *corpus*, ainda que seja fundamental o fato de ser entendido como suficientemente representativo. Além disso, embora não tenha extensão pré-definida, a representatividade é, logicamente, melhor caracterizada quanto maior for o *corpus*. Quando específicos (e não abertos), os *corpora* são de acesso exclusivo do pesquisador, que os produzem para uma finalidade pré-determinada, não sendo disponíveis para outros pesquisadores e, consequentemente, acabam não sendo "verificáveis, o que compromete a pesquisa em termos de sua replicabilidade e generabilidade" (SARDINHA 2000, p. 348). Acreditamos que os 95 textos do nosso *corpus* sejam suficientemente representativos para uma compreensão

geral do gênero em questão, inclusive porque, só do ano de 2019, há 44 textos no *corpus*, dos 53 possíveis.

De qualquer modo, se as ferramentas tecnológicas são meio para análise dos gêneros, poderíamos pensar, igualmente, que os próprios gêneros (em especial os escritos) são também uma tecnologia. Se tomarmos a perspectiva da língua humana como "instrumento" de interação e ação do homem no mundo, podemos atribuir-lhe um caráter tecnológico: como instrumento, a língua é um meio para se alcançar um determinado objetivo, exigindo dos falantes um amplo conhecimento de suas capacidades para cumprir com suas funções.

Nesse contexto, assumimos que os gêneros escritos são um tipo de tecnologia, não só por exigirem a escrita (também uma tecnologia, como sustenta Auroux (2014)), como pelo caráter de planificação (próprio da tecnologia, como defende Vieira Pinto (2005)), uma vez que pesquisas têm demonstrado que o planejamento é essencial para construção de textos escritos (CABRAL, 2013; OLIVEIRA, 2018).

A respeito disso, Campos e Ribeiro (2013, p. 25) afirmam que a natureza do gênero é ser

> [...] um instrumento semiótico que é apropriado por sujeitos, os quais, uma vez instrumentalizados, podem adquirir novos conhecimentos e saberes. Em outras palavras, instrumentalizado em alguns gêneros, o sujeito é capaz de alçar voo de forma independente porque desenvolve habilidades que lhe permitem usar a língua nas mais variadas práticas sociais, que são mediadas pelos gêneros.

Para tal instrumentalização, a escola tem um papel essencial, já que deveria possibilitar aos estudantes refletirem sobre o fazer linguístico em situações diversas, especialmente aquelas de maior complexidade, que exigem graus de letramento mais aprofundados (como a escrita). Por isso, Wachowicz (2010) defende que cabe à escola promover práticas em que os gêneros sejam os instrumentos por meio dos quais os alunos se relacionem com a cultura letrada. Por exemplo, o gênero artigo de opinião de um jornal é o instrumento socialmente elaborado que o jornal utiliza para estabelecer sua relação com o leitor; caberia, assim, à escola instrumentalizar o aluno para a ação sobre esse gênero, seja na leitura, seja na produção, seja na interação com o jornal que produziu esse artigo. Da mesma maneira que compreendemos a técnica da fabricação

de artefatos como mediação das ações humanas no mundo do trabalho, e tecnologia como um pensar sobre essas ações, tornando-as mais eficientes, compreendemos os gêneros como mediadores da produção de conteúdo linguístico, que permitem aos homens agir no mundo letrado. Acreditamos, portanto, que a relação entre linguagem e tecnologia se dá na medida em que ambas se utilizam do aprimoramento de instrumentos para se concretizarem como práticas sociais.

2.2 O gênero *redação do Enem*

A redação do Enem é chamada de texto dissertativo-argumentativo, uma mescla de tipos textuais. Cantarin *et al*. (2017, p. 83) consideram que dissertar é "[...] fazer uma reflexão teórica sobre um assunto", o que pressupõe que o autor lance mão de seu repertório sociocultural e utilize recursos de diferentes naturezas na exposição. Já a argumentação, numa perspectiva retórica, por exemplo (FIORIN, 2017; PERELMAN; OLBRECHTS-TYTECA, 2014), está ligada à capacidade de convencer um determinado auditório, a partir de relações possíveis, prováveis e plausíveis (e não por demonstrações lógicas, por exemplo). O candidato usa o seu conhecimento para expor uma questão; mas, ao articular as informações escolhidas, acaba por construir um argumento em torno de um ponto de vista.

Em textos argumentativos, como o que analisamos nesta pesquisa, o planejamento pode contribuir para o projeto do autor na defesa de um ponto de vista, já que um produtor "instrumentalizado" no gênero seria capaz de traçar um caminho argumentativo claro em seu texto. Coroa (2017, p. 52) sugere que a escrita desse tipo de texto segue um mapeamento por parte do autor: "a cada marca ou pista, ele avança, recua ou reorienta seu caminho." No caso dos textos nota 1000, considerando a avaliação criteriosa, o planejamento se revela na completa coerência do texto com relação ao ponto de vista defendido, aos argumentos que o sustentam e às propostas de intervenção sugeridas pelo candidato. Assim é essa materialização que se realiza no gênero e que propomos pode ser de alguma forma descrita pelo Tropes.

Ao tratar dos tipos textuais, Garcez (2017a, p. 45) ressalta que eles "se definem pela natureza linguística intrínseca de sua composição. As escolhas lexicais, os aspectos sintáticos, o emprego de tempos verbais, as relações lógicas estabelecidas definem o tipo textual." Por isso,

levantamos a hipótese de que haja recorrência de elementos linguísticos nos textos prototípicos do Enem (nota 1000).

De início, um ponto importante a ser reforçado é o fato de o texto do Enem ser considerado um gênero misto (dissertativo-argumentativo), o que significa dizer que a argumentação, ali, exige uma exposição do tema – como se disse, é aqui que o repertório vai contribuir para a dissertação. Coroa (2017) e Cantarin *et al*. (2017) mostram que a dissertação, entendida como exposição ou aprofundamento do tema, é essencial no texto do Enem. Nesse sentido, a redação também requer uma descrição de diferentes aspectos e relativamente aprofundada do tema, de tal forma que os argumentos se sustentem nesse contexto. De modo bem específico sobre a exposição, Cantarin *et al.* (2017, p. 85) afirmam que

> linguisticamente, há elementos básicos para a exposição, como a utilização de verbos no presente do indicativo (“exercem”, “surge”, “é”), especialmente os de ligação (“ser” e “estar”, por exemplo), e marcadores de qualidade e classe, como adjetivos (“diferentes”, “grande”, “infantil”) e substantivos (“processo”, “poder”, “hábitos”), entre outros elementos.

O uso de verbos estativos (“ser” e “estar”, por exemplo) ou de adjetivos, por exemplo, são importantes nesse gênero e devem indicar que o texto está dissertando sobre o tema, e ao mesmo tempo, argumentando, pela escolha de tais itens, conforme sugerem os mesmos autores.

Passando para a argumentação, é preciso dizer que o tipo argumentativo é bastante complexo. Coroa (2017), ao comparar a possibilidade de caracterização dos tipos, sustenta que as características de organização textual são de difícil identificação para o tipo argumentativo, embora recorra,

> com muita frequência às relações lógicas para demonstrar a verdade daquilo que diz, como as de causa e consequência e as de condição. Comumente tais relações são expressas por conectivos de finalidade, de causa, de justificativa, como em “por causa das múltiplas realidades” ou em “a fim de que, não importa quais sejam os resultados de nossos estudos, nosso compromisso”. Mas também a seleção lexical, como “complexa”, “múltiplas realidades”, “elementos significativos”, deve estar a serviço do objetivo. (COROA, 2017, p. 61.)

Como se vê, a autora indica aspectos importantes para a caracterização do tipo argumentativo no texto do Enem, em especial o cuidado com as relações lógicas e a seleção lexical, que pode indicar o repertório do candidato. Para Peixoto (2017, p. 163), são exatamente os conectivos os elementos responsáveis pela introdução da argumentação no texto: "argumentos com diferentes forças e orientações discursivas são colocados na cena textual pelos chamados operadores e conectivos argumentativos."

Outro fator importante é a análise com relação aos modalizadores. Autores como Koch (2011) defendem que esses elementos são formas importantes de mostrar um posicionamento do autor em relação ao conteúdo que expressa. Por isso, além de verbos modais, como "dever" e "poder", por exemplo, é importante que se verifique a ocorrência de outros marcadores de modalidade no texto, em especial advérbios. Neste trabalho, procuramos descrever a frequência de modalizadores presentes no *corpus*, imaginando que muitos deles serão aqueles adequados para caracterizar a argumentação no texto (uma vez que é prototípico do gênero).

Um último ponto de destaque vem da (im)pessoalidade no gênero sob análise. Garcez (2017b, p. 276, grifos nossos) afirma que "o Enem, tradicionalmente, exige um texto dissertativo-argumentativo, ou seja, aquele que apresenta ideias e informações e é impessoal (**preferencialmente deve-se evitar a primeira pessoa**) (...). Um bom modelo desse gênero é o editorial jornalístico". Essa "preferência" pela impessoalidade não significa, no entanto, que o candidato que usar primeira pessoa possa ser prejudicado pela escolha de registro, como afirmam Sandoval *et al*. (2017). Ainda assim, o uso de pronomes de alguma pessoa que não a terceira é um fator a ser investigado também neste trabalho.

Tudo isso vai ao encontro do que se lê em uma cartilha do participante do Enem (BRASIL, 2019), a qual detalha que o candidato precisa defender um ponto de vista com argumentos consistentes e elencar uma proposta de intervenção para o problema discutido. Concluímos que o conjunto de textos aqui analisados cumpriu de modo exemplar essas indicações, já que foram avaliados com a nota máxima pela banca de correção.

O mesmo material detalha aos candidatos as cinco competências avaliadas no processo, indicando inclusive uma matriz de referência de notas, em que se apontam os motivos pelos quais os avaliadores atribuem

notas de 0 a 200 para cada competência. O Quadro 2 a seguir resume os objetivos de cada competência e a descrição do modo como se avalia um texto com nota máxima em cada uma delas.

QUADRO 2 – Competências avaliadas na redação do Enem

Competência	Objetivo	Para nota máxima (200 pontos)
1	Demonstrar domínio da modalidade escrita formal da língua portuguesa.	Demonstra excelente domínio da modalidade escrita formal da língua portuguesa e de escolha de registro. Desvios gramaticais ou de convenções da escrita serão aceitos somente como excepcionalidade e quando não caracterizarem reincidência.
2	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.	Desenvolve o tema por meio de argumentação consistente, a partir de um repertório sociocultural produtivo, e apresenta excelente domínio do texto dissertativo-argumentativo.
3	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.	Apresenta informações, fatos e opiniões relacionados ao tema proposto, de forma consistente e organizada, configurando autoria, em defesa de um ponto de vista.
4	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.	Articula bem as partes do texto e apresenta repertório diversificado de recursos coesivos
5	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.	Elabora muito bem proposta de intervenção, detalhada, relacionada ao tema e articulada à discussão desenvolvida no texto.

Fonte: Adaptado de Brasil (2019)

Novamente, podemos sustentar que o conjunto em análise na presente pesquisa é composto de textos que cumpriram com excelência todos os requisitos indicados pelas competências. Sendo exemplares prototípicos, levantamos a hipótese de que devem ser formados por itens recorrentes, os quais podem estar intimamente ligados às características do gênero textual em questão. É o que pretendemos verificar a partir da próxima seção, em que apresentamos a metodologia e análise dos dados. Vamos descrever a ferramenta Tropes, os textos selecionados e os resultados obtidos.

3 Metodologia e análise dos dados

3.1 Dados, ferramenta e hipóteses

Para esta pesquisa, selecionamos um conjunto de 95 textos nota 1000 do Enem, sendo divididos desta forma: 20 textos do exame de 2014, disponíveis na imprensa;[3] 31 redações de 2018, disponibilizados por Felpi (2019); e 44 redações de 2019, disponibilizadas por Felpi (2020). O objetivo era verificar os elementos recorrentes no conjunto que pudessem descrever o gênero redação do Enem.

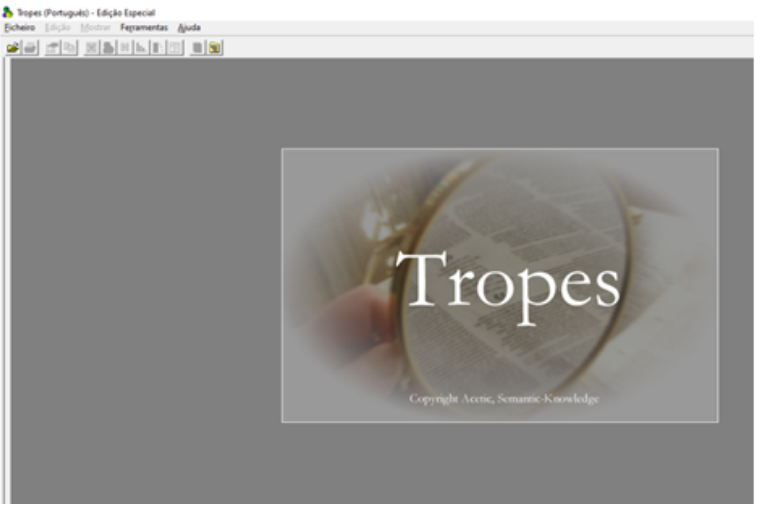
A análise foi realizada por meio do Tropes, uma ferramenta de análise textual em versão gratuita. Criado nos anos 90, pela Semantic-Knowledge (Acetic), com sede em Paris, o Tropes é uma das ferramentas que a empresa desenvolveu para análise textual. Por meio da incorporação de conhecimentos advindos da área de Processamento de Linguagem Natural, o objetivo do *software* era servir para estudos de áreas diversas, como de Sistemas de Informação, Sociologia e Linguística, por exemplo. Araújo (2017, p. 300) observa que

> [este] *software* destaca-se pelo processamento semântico de textos em línguas naturais. Para descrever as características dos enunciados em análise, o Tropes 7.2.3 vale-se de critérios linguísticos pré-programados e os associa às estruturas linguísticas encontradas nos textos processados.

[3] Entre os exemplos está o portal G1: Redações nota 1000: Disponível em: <http://g1.globo.com/educacao/enem/2015/noticia/2015/05/leia-redacoes-do-enem-que-tiraram-nota-maxima-no-exame-de-2014.html > Acesso em: 17 jul. 2018.

Por suas características de análise lexical, o programa faz um processamento do qual decorrem informações como o "estilo textual", o contexto básico relativo ao texto, denominado de "Universo de referências" (sendo uma espécie de mapeamento do repertório mobilizado no texto), e dados quantitativos referentes a classes lexicais. A Figura 1 apresenta a tela de abertura da ferramenta.

FIGURA 1 – Tela inicial do Tropes



Fonte: Tropes.

Neste trabalho, vamos enfatizar dois pontos: o universo de referência, pela relação com o repertório e o tema; e a frequência de classes lexicais, que tem relação com os elementos linguísticos que constituem o gênero. No primeiro caso, o software relaciona palavras de mesmo campo semântico, criando conjuntos de referência. Para a referência de "sociedade", por exemplo, ele agrupa expressões como *família*, *qualidade de vida* e *população* entre muitas outras. Nesse caso, acreditamos que o aplicativo poderá indicar o universo mais próximo do tema proposto bem como o repertório utilizado nas redações.[4]

[4] O aplicativo oferece subdivisões do Universo de referência; aqui, apresentaremos a mais geral, intitulada "Universo de referência 1".

No segundo caso, verificaremos a ocorrência das classes. Entre os conectivos, o Tropes divide a classe em *condição, causa, finalidade (escopo), adição, disjunção, oposição, comparação, tempo* e *lugar*; os modalizadores são divididos em *tempo, lugar, modo, afirmação, negação, dúvida* e *intensidade*; os verbos em *factivos, estativos, declarativos* e *performativos*; os adjetivos entre *subjetivos, objetivos* e *numéricos*; os pronomes entre *primeira, segunda, terceira pessoa* e o clítico *se*.

Considerando as funcionalidades do *software*[5] e as características da redação do Enem, apresentadas anteriormente, levantamos as seguintes hipóteses a serem testadas no conjunto de dados por meio do Tropes:

1. Espera-se um volume grande de repertório (universo de referência) mobilizado pelos autores, o que favorece a exposição e, articulado, leva à argumentação – Cantarin *et al*. (2017).
2. Haverá uma variedade nos tipos de conectivos, já que as relações presentes no texto argumentativo os exigem – Coroa (2017) e Peixoto (2017).
3. Haverá variedade de modalizadores, com enfoque para aquelas classes que contribuem mais para processos argumentativos, como os de modo e intensidade – Koch (2011).
4. Espera-se uma alta frequência de estativos pelo caráter de exposição (*ser* e *estar*, por exemplo) e modalização (*poder* e *dever*, por exemplo) – Cantarin *et al*. (2017).
5. Haverá alta frequência de adjetivos pela exigência da dissertação/descrição – Cantarin *et al.* (2017).
6. Espera-se um predomínio da terceira pessoa – Garcez (2017b).

A seguir, apresentamos os resultados e verificamos se as hipóteses foram corroboradas ou não pelos dados analisados no *corpus* com as 95 redações nota 1000.

[5] Para mais informações sobre o aplicativo, sugerimos Araújo (2017) e Bertucci (2020).

3.2 Resultados e análise

3.2.1 Universo de referência/repertório

Os primeiros dados que colhemos do conjunto de redações diz respeito ao repertório utilizado nos textos. A especificação "Universo de referência" no aplicativo apresentou os seguintes resultados, conforme os Quadros 3 a 5.

QUADRO 3 – Principais classes no universo de referência (2014)

Referência/Classe	Ocorrências	Exemplos
vida_humana	475	idade, crianças, propaganda
conceitos_gerais	377	padrões, necessidade, produto
negócios	147	consumo, publicidade, mercado
sociedade	78	pais, filhos, família

Fonte: Elaboração própria com dados do Tropes.

QUADRO 4 – Principais classes no universo de referência (2018)

Referência/Classe	Ocorrências	Exemplos
vida_humana	761	liberdade, indivíduo, manipulação
conceitos_gerais	691	comportamento, conteúdo, perfil
comunicação_e_mídia	170	linguagem, internet, comunicação
educação_e_ensino	79	educação, ensino, escola

Fonte: Elaboração própria com dados do Tropes.

QUADRO 5 – Principais classes no universo de referência (2019)

Referência/Classe	Ocorrências	Exemplos
conceitos_gerais	1278	inclusão, desenvolvimento, acesso
vida_humana	757	poder, conhecimento, construção
arte_e_cultura	557	cinema, cultura, arte
sociedade	217	classes sociais, sociedade, população

Fonte: Elaboração própria com dados do Tropes.

Como se percebe nos dados, os itens 'vida_humana" e "conceitos_gerais" são os mais frequentes nos três conjuntos de redações nota 1000 analisados. Isso parece indicar que os textos tratam de um conjunto de elementos relativos à sociedade (o que pode ser reforçado pela presença do item "sociedade" em dois conjuntos), cuja abordagem é de caráter teórico/conceitual. É uma sugestão de leitura dos dados e que, na prática, pode ser confirmada por uma análise qualitativa adicional. Outro ponto de destaque é que o universo de referência pode ser considerado amplo nos três casos: verificamos que, em média, 35 conjuntos de relações no universo de referência foram mobilizados nas redações. Sem dúvida, o ideal seria uma análise individual de cada redação para uma média por redação, tal como fez Pereira (2018). Analisando redações nota 1000, ela observou um total de 33 classes no universo de referência em seu *corpus* – o que se aproxima do nosso número. Além disso, verificou que, em média, cada candidato mobilizou cerca de 16,5 universos distintos. No nosso caso, quando se observam os dados de forma mais específica, verificam-se distintas categorias tais como *filosofia, história, economia, ciência*, entre muitas outras, o que revela um amplo repertório mobilizado. Esse é um ponto que parece ser decisivo para a categorização das redações nota 1000.

Com isso, acreditamos que a Hipótese 1, relativa ao repertório, foi parcialmente corroborada pelos dados – mas apenas uma análise individual das redações poderia dar uma noção mais precisa. A variedade observada favorece, portanto, a exposição, um dos elementos essenciais para a construção do gênero. Vale acrescentar ainda que a frequência observada nos Quadros 3 a 5 sugere uma relação próxima dos textos com a temática proposta, ou seja, os candidatos se mantiveram no escopo do tema solicitado: em 2014, o tema foi "A publicidade infantil em questão no Brasil", por isso *negócios* e *sociedade* aparecem também entre os itens mais frequentes do universo de referência; em 2018, o tema solicitado foi "A manipulação do comportamento do usuário pelo controle de dados da internet", por isso *comunicação_e_mídia* foi bastante recorrente; finalmente, o tema de 2019 foi "Democratização do acesso ao cinema no Brasil", o que explica a recorrência do item *arte_e_cultura*.

Finalmente, cabe a observação de que a mobilização de um amplo repertório parece estar ligada, primeiramente, às articulações expositivas no texto. Para Coroa (2017), o tipo dissertativo privilegia a exposição de uma ideia, enquanto, enquanto o argumentativo, o modo como se

defende um ponto de vista. Logo, no tipo dissertativo-argumentativo, essa exposição é mobilizada em favor de uma argumentação. Nesse sentido, o repertório estaria ligado à noção de legitimação e produtividade que aparecem nos materiais referentes à avaliação da redação do Enem (INEP, 2020). Assim, como os dados apontam para uma recorrência do fato, é possível que os candidatos treinem essa mobilização de diversos universos de referência com o objetivo de dar credibilidade à sua redação dissertativo-argumentativa.

3.2.2 Conectivos

A Tabela 1 resume o resultado obtido com a análise das 95 redações nota 1000.

TABELA 1 – Ocorrência de conectivos

	2014		2018		2019		Geral	
Conectivo	Total	%	Total	%	Total	%	Total	%
Adição	190	59	294	58,3	361	51,5	845	55,4
Causa	8	2,5	33	6,6	49	6,9	90	6
Comparação	22	6,9	49	9,7	97	13,9	168	11
Condição	30	9,4	42	8,4	78	11,1	150	9,9
Disjunção	15	4,6	16	3,2	23	3,3	54	3,5
Escopo (fim)	15	4,6	31	6,2	21	3	67	4,4
Lugar	2	0,6	0	0	8	1,2	10	0,6
Oposição	27	8,4	30	6	52	7,4	109	7,1
Tempo	12	3,7	8	1,6	12	1,8	32	2,1
TOTAL	321	100	503	100	701	100	1525	100

Fonte: Elaboração própria com dados do Tropes.

Os dados revelam que a categoria de adição é a mais frequente, com mais de 55% das ocorrências entre os conectivos. Isso é fruto, certamente, da alta frequência da conjunção em português, em especial pela possibilidade de ser usada em posição interior ao sintagma e não

apenas para ligação de orações, períodos ou parágrafos.[6] Outras conjunções que merecem destaque são as de comparação (11%), de condição (9,9%), oposição (7,1%) e causa (6%), exemplificadas respectivamente por *como*, *mesmo que, no entanto* e *uma vez que*. Todas elas são essenciais para a construção do teor argumentativo, para o direcionamento que o autor pretende dar no texto. Além disso, cabe ressaltar ainda dois pontos: primeiro, que todas as categorias foram contempladas na análise *corpus*, sendo as de lugar e tempo as menos frequentes – o que é de se esperar em textos de teor argumentativo. De fato, conforme observa Coroa (2017), espera-se mais elementos de referência temporal na narração e de referência espacial na descrição. Em segundo lugar, que, dado o total de 1.525 conectivos, tem-se uma média de 16 categorias distintas para cada uma das 95 redações analisadas. Pereira (2018) havia observado um total de 308 conectivos nas 16 redações analisadas, o que dá uma média próxima de 19 por texto. Nesse caso, também se observou uma alta frequência do conectivo de adição (acima de 54%) e porcentagens similares àquelas encontradas aqui para as outras categorias.

Isso parece indicar que a quantidade de conectivos no texto é tão relevante quanto sua variedade. E, de fato, isso está de acordo os materiais sobre a Competência 4 para avaliadores (GARCEZ; CORRÊA, 2017; INEP, 2020), que orientam sobre a diversidade e sobre a frequência dos elementos de coesão (como os conectivos) na redação do Enem. Em geral, tais materiais sugerem que uma redação do Enem deve apresentar pelo menos dois conectivos interparágrafos e pelo menos um intraparágrafos, o que já poderia contabilizaria um número de 6 conectivos (no mínimo) em uma redação de 4 parágrafos. Mas, quando se olha para a especificação das competências (INEP, 2020), percebe-se uma orientação para atribuição de nota máxima apenas aos casos de presença expressiva, com adequação e sem repetições, de elementos coesivos, o que deve explicar a recorrência nos textos nota 1000. Neste caso, reforçamos, como se trata de textos avaliados com a nota máxima, pode-se dizer que a quantidade e a qualidade do emprego foram importantes para a argumentação, sobretudo pela necessidade do uso dos conectivos para a

[6] Uma rápida pesquisa no site "Corpus do Português", com a entrada "e", devolveu mais de 34 milhões de ocorrências no termo (num *corpus* de mais de 1 bilhão de palavras). A título de comparação, a conjunção "mas" apareceu em mais de 3 milhões de casos, o que corresponde a apenas 10% de "e".

construção da argumentação. Portanto, podemos concluir que os dados corroboram a Hipótese 2, sobre a variedade com relação aos tipos de conectivos exigidos no gênero.

3.2.3 Modalizadores

A Tabela 2 resume o resultado sobre os modalizadores.

TABELA 2 – Ocorrência de modalizadores

	2014		2018		2019		Geral	
Modalizador	Total	%	Total	%	Total	%	Total	%
Afirmação	08	3	12	3,2	19	3	39	3
Dúvida	02	0,7	2	0,5	0	0	04	0,3
Intensidade	91	33,7	98	26,2	238	35	427	32,3
Lugar	07	2,6	6	1,5	17	2,6	30	2,3
Modo	65	24.1	142	38,1	204	30	411	31.1
Negação	52	19,3	44	12,4	107	15,5	203	15,4
Tempo	45	16,6	66	18,1	96	14	207	15,6
TOTAL	270	100	370	100	681	100	1321	100

Fonte: Elaboração própria com dados do Tropes.

Os dados mostram um fenômeno interessante: modalizadores de modo (como advérbios terminados em *-mente*) e intensidade (como *sobretudo*) predominam no tipo de texto em análise, revelando um posicionamento do autor em relação ao fato apresentado. Ainda que não esperada, a alta frequência da categoria de negação é explicada pelo amplo uso de *não* nos textos, outra palavra frequente na língua (quase 10 milhões indicadas no site "Corpus do Português"). Igualmente, a frequência da categoria de tempo parece estar relacionada com o uso de advérbios como *atualmente* e *hoje*. No entanto, uma pesquisa mais específica, com cada uma das redações, seria necessária para uma explicação mais minuciosa.

Com relação aos dados totais, temos uma média de 13,9 modalizadores para cada uma das 95 redações analisadas, um número

próximo dos 15,7 encontrados por Pereira (2018). Novamente, percebe-se que a maior frequência aliada a uma variedade de modalizadores parece constituir o gênero do ponto de vista da sua prototipicidade. Assim, tais dados revelam que os modalizadores, além de indicarem a capacidade do candidato em articular elementos linguísticos importantes para a construção da sua argumentação, se constituem como indícios de autoria importantes no texto dissertativo-argumentativo: é por meio do uso de elementos como esses que o autor modaliza as vozes/discursos que apresenta no texto (COSTA; GUEDES, 2017). Portanto, os dados apresentados corroboram para a Hipótese 3 a respeito da variedade de modalizadores, com enfoque para modo e intensidade.

De modo adicional, podemos dizer que o emprego dos modalizadores estaria ligado mais ao caráter argumentativo que dissertativo, na diferenciação dos tipos proposta por Coroa (2017). São esses elementos que vão contribuir para que o texto tenha um caráter de defesa de ponto de vista sobre o tema, a partir das informações e fatos mobilizados pelos candidatos. Ou seja, são eles que contribuem para a formação das estratégias argumentativas que sustentarão o texto.

3.2.4 Verbos

Na Tabela 3 são apresentados os dados relativos à frequência dos tipos de verbos.

TABELA 3 – Frequência de verbos

	2014		2018		2019		Geral	
Verbo	Total	%	Total	%	Total	%	Total	%
Factivo	727	69,2	1225	73	1642	68,9	3594	70,3
Estativo	302	28,7	418	24,9	689	28,9	1409	27,5
Outros	23	2,1	36	2,2	52	2,2	111	2,2
TOTAL	1.052	100	1679	100	2383	100	5114	100

Fonte: Elaboração Própria com dados do Tropes.

Os números relativos ao *corpus* em análise demonstram que os verbos classificados como *factivos* (verbos de ação) são predominantes. Ainda que o índice de 27,5% de estativos seja significativo, o resultado

geral parece ir na contramão da Hipótese 4, que afirmava sobre a alta frequência dos estativos. Pereira (2018), que também encontrou uma alta frequência de factivos em seu *corpus*, recorre a Araújo e Cunha (2009, p. 33-34) para explicar a relação dessa classe de verbos com a argumentação: os autores consideram que essa categoria "é aquela em que um sujeito intencional e animado realiza uma ação e essa ação afeta (ou efetua) um objeto paciente". Para Pereira (2018), essa ação com intencionalidade que afeta um paciente é percebida no encadeamento lógico (de ações/reflexões que se sucedem), exigido na redação do Enem, e que culmina com a proposta de intervenção. Vê-se como adequada, portanto, a presença de verbos factivos. Apesar disso, decidimos verificar no *corpus* quais verbos eram os mais frequentes desse total de 5114. O resultado se vê na Tabela 4.

TABELA 4 – Verbos mais frequentes

	Geral	
Verbo	**Total**	**% (do total de 5114)**
ser	792	15,5
dever	142	2,8
poder	141	2,8
ter	111	2,2
tornar	92	1,8
TOTAL	1278	25

Fonte: Elaboração Própria com dados do Tropes.

À primeira vista, os dados parecem contradizer a alta frequência de verbos factivos observada nos dados anteriores, uma vez que os cinco verbos mais frequentes no *corpus* são todos estativos: *ser, dever, poder, ter* e *tornar*, sendo que apenas o primeiro agrega mais de 15% do total de 5114 verbos etiquetados pelo Tropes. No entanto, a verdade é que tais números atestam a hipótese levantada a respeito da alta frequência de estativos esperada no *corpus*, sobretudo pelo papel que esses verbos têm na língua. Por isso, três outros pontos podem ser relacionados aos dados. O primeiro sobre o uso de modalizadores, como *poder* e *dever*.

Assim como se viu antes, a modalização é um processo esperado em textos argumentativos e, além dos advérbios, o uso de verbos modais é uma marca para essa ação linguística. Tal fato explica que esses verbos tenham uma frequência alta nos textos. O segundo ponto é com relação ao uso de *ser*. Bertucci (no prelo) mostra que o uso do argumento de definição é muito recorrente em textos do Enem e, para a construção desse tipo de argumento, um dos verbos mais recorrentes é exatamente o verbo *ser*. Finalmente, Pereira (2018) também observou índices gerais muito parecidos com o que apresentamos aqui, seja das classes verbais, seja dos exemplos em si. Assim, os números que apresentamos vão ao encontro da proposta de que as redações do Enem se caracterizam pela alta frequência de estativos específicos e de uma predominância de factivos.

Nesse sentido, podemos sugerir que, apesar de percentualmente inferiores, os verbos estativos são essenciais tanto para o caráter expositivo do texto quanto para o teor argumentativo pretendido pelo autor, em especial no uso de verbos modais e na formação de estruturas opinativas, como aquelas formadas com adjetivos (PADILHA, em preparação). Logo, no contexto de ensino, por exemplo, não se deve concluir que um tipo de verbo deve predominar: o que se espera é um emprego adequado (e não quantificado), ainda que os dados revelem pontos interessantes sobre a redação do Enem como gênero textual.

3.2.5 Adjetivos

O resultado da análise sobre adjetivos pode ser visto na Tabela 5.

TABELA 5 – Frequência de adjetivos

	2014		2018		2019		Geral	
Adjetivo	**Total**	**%**	**Total**	**%**	**Total**	**%**	**Total**	**%**
Objetivo	301	48,6	418	46,3	670	45,7	1389	46,5
Subjetivo	300	48,5	400	44,7	644	43,8	1344	45
Numérico	18	2,9	81	9	155	10,6	254	8,5
TOTAL	619	100	899	100	1469	100	2987	100

Fonte: Elaboração própria com dados do Tropes.

No que diz respeito à subcategorização de adjetivos, os dados revelam um relativo equilíbrio entre adjetivos classificados como objetivos (*infantil, social, cultural* entre outros) e aqueles marcados como subjetivos (*necessário, grande, importante* entre outros). Observa-se que o número de adjetivos utilizados (2987) é bastante alto, comparado ao de verbos (5114), ainda mais se se considerar que o verbo é um elemento essencial numa oração, mas o adjetivo não. O que se percebe é que os adjetivos têm um papel relevante em textos como do Enem, uma vez que são um meio também de revelar o posicionamento do autor. Isso fica claro, sobretudo, com os subjetivos, que são articulados em torno de um ponto de vista em relação a um fato apresentado, tal como observa Padilha (em preparação). Para a autora, a forte presença dos subjetivos, em especial aqueles de caráter modal e gradual, contribui para a noção argumentativa no texto, muito mais do que para uma simples exposição do tema. Voltando à Hipótese 5, podemos constatar que a média de adjetivos para cada uma das 95 redações é de 31,4, o que daria cerca de um adjetivo por linha, na redação do Enem. Isso mostra a importância da categoria nesse gênero e corrobora a referida hipótese.

3.2.6 Pronomes

Finalmente, a última categoria sob análise é dos pronomes. O Tropes destacou uma única referência à primeira pessoa do singular e observamos nos textos que se referia ao trecho seguinte, presente na redação 3 da cartilha referente ao Exame de 2018.

> Em segundo lugar, o ser humano perde a sua capacidade de escolha. Conforme o conceito de "Mortificação **do Eu**", do sociólogo Erving Goffman, é possível entender o que ocorre na internet que induz o indivíduo a ter um comportamento alienado. (FELPI, 2019, p. 11, grifos nossos)

Nesse caso, o uso do pronome é feito como citação/menção e não como uso por parte do candidato. Vale destacar ainda que o aplicativo destacou 276 ocorrências do clítico *-se*, que tem a função de impessoalizar a ação verbal (ou seja, é também um item de terceira pessoa). Pereira (2018), com o já referido *corpus* de 16 redações nota 1000, fez observações semelhantes às nossas. Tudo isso, então, corrobora a Hipótese 6, sobre a predominância da terceira pessoa em textos do Enem.

Nesse ponto, é importante destacar que os materiais relativos à avaliação da redação do Enem não excluem a possibilidade de escrita do texto em primeira pessoa (BRASIL, 2019; GARCÊZ; CORREA, 2017; INEP, 2020; SANDOVAL *et al*., 2017). No entanto, o que se vê nos dados é que os textos avaliados com a nota máxima não apresentam pronomes dessa categoria, o que é um indício de que a avaliação se dá num nível bastante restrito de impessoalização do texto, valorizando a escrita em terceira pessoa. Não nos parece justo dizer que a argumentação exija, em si, uma impessoalização; pelo contrário: pessoalizar pode ser uma forma de persuasão bastante válida, como mostram Freitas e Marra (2016) a respeito de anúncios publicitários. Este ponto, aliás como todos os já comentados aqui, merece mais espaço de estudo em português brasileiro, em especial no que tange à caracterização dos gêneros e, por consequência, suas possibilidades de reflexão e ensino.

4 Considerações finais

Os resultados da presente pesquisa nos permitem defender que ferramentas computacionais contribuem na descrição de gêneros textuais/discursivos, porque são capazes de analisar volumes de dados maiores do que aqueles que se faz manualmente (impensáveis para um pesquisador ou professor). Nesse sentido, pensamos ter justificado a importância da relação entre linguagem, tecnologia e um estudo de *corpus* com gêneros escolares, uma vez que pode ter impacto no modo como se analisam e se produzem os textos em ambientes de letramento.

Isso não significa, obviamente, que o ensino de produção textual deva ser comparado a uma receita pronta, uma vez que a ação linguística é extremamente complexa. No entanto, compreender as recorrências de um gênero pode contribuir para ampliar a capacidade de reflexão de professores e estudantes a respeito do fazer linguístico, especialmente de textos que se mostram bastante formatados, como a redação do Enem. Por isso, tentamos discutir os dados à luz dos materiais oficiais referentes à prova, além de outras pesquisas, como uma forma de vincular os dados ao modo como se interpreta (ou avalia) o texto escrito no Exame.

A partir da proposta de utilizarmos o Tropes para análise dos dados, podemos concluir que ele apresentou um panorama geral relativo às propriedades linguísticas das redações nota 1000 que são coerentes

com aquilo que se vê na literatura sobre esse tipo de texto. Com isso, concordando com Araújo (2017) e Bertucci (2020), pensamos que o Tropes contribui para a identificação de elementos linguísticos que caracterizam um gênero. Deixamos em aberto, no entanto, pesquisas que possam analisar suas limitações e possíveis inconsistências de análise. Entre as que identificamos, há a marcação dos artigos *o* e *a* como pronomes oblíquos de terceira pessoa, o que, no entanto, não prejudica as conclusões gerais apresentadas, em especial porque o número de pronomes de primeira e segunda pessoa foi nulo no *corpus*.

No presente trabalho, também não tratamos de questões a respeito do debate sobre a padronização do gênero, nem de como essa descrição realizada pode contribuir para o ensino de produção textual, ou seja, para a discussão do "saber-fazer" que a tecnologia metalinguística abarca na escola, uma vez que não era nossa intenção fornecer um manual de ensino do gênero, mas analisar um conjunto significativo de textos prototípicos do Enem e mostrar que a ferramenta Tropes pode contribuir para a descrição de textos. Ao leitor interessado nesse tópico, no entanto, a divulgação do material de apoio para avaliadores, divulgado em 2017 pelo Inep (GARCEZ; CORRÊA, 2017), e dos materiais para formação de avaliadores (INEP, 2020), recentemente disponibilizados, podem auxiliar a entender as minúcias desse gênero textual/discursivo exigido no Enem, com ênfase na produção e avaliação dos textos.

Por fim, entendemos que será uma grande contribuição, tanto para a área acadêmica, quanto para a pedagógica, que profissionais se debrucem sobre questões como as apresentadas aqui e desenvolvam estratégias que contribuam para um ensino da língua que, realmente, ajudem os alunos a fazer escolhas linguísticas relevantes na construção de seus textos. Reiteramos que isso não significa haver uma receita, mas pode ser um olhar diferente para a constituição do gênero. Igualmente relevantes são pesquisas com aplicativos como o aqui usado e com *corpora* similares, a fim de se contribuir para o entendimento de gêneros e para sua instrumentalização.

Referências

ALVES, W. M.; BESSA, J. C. R. Orientações para escrita da redação do Enem em vídeos do Youtube. *Hipertextus*, Recife, v. 19, n. 1, p. 1-23, 2018. Disponível em: https://periodicos.ufpe.br/revistas/hipertextus/article/view/247974/36463. Acesso em: 1 out. 2020.

ARAÚJO, L. S. de. O gênero entrevista radiofônica em comunidades hispânicas: um aporte da Análise Textual Automática. *Domínios de Linguagem*, Uberlândia, v. 11, n. 2, p. 289-312, 2017. Disponível em: https://doi.org/10.14393/DL29-v11n2a2017-2. Acesso em: 1 out. 2020.

ARAÚJO, F. de C.; CUNHA, M. A. F. da. A estrutura argumental dos verbos de ação. *PublICa*, Natal, v. 3, n. 1, p. 28-35, 2009. Disponível em: https://periodicos.ufrn.br/publica/article/view/106. Acesso em: 1 out 2020.

AUROUX, S. *A revolução tecnológica da gramatização*. 3. ed. Campinas: Editora da UNICAMP, 2014.

AZEVEDO, I. C. M. Organização de textos dissertativo-argumentativos em prosa: o que se percebe em dez anos de realização do Enem? *In:* SILVA, L. R. da; FREITAG, R. M. K. (org.). *Linguagem, interação e sociedade:* diálogos sobre o Enem. João Pessoa: Editora do CCTA, 2015. p. 33-50.

BARROS, M.; ALBUQUERQUE, M. G. As técnicas argumentativas e a construção de sentidos em redações do Enem. *In:* SEMINÁRIO DE ESTUDOS SOBRE DISCURSO E ARGUMENTAÇÃO, 2., 2015, Belo Horizonte. *Anais...* Belo Horizonte: Editora Fale, 2015. p. 545-559.

BARTON, D.; LEE, C. *Linguagem online*: textos e práticas digitais. Trad. Milton Camargo Mota. 1. ed. São Paulo: Parábola Editorial, 2015.

BERTUCCI, R. A. Aplicação de ferramentas para coleta e análise de dados em Linguística. *Diacrítica*, Braga, Portugal, v. 32, n. 3, p. 129-155, 2020. DOI: https://doi.org/10.21814/diacritica.576

BERTUCCI, R. A. Análise do argumento por definição em redações do Enem. *Acta Scientiarum,* Maringá, PR. (no prelo).

BERTUCCI, R. A.; MALHEIROS, A. J.; LOPES, W. de S. Ocorrências de anáforas encapsuladoras em redações do Enem. *Filologia e Linguística Portuguesa*, São Paulo, v. 22, n. 1, p. 81-102, 2020. DOI: 10.11606/issn.2176-9419.v22i1p81-102. Disponível em: https://www.revistas.usp.br/flp/article/view/164142. Acesso em: 23 nov. 2020.

BRASIL. *Redação no Enem 2019*: cartilha do participante. Brasília: Daeb/Inep/MEC, 2019. Disponível em: http://download.inep.gov.br/educacao_basica/enem/downloads/2019/redacao_enem2019_cartilha_participante.pdf. Acesso em: 1. out. 2019.

CABRAL, A. L. T. O conceito de plano de texto. *Linha d'Água*, São Paulo, n. 26, v. 2, p. 241-259, 2013. Disponível em: https://www.revistas.usp.br/linhadagua/article/download/64266/71562/. Acesso em: 1 ago. 2019.

CAMPOS, L. V. Mais de 143 mil participantes tiraram zero na redação do Enem 2019. *Brasil Escola*, *Site Uol*, 17 jan. 2020. Disponível em: https://vestibular.brasilescola.uol.com.br/enem/mais-143-mil-participantes-tiraram-zero-na-redacao-enem-2019/347183.html. Acesso em: 28 set. 2020.

CAMPOS, C. M.; RIBEIRO, J. Gêneros. *In:* COSTA, I. B.; FOLTRAN, M. J. (org.). *A tessitura da escrita.* São Paulo: Contexto, 2013. p. 23-44.

CANÇADO, M.; AMARAL, L.; AMORIN, E.; VELOSO, A.; MELLO, H. Subjetividade em correções de redações: detecção automática através de léxico de operadores de viés linguístico. *Linguamática*, Braga, v. 12, n. 1, p. 63-79. 2020. DOI: https://doi.org/10.21814/lm.12.1.313.

CANTARIN, M.; BERTUCCI, R. A.; ALMEIDA, R. C. de. A análise do texto dissertativo-argumentativo. *In:* GARCEZ, L. H.; CORRÊA, V. R. (org.). *Textos dissertativo-argumentativos*: subsídios para a qualificação de avaliadores, Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017. p. 81-91.

CASSIRER, E. *Ensaio sobre o homem:* introdução a uma filosofia da cultura humana. São Paulo: Martins Fontes, 1994.

COROA, M. L. O texto dissertativo-argumentativo. *In:* GARCEZ, L. H. do C.; CORRÊA, V. R. (org.). *Textos dissertativo-argumentativos:* subsídios para a qualificação de avaliadores. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017. p. 59-71.

CORPUS do Português. Disponível em: https://www.corpusdoportugues.org/. Acesso em: 30 set. 2020.

COSTA, J. de R. O.; GUEDES, M. A. A avaliação dos indícios de autoria. *In:* GARCEZ, L. H. do C.; CORRÊA, V. R. (org.). *Textos dissertativo-argumentativos:* subsídios para qualificação de avaliadores Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017. p. 101-108.

CUPANI, A. *Filosofia da tecnologia:* um convite. Florianópolis: Editora da UFSC, 2016.

FELPI, L. (org.). *Cartilha redação a mil*. 2019. Disponível em: https://www.lucasfelpi.com.br/redamil. Acesso em: 1 out. 2020.

FELPI, L. (org.). *Cartilha redação a mil 2.0*. 2020. Disponível em: https://www.lucasfelpi.com.br/redamil. Acesso em: 1 out. 2020.

FINATTO, M. J. B. Apresentação: descrição dos gêneros textuais/discursivos com apoio computacional. *Domínios de Lingu@gem*, Uberlândia, v. 11, n. 2, p. 282-288. 2017. DOI: https://doi.org/10.14393/DL29-v11n2a2017-1.

FIORIN, J. L. *Argumentação*. *São Paulo: Contexto*, 2017.

FREITAS, H. C.; MARRA, M. N. A. Tipos de argumentos utilizados nos anúncios publicitários das Havaianas. *Domínios de Lingu@gem*, Uberlândia, v. 10, n. 1, p. 304-329, 2016. DOI: https://doi.org/10.14393/DL21-v10n1a2016-16.

GARCEZ, L. H. do C. Gênero e tipo de texto. *In:* GARCEZ, L. H. do C.; CORRÊA, V. R. (org.). *Textos dissertativos-argumentativos:* subsídios para qualificação de avaliadores. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017a. p. 51-58.

GARCEZ, L. H. do C. O ensino de redação. *In:* GARCEZ, L. H. do C.; CORRÊA, V. R. (org.). *Textos dissertativos-argumentativos:* subsídios para qualificação de avaliadores. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017b. p. 275-279.

GARCEZ, L. H. C.; CORRÊA, V. R. (org.). *Textos dissertativo-argumentativos:* subsídios para a qualificação de avaliadores. Brasília: Inep, 2017. Disponível em: http://portal.inep.gov.br/documents/186968/484421/TEXTOS+DISSERTATIVO+ARGUME

NTATIVOS/7809ef0d-5a4a-4c24-9a03-9db15e0bdacf?version=1.0. Acesso em: 28 set. 2020.

INEP disponibiliza material inédito sobre correção da redação do Enem. *Portal do MEC*, Brasília, 2020. Disponível em: http://portal.mec.gov.br/pronatec/oferta-voluntaria/418-noticias/enem-946573306/90611-inep-disponibiliza-material-inedito-sobre-correcao-da-redacao-do-enem. Acesso em: 28 set. 2020.

KOCH, I. G. V. *Argumentação e linguagem*. 13. ed. São Paulo: Cortez, 2011.

LIMA, L. I. *Mapeamento semântico da construção de autoria no Ensino Médio*. 2019. 163f. Tese (Doutorado em Letras) - Setor de Ciências Humanas, Letras e Artes, Universidade Federal do Paraná, 2019. Disponível em: http://www.prppg.ufpr.br/siga/visitante/trabalhoConclusaoWS?idpessoal=29931&idprograma=40001016016P7&anobase=2019&idtc=1533. Acesso em: 01 out. 2020.

MAGALHÃES, M. M. A argumentação em redações escolares. *In:* SILEL – SIMPÓSIO NACIONAL E INTERNACIONAL DE LETRAS E LINGUÍSTICA, 2013, Uberlândia. *Anais*... Uberlândia: EDUFU, 2013, p. 1-13. Disponível em: http://www.ileel.ufu.br/anaisdosilel/wp-content/uploads/2014/04/silel2013_645.pdf. Acesso em: 8. Jun. 2020.

OLIVEIRA, F. C. C. de. *Um estudo sobre a caracterização do gênero redação do ENEM*. 2016. 166f. Tese (Doutorado em Linguística) – Centro de Humanidades, Universidade Federal do Ceará, 2016.

OLIVEIRA, W. R. de. *Planejamento de escrita em meio digital e analógico*. 2018. 175f. Dissertação (Mestrado em Estudos de Linguagens) - Departamento Acadêmico de Linguagem e Comunicação, Universidade Tecnológica Federal do Paraná, 2018. Disponível em: http://repositorio.utfpr.edu.br/jspui/handle/1/3319. Acesso em: 1 out. 2020.

OLIVEIRA, M. I. S.; CABRAL, A. L. T. Política de Língua Portuguesa para o ensino de Redação no nível médio da educação brasileira: o texto argumentativo dos PCN's à redação do Enem. *Verbum*, São Paulo, v. 6, n. 2, p. 6-30. 2017. Disponível em: https://revistas.pucsp.br/verbum/article/view/30274. Acesso em: 1 out. 2020.

PADILHA, M. P. *O papel modalizador dos adjetivos em redações do Enem*. Dissertação (Mestrado em Estudos de Linguagens) – Departamento Acadêmico de Linguagem e Comunicação, Universidade Tecnológica Federal do Paraná. Em preparação.

PAIVA, R. I. de S. *Redações nota mil no ENEM*: um estudo analítico da massificação se sua estrutura e conteúdo. 2020. 132f. Dissertação (Mestrado em Estudos de Linguagem) – Centro de Ciências Humanas, Letras e Artes, Universidade Federal do Rio Grande do Norte, 2020. Disponível em: https://repositorio.ufrn.br/handle/123456789/29929. Acesso em: 1 out. 2020.

PEIXOTO, J. dos S. A avaliação do emprego de operadores e conectivos argumentativos. *In:* GARCEZ, L. H. do C.; CORRÊA, V. R. (org.). *Textos dissertativos-argumentativos:* subsídios para qualificação de avaliadores. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017. p. 163-171.

PEREIRA, K. A. P. L. *A contribuição de um analisador automático para a caracterização de gêneros textuais*. 2018. 83f. Trabalho de Conclusão de Curso (Graduação em Letras) - Universidade Tecnológica Federal do Paraná, 2018. Disponível em: http://repositorio.roca.utfpr.edu.br/jspui/handle/1/11580. Acesso em: 1 out. 2020.

PERELMAN, C.; OLBRECHTS-TYTECA, L. *Tratado da argumentação*: a nova retórica. 3. ed. São Paulo: Martins Fontes, 2014.

PINHEIRO, C.; CORTEZ, J. Teorias da argumentação na prova de redação do ENEM. *Linguagem & Ensino*, Pelotas, v. 20, n. 1, p. 61-80. 2017. Disponível em: https://doi.org/10.15210/rle.v20i1.15215. Acesso em: 1 out. 2020.

RASO, T.; MELLO, H. (org.). C-*ORAL-BRASIL I*: corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.

SANDOVAL, A. N.; ALCÂNTARA, S. S.; ZANDOMÊNICO, S. C. M. de R. Notas sobre a avaliação de desvios de registro. *In:* GARCEZ, L. H. do C.; CORRÊA, V. R. (org.). *Textos dissertativos-argumentativos:* subsídios para qualificação de avaliadores. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2017. p. 29-35.

SARDINHA, T. B. Corpus Linguistics: History and problematization. *DELTA*, São Paulo, v. 2, n. 16, p. 323–367, 2000. Disponível em: http://dx.doi.org/10.1590/S0102-44502000000200005. Acesso em: 26 abr. 2018.

SILVA, N. S. *Análise textual mediada por ferramenta computacional*: um estudo sobre redações estilo Enem. 2018. 73f. Trabalho de Conclusão de Curso (Graduação em Letras) - Universidade Tecnológica Federal do Paraná, 2018.

TROPES. *Semantic-Knowledge*: text analysis, qualitative analysis & text mining. [s.d.]. Disponível em: https://www.semantic-knowledge.com/tropes.htm. Acesso em: 30 jun. 2018.

VIEIRA PINTO, Á. *O conceito de tecnologia*. Rio de Janeiro: Contraponto, 2005.

WACHOWICZ, T. C. *Análise linguística nos gêneros textuais*. Curitiba: Intersaberes, 2010.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1033-1058, 2021

Sujeito oculto às claras: uma abordagem descritivo-computacional

Omitted subjects revealed: a quantitative-descriptive approach

Cláudia Freitas
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Rio de Janeiro / Brasil
claudiafreitas@puc-rio.br
https://orcid.org/0000-0001-6807-8558

Elvis de Souza
Pontifícia Universidade Católica do Rio de Janeiro (PUC-Rio), Rio de Janeiro, Rio de Janeiro / Brasil
elvis.desouza99@gmail.com
https://orcid.org/0000-0001-9373-7412

Resumo: Neste trabalho, apresentamos estudos descritivos e computacionais relacionados ao sujeito oculto. Em um primeiro momento, realizamos uma descrição de cunho quantitativo, tomando por base três *corpora* dos gêneros jornalístico, literário e enciclopédico. Especificamente, quantificamos o sujeito oculto em cada um dos *corpora*, e encontramos sujeitos omitidos em 24%, 41% e 46% das orações, respectivamente. Em um segundo momento, por meio de uma estratégia baseada em regras, reconstituímos esses sujeitos e os devolvemos aos *corpora*, com o objetivo de avaliar o quanto a omissão do sujeito é capaz de impactar o aprendizado automático de dependências sintáticas. Os resultados indicam que a reconstituição formal do sujeito pode melhorar a aprendizagem das dependências sintáticas em até 2% quando consideramos a métrica CLAS, evidenciando o papel relevante da modelagem linguística no aprendizado automático.

Palavras-chave: descrição linguística; sujeito oculto; omissão de sujeito; dependências sintáticas; linguística computacional; aprendizado de máquina; linguística de *corpus*.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1033-1058

Abstract: In this paper, we present descriptive and computational studies related to omitted subjects. Firstly, we develop a quantitative descriptive study based on three *corpora*, which consist of journalistic, literary and encyclopedic genres. Specifically, we quantify the omitted subjects in sentences for each of these *corpora;* omitted subjects were found in 24%, 41% and 46% of their sentences, respectively. Secondly, applying rule-based strategies, we reconstitute those subjects and place them back to the *corpora*, with the goal of evaluating how much the omission of subjects can impact the automatic learning of syntactic dependencies. The results indicate that the formal subject reconstitution can enhance the learning of syntactic dependencies in up to 2% according to the CLAS metric, highlighting the relevant role of linguistic modeling in the automatic learning process.

Keywords: linguistic description; omitted subject; syntactic dependencies; computational linguistics; machine learning; corpus linguistics.

Submetido em 07 de outubro de 2020

Aceito em 14 de dezembro de 2020

1 Introdução

A articulação entre o Processamento (automático) de Linguagem Natural (PLN) e os estudos linguísticos vem ganhando força nos últimos anos, alterando um pouco o quadro descrito em 2007 por Karen Sparck-Jones quando constata o distanciamento entre a linguística e a linguística computacional. Muito dessa reaproximação se deve ao trabalho de anotação de *corpora* que, como já apontado em Sampson (2001), é, também, um trabalho de descrição linguística.

Neste artigo, contribuímos com mais um elemento na aproximação entre os dois campos, e o fazemos não pelo viés da anotação, mas partindo de *corpora* já anotados para a descrição de um fenômeno linguístico de grande relevância para uma série de tarefas de PLN em português: o sujeito oculto. Após uma caracterização linguística do fenômeno, voltamo-nos para o PLN, a fim de medir o quanto a ausência de sujeitos em uma oração pode dificultar o processamento sintático automático.

Uma das áreas de atuação do PLN é a extração de informação (EI). Ainda que, tradicionalmente, a extração de informação consista na detecção automática de informações relativa a certos atores pré-definidos, como pessoas, lugares e organizações para indicar, simplificadamente, *quem faz o quê*, tomada em um sentido amplo, várias tarefas do PLN podem

ser agrupadas como variações de um jogo que consiste em vasculhar uma imensa quantidade de textos com o objetivo de encontrar algum tipo de conteúdo (ou informação), e incluímos aqui tanto opiniões quanto dados factuais sobre alvos pré-determinados. A partir daí, com a concatenação dos conteúdos extraídos, podem-se construir novos fatos, ou hipóteses, que serão explorados posteriormente. Em termos metodológicos, a busca pelo conteúdo se dá por meio da identificação de padrões que podem ser codificados como relações (proposições) com um número variável de argumentos e/ou modificadores, como indicado abaixo:

Maria$_{ARG1}$ *estava triste*$_{Pred}$
Maria$_{ARG1}$ *sorriu*$_{Pred}$
Maria$_{ARG1}$ *comprou*$_{Pred}$ *uma bicicleta*$_{ARG2}$
Maria$_{ARG1}$ *emprestou*$_{Pred}$ *dinheiro*$_{ARG2}$ *para o irmão*$_{ARG3}$ mês passado$_{MOD1}$

Ainda que consideremos os diferentes focos de interesse envolvidos na identificação e extração de conteúdos em textos, em boa parte deles a identificação do agente responsável pelas ações/atividades detectadas é crucial. Este *quem*, em geral, se manifesta como o sujeito da oração, e por isso a identificação do sujeito é de extrema relevância para uma boa parcela de tarefas, como identificação de papéis semânticos, identificação e análise de opiniões e sentimentos, extração de citações, além da extração de informação propriamente.

No entanto, na língua portuguesa, diferentemente do inglês (língua que concentra boa parte das pesquisas em PLN), a possibilidade de omissão do sujeito em contextos em que este é facilmente recuperável, o chamado "sujeito oculto", é um dado a mais a ser considerado. Trata-se de um fenômeno cuja resolução, por pessoas, costuma ser trivial, mas que, para as máquinas, impede a construção de relações (ou proposições) adequadas, porque deixa de levar em conta um dos argumentos da proposição. Marcus *et al*. (1993), no contexto do *Penn Treebank*, afirmam que a maneira mais simples de incluir informação sobre a estrutura de predicado-argumento é permitindo que a árvore sintática contenha, de maneira explícita, os elementos nulos.

Além disso, como apontam Hartmann *et al*. (2014) no contexto da língua portuguesa, a inserção de elementos nulos contribuiria para reduzir o problema da escassez de dados, sendo, portanto, um aspecto altamente relevante em abordagens de aprendizagem de máquina.

Ou seja, a presença de sujeitos ocultos traria um desafio adicional ao processamento sintático automático.

Por fim, no que se refere ao processamento humano e ao letramento, é possível que a omissão do sujeito – como, aliás, a omissão de qualquer termo na estrutura frasal, conforme sugerido em Finatto *et al.* (2011) – seja mais um elemento a trazer complexidade aos textos. Deste modo, estratégias capazes de identificá-lo com precisão contribuem para a verificação do grau de dificuldade de um texto. E, do mesmo modo, estratégias capazes de reconstituí-los promoveriam uma simplificação textual.

Do ponto de vista do PLN de língua portuguesa, o quanto de informação perdemos quando não explicitamos o sujeito? Precisamos de algum tratamento especial para resolver a questão, ou se trata de um fenômeno periférico? Tais perguntas poderiam ser facilmente respondidas pelos estudos descritivos, mas não sabemos de nenhum que se dedique ao tema de um ponto de vista quantitativo. É neste espaço que nos colocamos, trazendo, em uma primeira etapa, dados de grandes *corpora* com características textuais distintas: um *corpus* de textos jornalísticos, um *corpus* de textos literários e um *corpus* de textos enciclopédicos, que juntos somam mais de 17 milhões de unidades/[1] 685 mil frases. Os resultados mostram que a ocorrência de sujeitos omitidos não é desprezível, indo de 24% a 46%, conforme o tipo de texto.

Em uma segunda etapa, informados de que a quantidade de orações que não apresentam um sujeito sintático explícito é relevante, realizamos um experimento de PLN no qual reintroduzimos os sujeitos omitidos e treinamos um modelo de aprendizagem de máquina, com o objetivo de verificar o quanto a omissão de sujeito prejudica o processamento automático. Os resultados mostram uma melhora de até 2% no aprendizado quando reconstituímos os sujeitos, sugerindo que um *dataset* cuja construção leva em conta características linguisticamente motivadas é capaz de contribuir no processamento automático no nível sintático.

Ao longo do artigo, apresentamos estudos sobre o sujeito oculto na língua portuguesa a partir de duas perspectivas distintas,

[1] Ao longo do texto, usamos a palavra *unidade* como uma tradução do inglês *token,* isto é, uma unidade mínima de anotação, já tendo sido separadas as contrações de verbos e pronomes, preposições e artigos, etc.

mas complementares: de um ponto de vista linguístico-descritivo, quantificamos o fenômeno do sujeito oculto; de um ponto de vista linguístico-computacional, medimos o quanto este fenômeno dificulta o processamento automático. Em ambos os casos, utilizamos *corpora* e ferramentas públicos.

O presente estudo é conduzido sob o *framework* do projeto *Universal Dependencies* (NIVRE *et al*., 2016), uma abordagem multilíngue para o processamento computacional das línguas. Ao apresentar aqui os dados e procedimentos para o português, esperamos também contribuir para um quadro mais geral de descrição de diferentes línguas no que se refere à omissão do sujeito.

O restante do artigo se organiza da seguinte maneira: na seção 2, apresentamos o fenômeno do sujeito oculto em português, especificando o objeto que nos interessa tratar; na seção 3, indicamos alguns trabalhos sobre o tema na perspectiva dos estudos com *corpus* e do PLN, e na seção 4 detalhamos a metodologia do estudo descritivo, cujos resultados são analisados na seção 5. Por fim, na seção 6 relatamos um pequeno experimento cujo objetivo é medir o impacto da omissão do sujeito na aprendizagem sintática; e na seção 7 tecemos algumas considerações finais.

2 A omissão do sujeito em português

A língua portuguesa licencia a omissão de sujeito nas frases de diferentes maneiras: (i) o sujeito oculto propriamente, ou sujeito elíptico; (ii) o chamado sujeito indeterminado e, ainda, (iii) as orações sem sujeito (veja-se por exemplo LUFT, 2002). Especificamente, nos interessam os casos do primeiro tipo (frases 1, 5, 6 e 7, abaixo), e frases do segundo tipo (frase 2). Deste modo, igualamos o que gramáticas tradicionais distinguem, já que, em ambos os casos, existe um sujeito a ser explicitado, ainda que não haja sujeito formal. Pelo mesmo motivo, não levamos em conta as chamadas orações sem sujeito, como frases com verbos impessoais (frase 3) e com verbos indicativos de fenômenos da natureza (frase 4), já que, nesses casos, não há sujeito a ser explicitado.[2]

[2] Todos os exemplos foram retirados do corpus Bosque, a parte revista do projeto Floresta Sintá(c)tica (AFONSO *et al*., 2002; FREITAS *et al*., 2008).

1. "Eu tentei, o senhor Vance tentou, se for respeitado, urrah!", comentou.
2. Sempre que surge um problema, chamam-na.
3. Há, no ar, uma certa ideia de invasão.
4. Desde há já alguns anos que chove «a eito» na centenária Igreja de Ceide.
5. A empresa norte-americana informou à PF que não tem filiais ou representantes no Brasil.
6. Só depois é que levanto a cabeça para fazer um lançamento», reclama Neto.
7. Herbert Berger, diretor-superintendente da empresa, diz que o Charade «se aproxima do Honda Civic em tamanho e custa bem menos».

Os casos 1 a 4 são aqueles tradicionalmente mencionados quando se aborda o fenômeno da omissão de sujeito em português. Em (5) e (6), temos omissão de sujeito em orações subordinadas; e, em (7), a omissão em uma oração coordenada. Diferentemente das frases (1) e (2), nesses casos o sujeito se encontra nos limites da frase. No entanto, se, do ponto de vista discursivo, (1) e (2) são problemas mais interessantes, pois a explicitação do sujeito envolve a busca por referentes além da frase, do ponto de vista do processamento automático, que privilegia os limites da oração, as frases (5), (6) e (7) podem ser igualmente complexas. Em nosso estudo, fazemos duas contagens: uma para medir os sujeitos de orações principais, e outra para os sujeitos ocultos de orações subordinadas.

Por fim, sabemos também que, em português, é possível que o *se* materialize uma indeterminação do sujeito, como nas frases (8) e (9). No entanto, e diferentemente dos demais casos de indeterminação, com o *se* não é possível identificar quem é o sujeito; não é possível determiná-lo. Como nosso interesse está em apenas distinguir o sujeito oculto, também excluímos esses casos de nossa contagem.

8. Tem que se demonstrar através de contas e de raciocínios que o expurgo significará perda
9. Diga-se, de uma vez e claramente, o que se quer ou o que se quer mais.

3 Trabalhos relacionados

O trabalho de Hartmann *et al*. (2014) é o único que conhecemos que se debruça sobre o sujeito oculto em português no contexto do PLN. Tendo como foco principal a anotação humana de papéis semânticos, os autores exploram a inserção de elementos artificiais para representar sujeitos omitidos. O objetivo da inserção é preencher algumas lacunas na estrutura sintática das frases, a fim de facilitar a atribuição de papéis semânticos e, assim, melhorar o material de treinamento da tarefa. O *corpus* utilizado foi o PropBank-BR (DURAN; ALUÍSIO, 2012), que foi então analisado pelo anotador PALAVRAS (BICK, 2000). A partir da observação do texto anotado pelo PALAVRAS, os autores criaram regras para inserção automática dos elementos nulos.

Assim como no presente trabalho, o processo de criação de regras foi exploratório e incremental. Os elementos nulos foram preenchidos com pronomes pessoais retos levando em conta a forma flexional do verbo ("eu", "nós", e um genérico *SUBJ* (sujeito) para os demais casos). No entanto, ao que parece, o trabalho considerou uma sintaxe linear, sem informação da dependência, o que dificultou sensivelmente a identificação dos sujeitos (ou de sua ausência), haja vista a quantidade de itens intervenientes que podemos encontrar entre o sujeito e o verbo e também a posição do sujeito em português, que pode estar anteposto ao verbo (posição preferencial) ou posposto. A partir da análise de erros de uma amostra de 200 frases, os autores relatam que a estratégia funcionou em cerca de 80% dos casos e que, quando considerada a inserção por tipo do sujeito, os resultados são heterogêneos: a inserção do sujeito oculto é bem-sucedida em 88% quando o verbo corresponde às primeiras pessoas, mas corresponde a apenas 55.8% dos demais casos, o que se deve, sobretudo, a erros anteriores decorrentes da análise automática.

Do ponto de vista descritivo, a situação não é diferente, e isto certamente se deve à ausência de material com as características técnicas necessárias para o estudo: um *corpus* sintaticamente anotado e uma interface de busca em árvores que permita procurar pela ausência, já que não temos a tradição de anotar elementos nulos em *corpora* – o *Penn Treebank* o faz, e trataremos dele a seguir.

Apesar de já dispormos de bons *corpora* em língua portuguesa, nem sempre estão anotados sintaticamente. O vasto material do projeto AC/DC (SANTOS; BICK, 2000) é uma saudável exceção, mas a interface

de busca não permite buscar por dependências sintáticas ou por elementos nulos, que não estão anotados. Já a última versão dos *corpora* do projeto Floresta Sintá(c)tica (FREITAS *et al*., 2008) – que, assim como projeto AC/DC, é criado e mantido pela Linguateca (SANTOS, 2011) e tem seus *corpora* sintaticamente anotados, também, pelo *parser* PALAVRAS – contém, como um elemento "procurável", duas etiquetas atreladas ao verbo que indicam a ausência de sujeito: <*nofsubj*>, que indica a ausência de um sujeito formal (usada para os casos de oração sem sujeito e verbos indicativos de fenômeno da natureza) e <*nosubj*>, para os casos de sujeito oculto. O *corpus* Bosque é a parte revista de todo o material que compõe a Floresta e descobrimos, buscando pelos respectivos *procuráveis* em uma ferramenta de pesquisa em *treebanks* específica do projeto Floresta, o Milhafre,[3] que há 3622 orações sem sujeito explícito e 266 orações sem sujeito formal. Neste trabalho, queremos verificar a distribuição de sujeitos em diferentes gêneros textuais, e sobretudo naqueles em que a presença de um sujeito agente é crucial para tarefas posteriores: obras literárias, na qual a distribuição de falas entre personagens é um objeto de pesquisa (ELSON; MCKEOWN, 2010; RUANO SAN SEGUNDO, 2016) e textos enciclopédicos – especificamente, uma enciclopédia biográfica sobre a história do Brasil, na qual se pode extrair informações sobre personagens da história política brasileira (HIGUCHI *et al*., 2019).

No que se refere à anotação de *corpus*, ainda que o fenômeno do sujeito oculto não aconteça na língua inglesa, o *tagset* do *corpus Penn Treebank* utiliza a etiqueta PRO para os casos onde existe um sujeito não especificado ou não realizado. Tais casos referem-se especificamente a elementos nulos no imperativo (10); e construções de alçamento e controle, como em (11-13).

10. Go away!
11. $John_i$ seems to PRO_i like Mary
12. $John_i$ promised Mary PRO_i to *write* the book
13. John persuaded $Mary_i$ PRO_i to *write* the book

O *tagset* sintático do projeto *Universal Dependencies* possui uma etiqueta – *xcomp* – para os casos em que

[3] Milhafre. Disponível em: https://www.linguateca.pt/Floresta/milhafre. Acesso em: 8 out. 2020.

> um verbo ou adjetivo é um predicativo ou complemento oracional sem seu próprio sujeito – [isso] não significa que uma oração seja um *xcomp* apenas porque seu sujeito não está omitido. O sujeito deve necessariamente ser herdado de uma posição fixa na oração superior.[4]

Ou seja, a etiqueta *xcomp* inclui (dentre outros fenômenos) casos como 11 e 12 – e apenas eles receberão esta etiqueta.

4 Metodologia

O principal desafio deste trabalho está na metodologia – como encontrar algo sem materialidade, dado que não temos anotação de elementos nulos nos *corpora*. Neste trabalho, usamos a abordagem gramatical do projeto *Universal Dependencies* (UD). O projeto, cujo objetivo é facilitar o desenvolvimento de *parsers* multilíngues e a pesquisa linguística, propõe esquemas de anotação compartilháveis entre línguas para a anotação de classes de palavras, de informação morfológica e sintática. Atualmente, UD conta com mais de 150 florestas (*treebanks*) em 90 línguas diferentes. Como mencionamos na seção 3, o *tagset* de UD conta com uma etiqueta específica para certos casos de omissão de sujeito, e apenas eles. Nos demais casos já mencionados aqui – que correspondem aos exemplos (1-2) e (5-9) – não há uma etiqueta especial. Uma vez que nosso interesse está em medir os casos de omissão de sujeito, não abordaremos, neste momento, as diferenças entre os casos 10-13. Com o auxílio de uma ferramenta desenvolvida especialmente para lidar com *corpora* anotados seguindo o formalismo UD, fomos iterativamente desenvolvendo estratégias e filtros até identificar as frases que nos interessam.

4.1 Os *corpora*

A pesquisa foi realizada em três *corpora* com características distintas. O primeiro deles é o já referido *corpus* Bosque, mas dessa vez em sua versão UD, o Bosque-UD (versão 2.6). Trata-se de um *corpus*

[4] "(…) a verb or an adjective is a predicative or clausal complement without its own subject – [this] does not mean that a clause is an *xcomp* just because its subject is not overt. The subject must be necessarily inherited from a fixed position in the higher clause." *Universal Dependencies guidelines*. Disponível em: https://universaldependencies.org/u/dep/xcomp.html. Acesso em: 8 out. 2020.

composto por textos jornalísticos, com 9.366 frases divididas igualmente entre as variantes do Brasil e de Portugal. Dos três *corpora* utilizados, apenas o Bosque-UD teve sua anotação gramatical revista por linguistas.[5] O fato de a versão original do Bosque conter informações relativas à omissão do sujeito funcionou também como um gabarito para nosso método de procura.

O segundo *corpus* é o DHBB (acrônimo de Dicionário Histórico Biográfico Brasileiro) (HIGUCHI *et al*., 2019). O DHBB é uma enciclopédia sobre a história política brasileira a partir de 1930, criada pelo Centro de Pesquisa e Documentação de História Contemporânea do Brasil, da Fundação Getúlio Vargas (CPDOC/FGV). É um material que interessa especialmente pelo seu conteúdo, configurando-se como uma importante fonte de pesquisa. Desde 2018, o DHBB foi convertido em um *corpus* anotado, integrando o acervo do AC/DC. A versão 6.1 do *corpus* contém 7.700 entradas (ou verbetes), 314 mil frases, cerca de 14 milhões de palavras/16 milhões de unidades e está disponível para consulta e *download* na página da Linguateca.[6]

Por fim, o *corpus* OBras (SANTOS *et al*., 2018). Criado para ser a contraparte brasileira do *corpus* Vercial, o OBras contém obras literárias brasileiras que já estão em domínio público. É um corpus dinâmico e lança novas edições a cada dois meses. Para este trabalho, utilizamos a versão 9.0, que contém 263 obras literárias, 6.8 milhões de palavras/9.7 milhões de unidades. O OBras, assim como o DHBB e o Bosque, integra o AC/DC, o que significa que está ricamente anotado com informação sintática e semântica e disponível para buscas pela internet.

Embora todo o material aqui utilizado já exista em uma versão linguisticamente anotada e disponível para consultas linguísticas e para download,[7] a anotação não contém nenhuma etiqueta relativa à ausência do sujeito, e o AC/DC, embora permita buscas sintáticas, não permite buscas sobre árvores sintáticas. Por isso, reanotamos o material com a ferramenta UDPipe (STRAKA *et al*., 2016).

---

[5] A criação do Bosque-UD está detalhadamente descrita em Rademaker *et al*. (2017).

[6] Disponível em: https://www.linguateca.pt/acesso/desc_dhbb.html. Acesso em: 8 out. 2020.

[7] O OBRAS se encontra disponível em: https://www.linguateca.pt/OBRAS/OBRAS.html, e o DHBB, em https://www.linguateca.pt/acesso/desc_dhbb.html. Acesso em: 8 out. 2020.

4.2 A anotação dos *corpora*

A ferramenta de anotação utilizada foi o UDPipe (STRAKA *et al.,* 2016), uma ferramenta de código aberto que realiza sequencialmente as etapas de tokenização (segmentação do texto em unidades básicas, como palavras e sinais de pontuação), anotação gramatical, lematização e análise de dependências em qualquer *corpus* que esteja no formato CoNLL-U.[8] O UDPipe fornece modelos para quase todos os *treebanks* do projeto UD.[9] O modelo fornecido para o português (versão 2.5) tem índices de acerto (F1) de 96.4%, 95%, 87.2% e 83.1% para os níveis de classes gramaticais (POS), características morfológicas (*feats*), dependência sintática (*unlabeled attachment score* (UAS) e relação de dependência sintática (*labeled atachment score* (LAS)), respectivamente.[10] O *corpus* usado para a geração do modelo de língua portuguesa é o corpus Bosque-UD, já mencionado na seção anterior. Por ser um corpus linguisticamente revisto, o Bosque-UD (assim como o Bosque original) se presta também ao treino e à avaliação de modelos sintáticos, e é esta característica que possibilita a realização do experimento em que medimos a dificuldade de realizar análises sintáticas no que se refere à ausência de sujeitos (seção 5). Para tanto, usamos um modelo criado a partir da versão do Bosque-UD disponibilizada no lançamento 2.6[11] e treinado por nós, utilizando os parâmetros padrão do UDPipe.

O Quadro 1 apresenta os três *corpora* conforme a sentenciação (separação do texto em frases) e tokenização feitas pelo UDPipe (e

[8] O formato CoNLL-U é uma adaptação do formato CoNLL-X. As anotações são codificadas em arquivos de texto simples, com um *token* por linha, e colunas (no máximo 10) que codificam diferentes informações linguísticas, como *lema*, *pos* etc. Uma explicação detalhada do formato pode ser encontrada em https://universaldependencies.org/format.html. Acesso em: 8 out. 2020.

[9] Disponível em: http://ufal.mff.cuni.cz/udpipe/models#universal_dependencies_25_models_publications. Acesso em: 8 out. 2020.

[10] Especificamente, as medidas UAS e LAS (*unlabeled attachment score* e *labeled attachment score*, respectivamente) se referem aos acertos de encaixe das dependências sintáticas, sendo que, na segunda métrica, além do encaixe (isto é, além de saber qual o núcleo sintático de um determinado elemento), a relação de dependência sintática também deve estar correta.

[11] Disponível em: https://github.com/UniversalDependencies/UD_Portuguese-Bosque. Acesso em: 8 out. 2020.

que podem não corresponder exatamente àquelas feitas no contexto do projeto AC/DC):

QUADRO 1 – Apresentação quantitativa dos corpora, conforme processado pela ferramenta UDPipe

	DHBB	OBRAS	BOSQUE
Tamanho (Mb)	960	480	14
Tokens	16037286	7863261	227825
Or. Principal	480218	353662	9364
Or. Subordinada	341133	316297	6842
Total De Orações	821351	669959	16206

4.3 A busca pelo sujeito oculto

Para identificar o sujeito oculto, utilizamos o Interrogatório, uma ferramenta para busca e revisão de corpora anotados (de SOUZA; FREITAS, 2019). Nela, é possível realizar buscas sobre árvores sintáticas, desde que estejam em arquivos no formato CoNLL-U. Elencamos, a seguir, o procedimento para identificação dos casos de sujeito oculto:

QUADRO 2 – Passos para a identificação de sujeitos ocultos

1. Encontrar sentenças em que não exista um sujeito (simples, oracional ou sujeito de oração passiva) dependente sintaticamente do núcleo da oração principal (*root*)
2. Das frases encontradas, eliminar as seguintes:
   a. Construções em que *root* é o verbo haver impessoal (3ª pessoa do singular)
   b. Construções em que *root* não é verbo[12]
   c. Construções em que *root* é verbo que indica fenômeno da natureza
   d. Construções em que o *se* é um índice de indeterminação do sujeito

Nas etapas acima, o principal desafio está no item (2.d): distinguir, nos casos de ausência de sujeito, aqueles em que o *se* corresponde a um índice de indeterminação (*Diga-se que*...; *Trata-se de*...), dos casos em

[12] Comum em manchetes jornalísticas ou interjeições.

que o *se* é complemento (*Cortou-se*) ou parte de um verbo pronominal (*Formou-se*). Nos primeiros não há o que ser explicitado, nos últimos, há sujeito a ser explicitado. Dois outros fatores também tornam este o filtro mais difícil: a anotação UD não diferencia os casos (14-15) de (16), ambos recebem a etiqueta *expl*.; e a distinção entre esses casos e o caso (17), no qual o *se* recebe a etiqueta de *obj*, ainda é pouco confiável de um ponto de vista automático.

14 Recorde-se aliás que, em Dezembro de 1992, quando a China realizou a maior explosão não nuclear de sempre.

15 Diga-se, de uma vez e claramente, o que se quer ou o que se quer mais.

16 Formou-se em engenharia em 1931.

17 Penteava-se segundo a moda do tempo, mas sem afetação.

A partir da análise manual, fizemos 3 filtros (d1; d2; d3) para identificar as ocorrências que nos interessam, e assim separamos os casos do tipo (14-15), que não contam como omissão do sujeito, dos demais casos:

d1. Casos em que o *se* recebe a anotação morfológica de gênero não especificado
Veja-se que «absoluta prioridade» não é simplesmente uma expressão, mas um princípio constitucional (...)

d2. Casos em que o *se* se associa a um verbo no infinitivo
Tem que se demonstrar através de contas e de raciocínios que o expurgo significará perda.

d3. Casos em que o *se* se associa a um verbo transitivo indireto ou intransitivo[13]
Pense-se em Kingsley Amis, Malcolm Bradbury e Albert Finney.

[13] Notamos que a forma de buscar as construções no *corpus* (isto é, o filtro) não corresponde, necessariamente, a uma análise correta. Neste exemplo, o *se* é exatamente do mesmo tipo do filtro d1, mas, como mencionamos, nem sempre podemos contar com uma análise sintática perfeita no caso do *se*. A forma de buscar indica apenas que, nesse caso, as ocorrências que gostaríamos de encontrar estão anotadas, na grande maioria das vezes, dessa maneira.

De forma complementar, fizemos também uma busca por sujeitos ocultos em orações subordinadas, utilizando as mesmas estratégias listadas no Quadro 2.[14]

5 Resultados e análise

Anterior à apresentação dos resultados, precisamos garantir que aquilo que recuperamos com as buscas e os filtros é o que desejamos. Esta validação é crucial no contexto do processamento automático, sobretudo porque em dois dos *corpora* analisados estamos lidando com o resultado de uma análise sintática que não foi revista. Procedemos a uma verificação manual de uma amostra, a fim de medir o grau de confiança que podemos ter nos resultados, já que apenas o Bosque-UD foi revisto. Foram analisadas até 20 frases por filtro (alguns filtros devolveram menos de 20 ocorrências), considerando cada *corpus*, totalizando 572 frases. A Tabela 1 traz os resultados da análise e a Tabela 2, complementar, indica a quantidade total de casos recuperados por filtro, bem como o quanto esses casos representam considerando o total de orações principais e subordinadas em cada *corpus*. Chamamos de *busca ingênua* a busca por qualquer frase que não tenha um sujeito. A coluna *Aval* (avaliados) da Tabela 1 indica o total de ocorrências de cada filtro; a coluna *Corr* indica a quantidade de ocorrências corretas, isto é, que atendem às especificações da busca/filtro.

A partir da Tabela 1, vemos que os resultados dos filtros variam por *corpus*, e o primeiro dado que chama a atenção é a importância de um material revisto, já que os números do Bosque superam os dos demais *corpora* em todos os cenários, e no que se refere às orações principais, isto é ainda mais evidente. Nos demais *corpora*, os resultados indicam que o que capturamos, quando tentamos encontrar o sujeito omitido, está correto em pouco mais da metade das vezes. Vemos, também, que é mais difícil acertar a procura nas orações subordinadas que nas principais, e isso se deve igualmente a limitações do processamento automático. Quando nos detemos nos resultados de cada um dos filtros, temos uma imagem mais nítida do que recuperamos.

[14] O único filtro não replicado nas orações subordinadas foi o 2b, relacionado às frases sem verbo, uma vez que há uma série de construções que atendem a essa especificação, como adjuntos adverbiais, que nada têm a ver com a omissão do sujeito.

TABELA 1 – Resultados da análise manual, por filtro

	DHBB				**OBras**				**Bosque-UD**			
	Principal		**Subordinada**		**Principal**		**Subordinada**		**Principal**		**Subordinada**	
	Aval.	**Corr**	**Aval.**	**Corr**	**Aval**	**Corr**	**Aval**	**Corr**	**Aval**	**Corr**	**Aval**	**Corr**
Busca ingênua	20	20	20	20	20	20	20	20	20	20	20	20
V. haver	20	20	20	20	20	17	20	20	20	20	20	20
Nominais	20	15	--	--	20	14	--	--	20	20	--	--
V. Natureza	0	0	4	4	20	20	20	20	1	1	2	2
D 1 (SE)	20	7	20	1	20	9	20	2	20	18	5	4
D 2 (SE)	20	5	20	1	20	4	20	1	2	2	20	3
D 3 (SE)	20	1	20	3	20	11	20	6	20	15	20	11
TOTAL	120	68	104	49	140	95	120	69	103	96	87	60
Total de acertos	**56%**		**47%**		**67%**		**57%**		**93%**		**69%**	

TABELA 2 – Distribuição das ocorrências por filtro, por corpus.

	DHBB				**OBras**				**Bosque-UD**			
	Principal		**Subordinada**		**Principal**		**Subordinada**		**Principal**		**Subordinada**	
Busca ingênua	226122	47%	190381	56%	172124	48%	156040	49%	2777	29%	2617	38%
V. Haver	1083	0,2%	861	0,2%	5181	1,5%	4395	1,4%	124	1%	148	2%
Nominais	31599	6,5%	735	0,2%	43326	12%	2766	0,8%	1145	12%	50	0,7%
V. Natureza	0	0%	5	0%	113	0%	147	0%	1	0%	2	0%
D 1 (SE)	59	0%	757	0,2%	215	0%	154	0%	31	0%	5	0%
D 2 (SE)	79	0%	2692	0,8%	229	0%	1609	0%	2	0%	24	0%
D 3 (SE)	26425	5,5%	13189	3,8%	4915	1,4%	6447	2%	57	0,6%	75	1%

Os filtros relativos à busca ingênua, aos verbos que indicam fenômenos da natureza e ao verbo *haver* impessoal trazem os resultados esperados, obtendo 100% de acertos, exceto pelo OBras, que contém 3 erros que se devem a construções do tipo *há de v-inf*, como *há de comer* e *há de saber*. Já o filtro relativo às frases sem verbo é mais dependente de uma boa análise sintática, e por isso mesmo o filtro é preciso quando aplicado ao Bosque. O grande responsável pelos erros é o filtro do *se*:

tanto no DHBB quanto no OBras, as frases recuperadas estão, na imensa maioria, erradas – e, portanto, onde esperaríamos orações sem sujeito, temos um sujeito omitido. Além disso, no caso do DHBB, que tem uma estrutura de texto previsível, vimos que muitos erros são, na verdade, fruto de uma mesma estrutura que se repete em vários verbetes. Uma construção como *transferindo-se*, por exemplo,[15] respondeu por 10 dos 19 erros e por 8 dos 13 erros encontrados em orações subordinadas e principais, respectivamente. No Bosque-UD, se temos bons resultados para o *se* quando este se encontra em orações principais, a qualidade da análise cai sensivelmente quando estamos diante de orações subordinadas, e o filtro d2 é responsável pela maior parte dos erros. Como indicamos na seção 4.3, devido à inconsistência (e dificuldade) de análise, mesmo no material revisto, fizemos uma busca não exatamente pelo que queríamos, mas por como nos parecia que o *se* estava anotado. Como podemos observar, essa estratégia foi pouco eficaz: ainda que seja precisa quando aplicada às orações principais, dá conta de poucos casos; no caso das orações subordinadas, recupera muitas construções, mas quase todas são casos de sujeitos omitidos. Por outro lado, quando levamos em conta os dados da Tabela 2, vemos que os casos de *se* respondem por uma porção muito pequena do total de casos filtrados no *corpus*.

Por outro lado ainda, como na omissão dos sujeitos só há duas possibilidades de resposta (estar diante de um sujeito omitido ou não), se simplesmente descartamos os filtros do *se* nos *corpora* não revistos todos os erros passam a acertos (e vice-versa), e desse ponto de vista os resultados melhoram sensivelmente. Ou seja, considerando os números do DHBB, e apenas com relação aos filtros do *se* tomados como um todo, passamos de 21% e 8% de acertos para orações principais e subordinadas, respectivamente, para 78% e 91%. No OBras, também apenas no que se refere ao *se*, quando eliminamos os filtros passamos de 40% e 15% de acertos para 60% e 85%. Diante dos resultados, optamos por prosseguir da seguinte maneira: eliminação de todos os filtros do *se* no DHBB e no OBras, e eliminação, no Bosque-UD, dos filtros d2 e d3 apenas nos casos das orações subordinadas. Com a alteração, a análise dos filtros traz resultados bastante positivos quanto aos resultados das buscas pelo sujeito oculto (TABELA 3): 89% de acertos no DHBB, 88%

[15] Como em *Transferindo-se para o Partido Social Cristão (PSC), em novembro de 1986 concorreu a deputado federal constituinte*.

no Bosque e 83% no OBras. Lembramos ainda que o *se*, responsável absoluto pelos erros, interfere pouco no resultado final, visto sua baixa frequência quando consideramos cada *corpus* na íntegra. Adicionalmente, a análise da amostra revelou que, no DHBB, os casos de *se* que não permitem omissão de sujeito (e que, portanto, são considerados erros na nova contagem) referem-se em sua imensa maioria a construções com *tratar-se de*, de modo que uma eliminação simples dessas construções torna os resultados ainda mais precisos. Especificamente, encontramos 347 casos de *tratar-se* no DHBB e 258 no OBras. Com isso, na prática, os resultados das buscas pelos sujeitos ocultos são ainda mais precisos do que indica a Tabela 3.

Diante dos resultados positivos, prosseguimos com a contagem. A Tabela 4 apresenta, finalmente, o resultado da quantificação do sujeito oculto nos três *corpora*. Começamos a contagem com a *busca ingênua* e, das ocorrências obtidas, fomos excluindo, com os filtros, os casos em que embora não haja sujeito, não há uma omissão.

TABELA 3 – Resultado final da análise dos filtros

	DHBB	**OBras**	**Bosque-UD**
Total de acertos	89%	83%	88%

TABELA 4 – Distribuição das ocorrências de sujeito oculto por *corpus*

	DHBB		OBras		Bosque-UD	
	Principal	**Subord.**	**Principal**	**Subord.**	**Principal**	**Subord.**
Busca ingênua	226122	190381	172124	156040	2777	2617
v. haver impessoal	1083	861	5181	4395	124	148
Or. Sem verbo	31599	0	43326	0	1145	0
Fenômenos da natureza	0	4	100	127	1	2
Filtros se	Não se aplica	Não se aplica	Não se aplica	Não se aplica	90	5
Filtro "tratar-se de"	181	163	150	123	Não se aplica	Não se aplica
Total	193259(40%)	189353(55%)	123367(34%)	151395(47%)	1417(15%)	2462(36%)
Total de frases com sujeito oculto	**382612 (46.5%)**		**274762 (41%)**		**3879 (24%)**	

Como suspeitávamos, o *corpus* DHBB é o que apresenta a maior quantidade de omissões de sujeito, o que não espanta dada a natureza de seu conteúdo: verbetes biográficos ou temáticos, nos quais o tema/foco da frase dificilmente se altera, e, por isso, a omissão é o recurso estilístico utilizado para deixar o texto não repetitivo. Do ponto de vista da identificação e da extração automática de informação, os resultados alertam para o fato de que em quase metade das orações (46.5%) a extração de relações entre argumentos de verbo fica prejudicada devido à ausência de um sujeito sintático. O *corpus* de obras literárias também tem um número expressivo de orações em que o sujeito foi omitido (41%), o que se explica, igualmente, em temos de estilística – lembremos das omissões de sujeito nos discursos relatados que introduzem falas de personagens, por exemplo. No Bosque-UD, composto por notícias de jornal, ainda que a frequência de sujeitos ocultos seja bem menor (24%) que nos demais materiais, não é insignificante, e traz impactos em tarefas do PLN como a extração de citação. Também vemos, na Tabela 4, que a presença de sujeito oculto é maior em subordinadas, mas que a diferença entre omissões quanto ao tipo de oração varia entre 15% e 20%. Levando em conta que frequentemente a omissão do sujeito na oração principal não se desfaz nas subordinadas, e que, nesses casos, o sujeito está fora do âmbito da sentença, esta característica é um desafio a mais para o processamento automático do texto e da informação.

Em resumo, não são poucos os casos de omissão do sujeito em português, independente de gênero, ainda que os números variem em função do tipo de texto. Do lado do PLN, o alto índice indica o desafio na construção de proposições informativas – e mesmo que em alguns casos, sobretudo o de orações subordinadas, a recuperação do elemento sujeito possa ser feita no âmbito da frase, isso envolve algum trabalho de pós-processamento. Outro desdobramento da alta frequência dos sujeitos omitidos é um aumento da dificuldade para o processamento automático no nível sintático, já que a ausência de um elemento para preencher a posição do sujeito pode desencadear outros erros. Na seção a seguir, realizamos um experimento simples para avaliar o problema.

6 Experimento: omissão do sujeito e aprendizagem automática

Dos três *corpora* utilizados em nosso estudo, apenas o Bosque é um *corpus* revisto. Por isso, este foi o material utilizado para verificar

em que medida a explicitação dos sujeitos poderia facilitar o aprendizado automático. Para tanto, reconstituímos os sujeitos e os devolvemos às frases, seguindo o procedimento do Quadro 3, no qual os elementos novos são as etapas 3 e 4, esta última inspirada na solução de Hartman *et al*. (2014).

QUADRO 3 – Etapas do experimento de busca e reconstituição de sujeitos ocultos

1. Encontrar sentenças em que não exista um sujeito (simples, oracional ou sujeito de oração passiva) dependente do núcleo da oração principal (*root*)
2. Das frases encontradas em (1) eliminar as seguintes:
   a. Construções em que *root* é o verbo haver impessoal (3ª pessoa do singular)
   b. Construções em que *root* é verbo que indica fenômeno da natureza
   c. Construções em que *root* não é verbo[16]
   d. Construções em que o *se* é um índice de indeterminação do sujeito
3. Das frases encontradas em (1), identificar aquelas em que *root* é verbo de oração adverbial, contém pessoa e número compatíveis com o sujeito de *root* e antecede *root*.
   a. Devolver este sujeito para à esquerda do verbo ao qual se relaciona.
4. Das frases que não foram eliminadas, incluir um elemento sujeito imediatamente anterior ao verbo a que se associa. O sujeito é um pronome pessoal reto e corresponde aos elementos flexionais indicados pelo verbo, ou seja, um verbo na 1ª pessoa do singular recebe o pronome *eu*. Nos casos de formas participiais, a informação de gênero também é levada em conta. No caso de infinitivo, inserimos um elemento genérico *SUBJ*.

Na etapa 1, identificamos todos os casos em que uma oração não tem um sujeito associado, e na etapa 2 eliminamos, desses casos, aqueles em que não há sujeito a ser encontrado. As etapas 3 e 4 têm o objetivo de reconstituir os sujeitos nas ocorrências restantes, sendo que a etapa 3 devolve o sintagma sujeito completo. A seguir apresentamos alguns exemplos de frases com seus sujeitos reconstituídos (sublinhados):

> Original: Quando o povo suíço recusou, em 92, a adesão ao Espaço Económico Europeu, como já fizera com a ONU, cometeu um grave engano.
>
> Reconstituído: quando o povo suíço recusou, em 92, a adesão ao Espaço Económico Europeu, como já fizera com a ONU, o povo suíço cometeu um grave engano.

[16] Comum em manchetes jornalísticas, como em "PT no poder".

> Original: Os dirigentes da Fenprof avisam no entanto desde já que se Couto dos Santos insistir nalgumas das directrizes dos seus antecessores arranjará lenha para se queimar.
>
> Reconstituído: Os dirigentes de a Fenprof avisam em o entanto desde já que se Couto de os Santos insistir em algumas de as directrizes de os seus antecessores Couto Santos arranjará lenha para se queimar.
>
> Original: Quando a gente se diz engajado, corre o risco de evocar modelos anteriores e o engajamento hoje deve encontrar formas novas.
>
> Reconstituído: Quando a gente se diz engajado, a gente corre o risco de evocar modelos anteriores e o engajamento hoje deve encontrar formas novas.

A etapa 4 devolve sujeitos "genéricos", isto é, formas pronominais informadas por traços morfossintáticos do verbo ou de predicadores, e um genérico *SUBJ* para o caso em que tais informações não estão disponíveis (verbos no infinitivo, por exemplo). Ainda que o desenvolvimento de estratégias para a reconstituição correta seja um exercício desafiador e relevante, nosso foco aqui está apenas em verificar se a presença de um sujeito é um elemento capaz de facilitar a aprendizagem automática. E, para isso, interessa devolver os sujeitos às orações que de fato precisam de um (em oposição a todos os casos excluídos pelo filtro 2). O índice de acerto acima de 80% de cada estratégia (cf. TABELA 2) sugere que os resultados da reconstituição, de um ponto de vista da estrutura da oração, são confiáveis.

Separamos o Bosque-UD reconstituído nas partições *treino* (incluímos a partição *dev* no treino), e *teste*, segundo a distribuição do Bosque-UD.[17] Utilizando a ferramenta UDPipe, criamos um novo modelo (com os sujeitos reconstituídos) e o avaliamos. Criamos, ainda, um cenário alternativo de avaliação, que corresponde ao Bosque-UD clássico. Isto é, no treino, utilizamos sujeitos reconstituídos, mas no

[17] A divisão de um corpus (ou de um *dataset*) em partições de treino (*train*), desenvolvimento (*dev*) e teste (*test*) são próprias para o aprendizado de máquina, e indicam respectivamente o conjunto de dados que será usado para treinar (ou aprender), para realizar ajustes e para avaliar o modelo criado.

teste consideramos o texto original, com sujeitos ocultos. Quisemos, com isso, reproduzir um cenário real, no qual os textos não terão seus sujeitos reconstituídos, porque para reconstituí-los é necessária uma análise sintática prévia, e é justamente isso que estamos medindo.

A Tabela 4 apresenta os resultados da aprendizagem com e sem a reconstituição do sujeito em termos de métricas específicas para a avaliação de dependências sintáticas: UAS, LAS e CLAS (*content-word labeled attachment score*). Como podemos observar, os resultados são melhores no *corpus* com o sujeito reconstituído, sugerindo facilidade na aprendizagem. A maior alteração está na métrica CLAS, com aumento de 2%, e este é um dado relevante, uma vez que CLAS representa melhor os resultados da análise sintática por tratar de relações entre palavras de conteúdo/classes abertas. Especificamente, a métrica CLAS não leva em conta as relações entre um sinal de pontuação e qualquer elemento (visto que serão sempre uma relação do tipo *punct)* e as relações entre determinantes, preposições, auxiliares e conjunções e qualquer outra palavra, visto que serão sempre relações de um mesmo tipo: *det, case*; *aux* e *mark*, respectivamente (NIVRE; FANG, 2017). Por isso, CLAS é uma medida mais sensível ao aprendizado das relações sintáticas. Por outro lado, quando observamos os resultados relativos à reconstituição apenas no treino, a melhora, ainda que permaneça, é menor. Se este último dado traz algum desapontamento no que se refere à aprendizagem de aspectos sintáticos, sinaliza também que, para tarefas cujo foco não é a análise sintática, mas que necessitem de análise sintática prévia, como a anotação de papéis semânticos, a reconstituição de sujeitos é uma estratégia que merece ser considerada.

TABELA 5 – Comparação da aprendizagem sintática em diferentes cenários relativos à omissão dos sujeitos

	Bosque-UD v.2.6 clássico	Bosque-UD v.2.6 com reconstituição no treino e na validação	Bosque-UD v.2.6 com reconstituição só no treino
UAS (F1)	84.81	85.66	85.34
LAS (F1)	80.63	81.85	81.01
CLAS (F1)	72.67	74.81	72.83

7 Considerações finais e desafios futuros

Dentre as várias utilidades que um *corpus* oferece, fazemos uso de duas: *corpus* para medir um fenômeno, e *corpus* para treinar um sistema e gerar um modelo de língua. Nossa conclusão é a de que o tratamento do sujeito oculto em português é relevante no PLN por dois motivos: (i) porque é um fenômeno altamente frequente, e não levá-lo em consideração tem como consequência limitar as possibilidades da extração automática de conteúdos; e (ii) porque a reconstituição do sujeito, quando possível, é capaz de facilitar a aprendizagem automática de dependências sintáticas, o que é positivo para todas as tarefas que, em algum momento, se utilizam deste tipo de análise linguística. Com relação a esse segundo aspecto, chamamos a atenção para a relevância da modelagem linguística na aprendizagem automática. Sem qualquer interferência na forma de aprendizagem, apenas lidando com informação linguística, conseguimos um impacto positivo que pode ser de até 2% no aprendizado automático. Um terceiro cenário que pretendemos testar consiste em aplicar as regras de reconstituição de sujeito nos resultados da anotação automática e, então, reanotar sintaticamente o material com os sujeitos reconstruídos. Isso nos dará outra medida relacionada aos benefícios da reconstituição do sujeito na análise sintática.

Diante dos resultados obtidos, outra tarefa se impõe: a reconstituição precisa do sujeito em termos discursivos (e não apenas sintáticos ou formais), o que esbarra também na resolução de correferência.

Como um resultado adicional, mas não inesperado, ratificamos a necessidade e relevância de *corpora* com anotação linguística de qualidade. Quando se trata de *treebanks* revistos (e públicos), a língua portuguesa dispõe apenas do Bosque, em suas variadas versões, desde 2002. E, ainda assim, temos fenômenos linguísticos que carecem de um tratamento sistemático, como é o caso da classificação do pronome *se*.

Apesar de não ser o foco deste trabalho, a identificação de sujeitos ocultos, por um lado, e sua reconstituição, por outro, são de grande valia também para a área de simplificação textual e de leitura. De um ponto de vista didático, um *corpus* com sujeitos artificialmente (bem) reconstituídos pode funcionar como "gabarito" para exercícios de leitura, por exemplo. Porém, para que os resultados sejam confiáveis, é crucial garantir a qualidade da anotação sintática subjacente. Embora

um índice de acertos que se aproxima dos 90% seja suficientemente bom para uma série de tarefas de PLN, reconhecemos que ainda há espaço para melhorias (veja-se, por exemplo a página *NLP Progress*,[18] um repositório para o monitoramento da evolução de tarefas de PLN que elenca o estado da arte para as tarefas mais comuns. Infelizmente, grande parte das tarefas têm como alvo a língua inglesa).

Por fim, lembramos que o ganho que tivemos pode ser ampliado por meio de técnicas mais elaboradas de aprendizado de máquina. Para isso, disponibilizamos o *corpus* Bosque-UD reconstituído, além das regras que permitem a identificação do sujeito, bem como sua reconstituição.[19] Nesse contexto, chamamos a atenção para a relevância da modelagem linguística no aprendizado automático, tendo em vista a resolução de tarefas de PLN.

Agradecimentos

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela bolsa de Iniciação Científica concedida a Elvis de Souza no âmbito do projeto "Construção de datasets para o PLN de Língua Portuguesa". Número do processo da bolsa: 128693/2019-3.

Contribuição dos autores

Cláudia Freitas foi responsável pela concepção do trabalho, e Elvis de Souza, pela preparação dos dados. A análise dos dados e a redação foram realizadas por ambos.

Referências

AFONSO, S.; BICK, E.; HABER, R.; SANTOS, D. Floresta sintá(c)tica: A Treebank for Portuguese. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2002), 3rd, 2002, Las Palmas de Gran Canaria. *Proceedings* […]. Las Palmas de Gran Canaria: ELRA, 2002. p. 1698-1703.

[18] Disponível em: http://nlpprogress.com/. Acesso em: 8 out. 2020.

[19] Disponível em: https://github.com/alvelvis/desocultando-sujeitos. Acesso em: 30 nov. 2020.

BICK, E. *The parsing system palavras*: Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Famework. Aahus, Dinamarca: Aarhus Universitetsforlag, 2000.

DURAN, M. S.; ALUÍSIO, S. M. Propbank-Br: a Brazilian Treebank Annotated with Semantic Role Labels. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 12), 8th., 2012, Istambul, *Proceedings* […]. Istambul: ELRA, 2012. p. 1862-1867.

ELSON, D.; MCKEOWN K. Automatic Attribution of Quoted Speech in Literary Narrative. *In:* CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI 10), 24th., 2010, Atlanta, *Proceedings* […]. Atlanta: The AAAI Press, 2010. p. 1013-1019.

FINATTO, M. J.; SCARTON, C.; ROCHA, A.; ALUÍSIO, S. Características do jornalismo popular: avaliação da inteligibilidade e auxílio à descrição do gênero. *In:* 8TH BRAZILIAN SYMPOSIUM IN INFORMATION AND HUMAN LANGUAGE TECHNOLOGY (STIL 2011), 8th., 2011, Cuiabá, *Proceedings* […]. Cuiabá: SBC, 2011. p. 49-58.

FREITAS. C.; ROCHA, P.; BICK, E. Um mundo novo na Floresta Sintá(c)tica – o treebank do Português. *Calidoscópio*, São Leopoldo, RS, v. 6, n. 3, p. 142-148, 2008. DOI: https://doi.org/10.4013/cld.20083.03

HARTMANN, N. S.; DURAN, M. S.; ALUÍSIO, S. M. Filling the Gap: Inserting an Artificial Constituent Where a Subject Is Omitted in Portuguese. *In:* WORKSHOP ON TOOLS AND RESOURCES FOR AUTOMATICALLY PROCESSING PORTUGUESE AND SPANISH (TORPOR), I., São Carlos, *Proceedings* […]. São Carlos: SBC, 2014. Disponível em: http://www.nilc.icmc.usp.br/semanticnlp/includes/projects/brazilis/artigos/ToRPorEsp,%202014.pdf. Acesso em: 8 out. 2020.

HIGUCHI, S.; SANTOS, D.; FREITAS, C.; RADEMAKER, A. Distant Reading Brazilian Politics. *In:* CONFERENCE OF THE ASSOCIATION DIGITAL HUMANITIES IN THE NORDIC COUNTRIES (DHN 2019), 4th., 2019, Copenhagen. *Proceedings* […]. Copenhagen: University of Copenhagen, 2019. p. 190-200.

JONES, K. S. Computational Linguistics: What about the Linguistics? *Computational Linguistics*, Cambridge, MA, v. 33, n. 3, p. 437-441, 2007. DOI: https://doi.org/10.1162/coli.2007.33.3.437

LUFT, C. P. *Moderna gramática brasileira*. Rio de Janeiro: Globo Livros, 2002.

MARCUS, M.; SANTORINI, B.; MARCINKIEWICZ, M. A. Building a Large Annotated Corpus of English: The Penn Treebank. *Computational Linguistics*, Cambridge, MA, v. 19, n. 2, p. 313-330, 1993. DOI: https://doi.org/10.21236/ADA273556

NIVRE, J.; de MARNEFFE, M.C.; GINTER, F.; GOLDBERG, Y.; HAJIČ, J.; MANNING, C.D.; McDONALD, R.; PETROV, S.; PYYSALO, S.; SILVEIRA, N.; TSARFATY, R.; ZEMAN, D. Universal Dependencies v1: A Multilingual Treebank Collection. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'16), 10th., Portorož, *Proceedings* [...]. Portorož: ELRA, 2016. p. 1659-1666.

NIVRE, J.; FANG, C. Universal Dependency Evaluation. *In:* UNIVERSAL DEPENDENCIES WORKSHOP (UDW 2017), 2017, Gothenburg, *Proceedings* [...]. Gothenburg: Association for Computational Linguistics, 2017. p. 86-95.

RADEMAKER, A. CHALUB, F.; REAL, L.; FREITAS, C.; BICK, C.; de PAIVA, V. Universal Dependencies for Portuguese. *In:* INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS (DEPLING 2017), 4th., Pisa, *Proceedings* [...]. Pisa: Linköping University Electronic Press, 2017. p. 197-206.

RUANO SAN SEGUNDO, P. A Corpus-Stylistic Approach to Dickens' Use of Speech Verbs: Beyond Mere Reporting. *Language and Literature,* [*S.l.*], v. 25, n. 2, p. 113-129, 2016. DOI: https://doi.org/10.1177/0963947016631859

SAMPSON, G. *Empirical Linguistics*. London: Continuum, 2001.

SANTOS, D.; BICK, E. Providing Internet Access to Portuguese Corpora: the AC/DC project. *In:* INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC 2000), 2nd., Atenas, *Proceedings* [...]. Atenas: ELRA, 2000. p. 205-210.

SANTOS, D. Linguateca's Infrastructure for Portuguese and How It Allows the Detailed Study of Language Varieties. *OSLa*: *Oslo Studies in Language,* Oslo, v. 3, n. 2, p. 113-128, 2011. DOI: https://doi.org/10.5617/osla.100

SANTOS, D.; FREITAS, C.; BICK, E. OBras: A Fully Annotated and Partially Human-Revised Corpus of Brazilian Literary Works in the Public Domain. 2018. Disponível em: https://opencor.gitlab.io/corpora/santos18obras. Acesso em: 8 de out. 2020.

de SOUZA, E.; FREITAS, C. ET: uma Estação de Trabalho para revisão, edição e avaliação de corpora anotados morfossintaticamente. *In:* WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (TILic 2019), VI., 2019. Salvador. *Proceedings* [...]. Salvador: SBC, 2019. p. 15-18.

STRAKA, M.; HAJIC, J.; STRAKOVÁ, J. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *In*: TENTH INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUATION (LREC'16), 10th., Portorož, *Proceedings* [...]. Portorož: ELRA, 2016. p. 4290-4297.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1059-1087, 2021

O papel do corpus de estudo no aprimoramento descritivo da complementaridade informacional multidocumento

The role of the study corpus in the descriptive improvement of multi-document informational complementarity

Jackson Wilke da Cruz Souza
Universidade Federal de Alfenas (UNIFAL-MG), Varginha, Minas Gerais / Brasil
jackcruzsouza@gmail.com
http://orcid.org/0000-0003-1881-6780

Resumo: Em subáreas do Processamento Automático de Línguas Naturais (PLN), como a Sumarização Automática Multidocumento (SAM), é necessário compreender o comportamento linguístico de determinados fenômenos, especialmente os de natureza semântica. A *Cross-document Structure Theory* (CST) é bastante utilizada em estudos do PLN por proporcionar um conjunto de relações semânticas que organizam a informação entre unidades de análise (comumente, pares de sentenças), agrupadas entre *conteúdo* (a saber, redundância, complementaridade e contradição) e *apresentação* (a saber, fonte/autoria e estilo). Até então, a caracterização das relações CST baseava-se em atributos *genéricos* (como a quantidade de palavras em comum entre as sentenças de um par) e *específicos* (como a presença de advérbios temporais) para as relações de Redundância e Complementaridade. Entretanto, percebe-se que a delimitação de tais atributos ainda é incipiente, pois não inclui atributos semânticos e pragmáticos, níveis linguísticos que são possíveis de recuperar manualmente entre as unidades de análise da CST. Nesse sentido, objetiva-se, neste artigo, reconstruir o percurso metodológico de Souza (2019) ao que se refere ao estudo em *corpus* das relações CST em textos jornalísticos do Português, já que o conjunto de atributos disponíveis, até o momento, ainda produzia equívocos na identificação dos subtipos de complementaridade multidocumento, a saber temporal e atemporal. Partindo do *corpus* CSTNews, organizou-se um subconjunto de estudo com os 10 primeiros *clusters*, o que contabilizou 204 pares de sentenças. Como resultado, foram obtidas a descrição detalhada da complementaridade CST e a criação de uma tipologia de sinalizadores das relações que traduzem esse fenômeno, além da proposição de uma metodologia específica para o estudo de relações CST.

Palavras-chave: Complementaridade informacional multidocumento; Processamento Automático de Línguas Naturais; *Corpus* de estudo.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1059-1087

Abstract: In sub-areas of Natural Language Processing (NLP), such as Automatic Multidocument Summarization (AMS), it is necessary to understand the linguistic behavior of certain phenomena, especially those of a semantic nature. Cross-document Structure Theory (CST) is widely used in NLP studies because it provides a set of semantic relations that organize information between units of analysis (commonly, pairs of sentences) organized between *content* (namely, redundancy, complementarity and contradiction) and *presentation* (namely, source/authorship and style). Until then, the characterization of CST relationships was based on *generic attributes* (such as the number of words in common between sentences of a pair) and *specific attributes* (such as the presence of temporal adverbs) for the relationships of Redundancy and Complementarity. However, the delimitation of such attributes is still incipient, as they do not include semantic and pragmatic attributes, linguistic levels that are possible to recover between the CST units of analysis. In this sense, the aim of this paper is to reconstruct the methodological path of Souza (2019) with regard to the study in *corpus* of CST relations in Portuguese journalistic texts, since the set of available attributes, until then, still produced mistakes in the identification of multi-document complementarity subtypes, namely temporal and timeless. Based on the CSTNews *corpus*, a subset of studies was organized with the first 10 clusters, that are represented by 204 pairs of sentences. As a result, a detailed description of CST complementarity was obtained, as well as the creation of a typology of signaling relationships that translate this phenomenon, in addition to proposing a specific methodology for the study of CST relations.

Keywords: Multi-document informational Complementarity; Processing of Natural Languages; Study *corpus*.

Recebido em 10 de outubro de 2020

Aceito em 04 de janeiro de 2021

1 Introdução

As pesquisas na área de Linguística de *Corpus* (doravante, LC) têm se dedicado a estudar fenômenos linguísticos a partir de textos produzidos, em sua maioria, por humanos. Assim, derivam-se os estudos em Terminologia (TAGNIN; BEVILACQUA, 2015), em Linguística Aplicada (VIANA; TAGNIN, 2011), em Linguística Descritiva (RODRIGUES, 2019) ou em Processamento de Automático de Línguas Naturais (PLN) (CASELI, 2015).

Os *corpora*, de maneira geral, permitem identificar características que destacam a evidência de certos fenômenos em um ambiente linguisticamente natural. Ao tentar delimitar um candidato a termo de uma área médica, por exemplo, a análise do texto que está ao seu redor dará pistas ao pesquisador se aquele candidato é um hiperônimo ou hipônimo de outro termo. É nesse processo de observação dos fenômenos linguísticos nos *corpora* que é possível corroborar teorias, desenvolvê-las ou mesmo refutá-las.

De acordo com Sardinha (2000), um *corpus* pode ter finalidades de estudo, referência e treinamento (ou teste). O *corpus* de estudo subsidia análises preliminares do objeto e/ou do fenômeno em observação. O *corpus* de referência oferece suporte a uma análise contrastiva entre este e o *corpus* de estudo. Já o *corpus* de treinamento é submetido a testes que se pautarão no conhecimento levantado a partir das observações realizadas nos *subcorpora* de estudo e de referência.

Dentre as diversas contribuições que a Linguística de *corpus* pode promover à Computação, destacam-se, aqui, o estudo, a descrição e a caracterização de fenômenos linguísticos em *corpus*. Ao analisarem o comportamento linguístico da complementaridade entre textos jornalísticos do português, Souza (2015) e Souza e Di-Felippo (2018) basearam-se em características de maior acurácia recomendadas pela literatura, como a presença de *expressões temporais*. A partir da proposição de um conjunto de atributos, os autores submeteram os pares de sentenças ao *Waikato Environment for Knowledge Analysis* (Weka) (HALL *et al.,* 2009), resultando em algoritmos de Aprendizado de Máquina (AM) que discriminam as relações semânticas de complementaridade (a saber, *Historical Background*, *Follow-up* e *Elaboration*) da *Cross-document Structure Theory* (CST) (RADEV, 2000). Como resultado, os algoritmos de AM propostos obtiveram cerca de 75% de precisão em identificar as relações dos pares de sentença.

Entretanto, observou-se que entre as relações *Follow-up* e *Elaboration*, classificadas por Maziero (2012) e Maziero, Jorge e Pardo (2010), ainda havia equívocos devido à similaridade entre elas. De acordo com os autores, *Follow-up* ocorre quando, em um par de sentenças (S1 e S2), S2 apresenta acontecimentos que ocorrem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si. Já a relação *Elaboration* ocorre quando, dado o par de sentenças, S2 detalha/refina/elabora algum

elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Para tanto, exemplificam-se esses apontamentos em (1).

(1)

S1: Confrontos entre o Exército e o grupo rebelde Tigres Tâmeis eclodiram na região de Muttur há duas semanas, após a guerrilha ter cortado o suprimento de água para alguns vilarejos.

S2: Os rebeldes afirmaram que consideram o novo bombardeio do Exército equivalente a "uma declaração de guerra".

De acordo com a definição proposta por Maziero (2012) e Maziero, Jorge e Pardo (2010), o par de sentenças em (1) pode ser identificado tanto como *Follow-up* (pois o evento narrado em S2 ocorre após S1), como *Elaboration* (pois S2 apresenta detalhes acerca da afirmação dos rebeldes indicados em S1). São em casos semelhantes a esse que os algoritmos desenvolvidos por Souza (2015) e Souza e Di-Felippo (2018) cometiam equívocos de classificação.

Além da similaridade entre as relações CST de complementaridade, até então, os estudos se baseavam em descrever as relações com base em atributos presentes (ou ausentes) na superfície textual. Assim, em um par de sentenças que ocorresse *advérbio ou locução adverbial de tempo*, potencialmente poderia ser classificado como complementaridade temporal. Isso muito se deve ao fato de a descrição estar bastante empenhada em promover algoritmos automáticos que pudessem identificar e classificar as relações CST. Nesse sentido, de certa maneira, foram deixados de lado atributos que não pudessem ser processados computacionalmente, sob a justificativa que tais atributos não deixam evidências de relações semânticas no texto.

Nesse contexto, Das e Taboada (2018) e Taboada e Das (2013) advogam que qualquer fenômeno semântico se expressa por meio do texto e sempre deixa marcas ou mesmo indícios para apontar informações adicionais para fora dele. Os autores reanotaram o *RST Signalling Corpus* (DAS; TABOADA; MCFETRIDGE, 2015), o qual contém textos em que as proposições estão estruturadas de acordo com o modelo *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987). O objetivo dos estudos foi identificar como os marcadores discursivos (como *conjunções ou locuções conjuntivas de adversidade*, por exemplo) contribuem para o sentido do discurso ao sinalizar relações no texto. Após esses estudos, os

autores organizaram tipologicamente os sinais em *genéricos* e *específicos*, que podem ocorrer sozinhos ou combinados.

Para estudar o comportamento linguístico-estrutural do modelo CST e identificar os sinalizadores de suas relações, é necessário conceber esse modelo como teoria semântica, ainda que suas relações não sejam intencionais. Assim, baseando-se no aprofundamento da análise da complementaridade proposta por Souza (2019), objetiva-se apresentar detidamente os procedimentos metodológicos acerca do estudo da complementaridade em *corpus*. Além do próprio estudo do fenômeno, objetiva-se salientar o lugar do *corpus* de estudo na LC como ferramenta de desenvolvimento de teorias linguísticas ou mesmo de suas possíveis revisões e/ou aprimoramentos.

Para tanto, este artigo está organizado em cinco seções. Na Seção 2, apresentam-se algumas reflexões sobre a LC, a fim de destacar os critérios de construção e anotação de *corpora* linguísticos. Na Seção 3, tem-se o panorama acerca dos sinalizadores linguísticos sob a perspectiva de duas teorias discursivas distintas, a RST e a CST. Na Seção 4, demonstra-se o estudo baseado em *corpus*, que resultou no levantamento de sinalizadores da complementaridade via modelo CST. Por fim, na Seção 5, tecem-se considerações finais, além de se delinearem trabalhos futuros.

2 Da construção à anotação de *corpus*

É notável como a LC passou por diversos aprimoramentos teóricos e metodológicos nos últimos anos, especialmente com a contribuição de abordagens computacionais. Tagnin (2018) aponta que, desde a publicação do *Brown University Standard Corpus of Present-Day American*, em 1964, o cenário na LC tem mudado. Essas mudanças são promovidas pela utilização de ferramentas e abordagens computacionais na área, o que resultou em transformações de perspectivas teóricas e técnicas sobre o conceito de *corpus* e como ele pode ser utilizado em pesquisas linguísticas.

Com relação ao conceito de *corpus*, Sardinha (2004) propõe que esse recurso linguístico pode ser definido como

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e

> profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (SARDINHA, 2004, p.18.)

A partir dessa definição, é possível resgatar alguns critérios relevantes e necessários para que dado conjunto de textos não seja tido apenas como um arquivo, coletânea ou biblioteca de textos.

O primeiro critério que Sardinha (2004) chama à atenção é o *tamanho*, a fim de garantir que o fenômeno a ser observado possa ocorrer no conjunto de textos. Assim, o autor classifica o conjunto em *pequeno* (menos de 80 mil palavras), *pequeno-médio* (80 a 250 mil palavras), *médio* (250 mil a 1 milhão), *médio-grande* (1 milhão a 10 milhões) e *grande* (acima de 10 milhões de palavras). O *Brown Corpus*, por exemplo, tinha 2.000 palavras (frente a cerca de 1 milhão atualmente), enquanto o *Corpus* do Núcleo Interinstitucional de Linguística Computacional (NILC) possui mais de 40 milhões de palavras (KUHN; ABARCA; NUNES, 2000) e o iWeb, mais de 14 bilhões (DAVIES; KIM, 2019). O aumento no tamanho dos *corpora* deve-se, certamente, às facilidades que os sistemas e ferramentas computacionais proporcionaram às pesquisas em LC nos últimos anos.

Atrelado ao tamanho, Sardinha (2004) também propõe que a profundidade é um critério relevante na construção de um *corpus*. Se o conjunto de textos não apresentar profundidade, pouco relevante será sua grande extensão. Nesse sentido, a escolha dos textos que farão parte do conjunto deve seguir critérios rigorosos, quanto ao balanceamento dos textos (de gênero textual, por exemplo). Para tanto, o autor apresenta uma tipologia que tenta estabelecer certa profundidade aos *corpora* linguísticos, a saber: *modo* (falado ou escrito), *tempo* (sincrônico, diacrônico, contemporâneo ou histórico), *seleção* (de amostragem, monitor, dinâmico, estático ou equilibrado), *conteúdo* (especializado, regional/dialetal ou multilíngue), *autoria* (de aprendiz ou de língua nativa), *disposição interna* (paralelo ou alinhado) e *finalidade* (de estudo, de referência ou de treinamento/teste).

Outro critério importante é a representatividade. Biber (2012) aponta que o conjunto de textos deve demonstrar o fenômeno estudado em seu ambiente natural de ocorrência, dadas suas especificidades. O autor ressalta que esse critério deve ser anterior ao planejamento do

corpus, reconhecendo "os parâmetros situacionais que variam entre os textos de uma comunidade discursiva e também os tipos de características linguísticas que serão examinadas no *corpus*" (BIBER, 2012, p.11). Conceber a representatividade dessa maneira influencia diretamente a construção do *corpus*, que é organizada, segundo o autor, em duas etapas: "o planejamento original baseado em análises teóricas e de estudo-piloto de uma coleta de textos, por investigações empíricas detalhadas da variação linguística, e por uma revisão do planejamento" (BIBER, 2012, p.11).

Além dos critérios advindos da própria natureza e do conceito de *corpus*, também se derivam critérios específicos à pesquisa a que ele subsidia. Di-Felippo e Souza (2012) indicam que há decisões de projeto que devem fazer parte dos critérios de construção, pautando-se na utilização de recursos linguístico-computacionais em tarefas de estruturação do conjunto de textos e métodos semiautomáticos de extração de conhecimento. Ademais, os autores salientam que os critérios de (i) definição do objeto *corpus*, (ii) seleção do tipo de recurso linguístico a ser construído e (iii) decisões de projeto contribuem para um *design* do *corpus* adequado à pesquisa.

Um dos produtos que resulta da construção de um *corpus* é a anotação linguística. Essa tarefa pode ser definida como o processo de enriquecimento do *corpus*, adicionando (manual ou automaticamente) informações linguísticas, com objetivos teóricos (acolher uma teoria linguística, por exemplo) ou práticos (treinar um etiquetador morfológico, por exemplo). Embora o *corpus*, por si só, já seja um recurso bastante importante, quando anotado, torna-se caro à pesquisa que o desenvolveu, bem como a outras que posteriormente dele podem se beneficiar. Isso ocorre porque as anotações acrescentam valor ao *corpus*, permitindo que sejam realizados buscas e processamentos mais refinados (PEDRO; VALE, 2018).

Hovy e Lavid (2010) defendem que a anotação evidencia a perspectiva sobre a língua e a teoria linguística adotadas no estudo, como quais são as unidades de análise que são anotadas na RST (proposições) e na CST (sentenças, palavras ou porções textuais), por exemplo. Os autores equacionam metodologicamente a anotação em oito tarefas, a saber: (i) selecionar textos que sejam representativos para um *corpus* de treinamento, (ii) selecionar a teoria ou o conceito linguístico que subsidie um conjunto de etiquetas que será aplicado na tarefa, (iii)

anotar um pequeno fragmento do *corpus* de treinamento, (iv) medir comparativamente a concordância entre os anotadores, (v) decidir qual o nível de concordância será adotado no trabalho e, caso, não seja satisfatória, voltar a partir da etapa (ii) e fazer as adaptações necessárias, (vi) anotar uma maior parte do *corpus*, (vii) utilizar aprendizado de máquina a fim de, posteriormente, automatizar o processo de anotação e (viii) caso o desempenho dos algoritmos seja satisfatório, anotar automaticamente uma porção de textos ainda não anotados.

Mesmo após refinar o modelo teórico, é durante a anotação que se verificam certas incongruências no *corpus*, como a falta de representatividade do fenômeno estudado devido ao seu tamanho, por exemplo. Nesse contexto, Taboada e Das (2013) realizaram uma nova anotação sobre outra já feita: ao verificar que os marcadores discursivos usualmente utilizados para caracterizar e identificar as relações do modelo RST eram insuficientes, propuseram-se a estudar outras pistas linguístico-estruturais que pudessem ser consideradas como tal. A partir de então, os autores passaram a denominar tais pistas como *sinalizadores* das relações semânticas do modelo.

Construindo um paralelo aos pressupostos metodológicos de anotação de Hovy e Lavid (2010) em perspectiva à proposta de Taboada e Das (2013), adotou-se, neste trabalho, a estratégia de investigar quais os possíveis sinalizadores das relações de complementaridade informacional do modelo CST a partir da construção de um *corpus* de estudo. A seguir, têm-se as reflexões acerca dos sinalizadores que permitem remontar as relações de sentido entre unidades de análise dos modelos RST e CST.

3 Panorama dos sinalizadores de relações semânticas

O estudo de relações semânticas, em sua maioria, baseia-se no levantamento de conjuntos de pistas que possam caracterizar linguistica ou estruturalmente tais relações e, *a posteriori*, subsidiar a identificação (automática) de cada uma. Ao que se refere aos estudos que abordam direta ou indiretamente a complementaridade informacional, destacam-se Das e Taboada (2018), Taboada e Das (2013), para o modelo RST, e Maziero (2012), Souza (2015) e Souza e Di-Felippo (2018), para o modelo CST.

Comumente, a identificação (manual e automática) de relações RST é feita com base em marcadores discursivos. Em dado texto, ao

utilizar *se* como conjunção entre duas proposições, por exemplo, o autor planeja evidenciar uma *relação condicional* entre unidades discursivas. Dessa maneira, a conjunção marca a relação em questão.

Entretanto, Das e Taboada (2018) e Taboada e Das (2013) advogam que a ideia mais difundida na literatura sobre *marcador discursivo*, na verdade, limita a identificação das relações RST, pois se baseia apenas naquilo que está expresso no texto. Dado que as sentenças "Por conta de consumirem muita lactose, João e Pedro não podem ingerir leite" e "Os jovens continuam consumindo leite em sua versão 'sem lactose'" tenham sido extraídas do mesmo texto, percebe-se que o autor deseja evidenciar *contraste* entre as informações das duas proposições, porém, sem utilizar um marcador discursivo para tanto. Ademais, para que essa análise hipotética seja verdadeira, é preciso assumir que "os jovens" retomam anaforicamente "João e Pedro".

Assim, os autores propõem a hipótese de que, se a relação preterida pelo autor do texto é compreensível pelo leitor, a relação, então, é recuperável, ainda que o marcador não ocorra na superfície textual. Nesse sentido, a relação semântica deve apresentar algum sinalizador para que o leitor interprete o mais próximo possível a intenção do autor. No exemplo dado, um possível sinalizador seria a *anáfora lexical* entre as duas sentenças, além das proposições negativa ("não podem consumir leite") e afirmativa ("continuam consumindo leite") na primeira e na segunda sentenças, respectivamente

Taboada e Das (2013), revisando um *corpus* anotado com o modelo RST, propuseram um conjunto de sinalizadores que não foram previamente identificados como sinalizadores das relações previstas no modelo teórico, como é o caso das *anáforas lexicais*. Como resultado, apresentaram uma taxonomia de sinalizadores, incluindo os marcadores discursivos recorrentemente utilizados em estudos dessa natureza. Essa taxonomia organiza-se em nove categorias que superordenam outras subcategorias, a saber:

a) *Marcador discursivo:* sinalizadores que se caracterizam por serem marcas específicas de cada uma das relações RST; em geral, são tidos como expressões léxicas ou conjunções;

b) *Entidade*: sinalizadores que se caracterizam por estabelecerem similaridade ou dissimilaridade entre entidades nomeadas de unidades discursivas;

c) *Semântico:* sinalizadores que manifestam relações lexicais (hiperonímia, por exemplo) entre duas entidades de unidades discursivas distintas;

d) *Léxico:* sinalizadores que são traduzidos em palavras que indicam algum tipo de relação, como acrescentar uma informação;

e) *Morfológico:* sinalizadores que auxiliam a identificar fatores temporais por meio de desinências verbais;

f) *Sintático:* sinalizadores que indicam relações RST por meio de construções sintáticas específicas, como o discurso indireto;

g) *Gráfico:* sinalizadores que podem indicar relacionamento semântico por meio de pontuações, como as vírgulas em elipses;

h) *Numérico:* sinalizadores que evidenciam especificações entre unidades discursivas, detalhando ou ressaltando alguma informação apresentada genericamente (p.ex. "João, Pedro e Paulo foram acompanhar Maria no aeroporto" e "A garota estava acompanhada de seus três amigos no aeroporto");

i) *Gênero (textual):* sinalizadores que evidenciam marcas textuais específicas de cada gênero, como o informativo, em que as primeiras sentenças de um texto desse gênero terão informações genéricas, as quais serão elaboradas/detalhadas nas sentenças subsequentes.

Os autores concluíram que (i) há sinalizadores que ocorrem mais recorrentemente em determinadas relações (como é o caso de *gênero textual*, que ocorre mais em *Elaboration*); (ii) há aqueles que caracterizam certas relações somente sob combinações com outros (como é o caso de *construção frasal* que ocorre juntamente com *sintático* para caracterizar a relação *Background*); e (iii) há sinalizadores que, até então, não tinham sido explorados (como é o caso de *pontuação*).

Em um estudo mais recente, Das e Taboada (2018) refinaram a análise prévia, e organizam os sinalizadores em *singulares* e *combinados*. Os singulares são os mesmos apresentados no trabalho anterior, e os combinados são *referenciais* (ou anafóricos), *semânticos* e *gráficos* em conjunto com algum do tipo sintático, cada um deles.

Já acerca da identificação das relações do modelo CST, há os trabalhos de Maziero (2012), Souza (2015) e Souza e Di-Felippo (2018).

Visando à identificação automática das relações CST no contexto da Sumarização Automática Multidocumento, Maziero (2012) propõe uma série de métodos que se baseiam em sinalizadores explícitos na sentença. O autor parte do princípio de que as relações desse modelo semântico sempre compartilham informações entre duas unidades de análise, manifestando, portanto, redundância em maior ou menor grau (RADEV, 2000). Especificamente, ao analisar duas sentenças (S1 e S2) advindas de textos distintos, mas que versam sobre o mesmo assunto, o autor identifica as relações CST com base em (i) Diferença de tamanho em palavras (S1-S2), (ii) Porcentagem de palavras em comum em S1, (iii) Porcentagem de palavras em comum em S2, (iv) Posição de S1 no texto (início, meio ou fim), (v) Número de palavras na maior *substring* entre S1 e S2, (vi) Diferença no número de substantivos entre S1 e S2, (vii) Diferença no número de advérbios entre S1 e S2, (viii) Diferença no número de adjetivos entre S1 e S2, (ix) Diferença no número de verbos entre S1 e S2, (x) Diferença no número de nomes próprios entre S1 e S2, (xi) Diferença no número de numerais entre S1 e S2 e (xii) Sobreposição de sinônimos entre S1 e S2.

Além desses métodos, Maziero (2012) utilizou alguns específicos para a identificação das relações *Identity*, *Contradiction*, *Attribution*, *Indirect Speech* e *Translation*. O método formulado para a identificação da relação *Contradiction*, por exemplo, prevê apenas os casos de contradição do tipo explícita, isto é, resultantes de diferenças numéricas entre as sentenças de um par.

Para avaliar os métodos propostos, o autor utilizou o *corpus* CSTNews (CARDOSO *et al.,* 2011; DIAS *et al.,* 2014). Esse *corpus* se caracteriza por ser um conjunto multidocumento de textos jornalísticos em português, e está anotado com as relações do modelo CST. Trata-se de 50 *clusters* de notícias (em média, 3 textos) que possuem um mesmo assunto, sendo provenientes de fontes jornalísticas *online* distintas. No total, o CSTNews possui 140 textos, somando 2.088 sentenças e 47.240 palavras.

Com base nos métodos descritos, o Maziero (2012) desenvolveu algoritmos de AM, cuja precisão geral foi de 68,13%. Essa precisão é a média ponderada da precisão dos métodos para a identificação das relações *Overlap*, *Subsumption*, *Elaboration*, *Equivalence*, *Historical Background* e *Follow-up*, *Identity*, *Contradiction*, *Attribution*, *Indirect Speech* e *Translation*. Segundo o autor, essa precisão é considerada boa,

devido à subjetividade inerente à tarefa de identificação das relações multidocumento.

Especificamente sobre a identificação da complementaridade informacional via CST, há os trabalhos de Souza (2015) e Souza e Di-Felippo (2018). De acordo com os autores, tal fenômeno pode ser identificado com base em informações linguístico-estruturais, capturadas por pistas que evidenciam a complementação temporal entre as sentenças de um par. A análise foi traduzida em métodos de identificação (automática) dos tipos de complementaridade (temporal e atemporal) e das relações CST que a codificam, a saber: (i) distância entre as sentenças, (ii) sobreposição de nome/substantivo, (iii) ocorrência de advérbios temporais em S1, (iv) ocorrência de advérbios temporais em S2, (v) ocorrência de expressões temporais em S1, (vi) ocorrência de expressões temporais em S2, (vii) sobreposição de subtópicos, (viii) ocorrência de marcador discursivo em S1 e (ix) ocorrência de marcador discursivo em S2. Entretanto, esses estudos ainda se baseiam nos sinalizadores presentes na superfície textual das sentenças, além de, na maioria, se restringirem à presença e à ausência de informação temporal, como demonstrado em (2).

(2)

S1: A seleção brasileira de vôlei voltou a fazer bonito, desta vez na final da Liga Mundial, disputada contra a Rússia neste domingo no ginásio de Spodekna, em Katowice, na Polônia.

S2: Sua última derrota em finais da Liga Mundial, aliás, ocorreu em 2002, coincidentemente para a Rússia.

Em (2), narra-se sobre a participação da seleção brasileira na Liga Mundial de Vôlei. Na primeira sentença do par, aborda-se o desempenho do time da edição do evento daquele ano, disputado contra a Rússia, na Polônia; já a segunda informa a derrota do Brasil sobre a seleção russa, mas disputando a edição de 2002 do mesmo campeonato. Assim, S2 em relação a S1 apresenta uma informação complementar do tipo histórica sobre a participação do Brasil em uma edição anterior do evento esportivo.

Tendo em vista que os estudos realizados por Souza (2015) e Souza e Di-Felippo (2018) baseiam-se apenas em sinalizadores que

auxiliam a recuperação da informação complementar somente em traços presentes na superfície textual, propõe-se o refinamento da análise do fenômeno linguístico. Para tanto, baseando-se em Taboada e Das (2013) e Das e Taboada (2018) a fim de identificar sinalizadores que possam recuperar a complementaridade informacional, analisou-se o fenômeno em um *corpus* de estudo, construído a partir do CSTNews, considerando os princípios da LC. Ademais, como produto dessa análise, realizou-se a revisão manual da anotação da complementaridade nos pares de sentenças disponíveis no *corpus*, bem como a anotação manual dos sinalizadores de cada uma das relações que traduzem o fenômeno.

4 Análise da complementaridade no *corpus* cstnews

Para a realização deste estudo, selecionou-se o CSTNews (CARDOSO *et al.,* 2011; DIAS *et al.,* 2014). Como dito, o *corpus* está organizado em *clusters*, os quais representam as seções dos jornais *online* de onde os textos foram coletados, a saber "mundo", "política", "cotidiano", "ciência" e "esporte". Além dos textos-fonte (dois ou três), o *corpus* também contém sumários monodocumento e multidocumento de referência (manuais) e automáticos, alinhamento manual das sentenças dos sumários multidocumento às respectivas sentenças dos textos-fonte e uma série de camadas de anotações linguísticas. Dentre elas, estão: (i) anotação de relações semânticas multidocumento via CST; (ii) anotação de expressões temporais dos textos-fonte; (iii) etiquetação morfossintática (ou *tagging*); (iv) anotação dos sentidos dos substantivos e verbos; (v) anotação de aspectos informacionais nos sumários multidocumento (o quê, onde, quando, por exemplo), (vi) anotação automática dos textos-fonte via RST e (vii) anotação manual de subtópicos informativos em cada texto-fonte do *corpus*.

A anotação CST, em especial, foi realizada semiautomaticamente. Aleixo e Pardo (2008) revisaram o conjunto de rótulos das relações CST proposto para o inglês (ZHANG; GOLDENSHON; RADEV, 2002) e, a partir dessa revisão, decidiram aglutinar em um mesmo rótulo relações que apresentaram redundância entre si, e excluíram aquelas que não ocorreram nos textos do *corpus*.

Neste estudo, foram selecionados somente os pares de sentenças anotados com as relações que traduzem a complementaridade, a saber, *Historical Background*, *Follow-up* e *Elaboration*, representado por 73,

260 e 319 pares, respectivamente. Tendo em vista que no CSTNews há 713 pares de sentenças anotadas com as relações de complementaridade, construiu-se um *subcorpus* de estudo, que abrangeu os 10 primeiros *clusters* do *corpus*, resultando em 204 pares de sentença, sendo: (i) 12 pares anotados com a relação *Historical Background*, (ii) 94 com a relação *Follow-up* e (iii) 98 com *Elaboration*.

A respeito da complementaridade, Maziero (2012) e Maziero, Jorge e Pardo (2010) definem que o fenômeno ocorre pela relação que é estabelecida entre duas sentenças, S1 e S2, sendo cada uma delas provenientes de textos distintos; S2 deve apresentar a informação complementar em relação a algum elemento presente em S1. Admite-se ainda que as sentenças do par podem compartilhar conteúdo informacional, mas uma delas deve ter alguma informação aditiva que não esteja presente na outra.

Os autores compreendem a complementaridade em dois tipos. A do tipo temporal envolve sobreposição de conteúdo entre as sentenças de um par, sendo que S2 apresenta informação adicional baseada na informação temporal, a qual trata de um acontecimento anterior ou posterior ao evento principal descrito em S1. As relações *Historical-Background* e *Follow-up* traduzem esse tipo de complementaridade

Ainda segundo os autores, a complementaridade atemporal também se caracteriza pela sobreposição de conteúdo entre as sentenças de um par, sendo que uma das sentenças fornece informação adicional sobre o tópico principal. No entanto, o que a diferencia da complementaridade temporal é o fato de a informação adicional não ser de natureza temporal (MAZIERO, 2012; MAZIERO; JORGE; PARDO, 2010), e nem sempre ser marcada linguisticamente na superfície textual (SOUZA, 2015; SOUZA; DI-FELIPPO, 2018). A relação *Elaboration* compreende esse tipo de complementaridade.

4.1 Procedimentos metodológicos para levantamento de sinalizadores da complementaridade

Para levantar os sinais que evidenciam a complementaridade entre as sentenças de um par, optou-se por dois procedimentos metodológicos. Inicialmente, partiu-se do conjunto de sinais levantados por Souza (2015) e Souza e Di-Felippo (2018) e Taboada e Das (2013), assumindo que ambos os trabalhos, embora se ancorem em teorias linguísticas diferentes

(CST e RST), compartilham o pressuposto de que as relações, ora entre sentenças, ora entre proposições, são de natureza semântica.

O outro procedimento baseou-se na análise manual das sentenças já anotadas para identificar possíveis sinalizadores não previstos pelos trabalhos prévios, ou que não eram usualmente utilizados na descrição dos fenômenos multidocumento. Tal estudo consistiu em, dado um par de sentenças, (i) delimitar o trecho da sentença em que se manifestava a complementaridade e (ii) registrar os sinalizadores que auxiliam na recuperação da relação CST de complementaridade. Em (3) exemplifica-se tal procedimento.

(3)

S1: O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.

S2: <HB>A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002, após o incêndio de um trem que cobria o trajeto entre Cairo e Luxor, lotado de passageiros, e que deixou 376 mortos</HB>, segundo números oficiais.

Em (3) tem-se um par de sentenças que narra um acidente ferroviário que causou a morte de 57 pessoas, no Egito. A segunda sentença apresenta a informação histórica em relação à primeira, no trecho delimitado por "<HB>" e "</HB>"[1]. Após a identificação da informação complementar, observaram-se os possíveis traços que marcavam essa relação, tal como (a) *construção superlativa* (no caso, "a maior tragédia ferroviária da história do Egito"), (b) informação temporal marcada por uma *expressão temporal* (no caso, "em fevereiro de 2002"), (c) *discurso reportado* como estratégia sintática que compreende a informação complementar (d) *similaridade entre os eventos* narrados nas sentenças

[1] Durante a anotação foram utilizados delimitadores para identificar a informação complementar nos pares de sentenças e, posteriormente, em análises computacionais, dinamizar a recuperação automática dos trechos, já que as marcações foram feitas com base em XML. Assim, foram utilizadas as siglas HB, para *Historical Background*, FU, para *Follow-up* e ELAB, para *Elaboration*.

e (e) o *aspecto pontual* em S2, já que não se trata de um evento que se repete.

Dos traços levantados a partir da análise do exemplo em (3), (b) já havia sido identificado por Souza (2015) e Souza e Di-Felippo (2018) e (d), por Taboada e Das (2013), como marcas de fenômenos semânticos em suas respectivas teorias.

Ao se deparar com um novo sinalizador, verificava-se se ele já tinha ocorrido nos pares de sentenças anteriormente analisados, em quaisquer das relações de complementaridade no *corpus* de estudo. Especialmente para a relação *Historical Background*, esse procedimento estendeu-se a todos os pares anotados com esse rótulo devido à baixa ocorrência de pares de sentenças no CSTNews. Esse estudo durou cerca de oito meses.

Um aspecto relevante nesse procedimento metodológico é a identificação dos trechos das sentenças que continham a informação complementar. Até então, o CSTNews não apresentava essa delimitação. Tendo em vista que a identificação dos sinais das relações semânticas deve ser realizada a partir do processamento cognitivo, ou seja, mapeando as intenções do autor (DAS; TABOADA, 2018), ter os trechos delimitados auxilia esse tipo de análise, pois busca-se perceber o que motivou os anotadores a atribuírem dada relação às sentenças. Por conta disso, todos os trechos de complementaridade foram identificados previamente ao estudo.

4.2 Sinalizadores da complementaridade informacional multidocumento

Os sinalizadores apresentados nesta seção estão organizados em dois subgrupos: aqueles que discriminam as relações CST de complementaridade entre si, e aqueles que são capazes de auxiliar na recuperação da complementaridade, porém ocorrem em pelo menos duas das três relações.

4.2.1 Sinalizadores de Historical Background

a) *Expressões superlativas de comparação* – Esse tipo de construção frástica ocorre sempre em que há eventos relacionados por sucessivas repetições que "se superam". Em (4), a primeira sentença narra sobre uma indenização financeira que a Igreja

Católica americana pagou a vítimas de abuso sexual, enquanto a segunda narra sobre "*o maior pagamento* já feito pela Igreja Católica".

(4)

S1: A Igreja Católica chegou a um acordo financeiro estimado em US$ 660 milhões (aproximadamente R$ 1,2 bilhão) com mais de 500 pessoas que alegam ter sido vítimas de abuso sexual por padres em Los Angeles, nos Estados Unidos.

S2: <HB>Este seria o maior pagamento já feito pela Igreja desde que surgiu o escândalo de abuso sexual envolvendo religiosos em 2002 e elevaria o total de indenizações pago pela Igreja desde 1950, nos Estados Unidos, a US$ 2 bilhões (R$ 3,7 bilhões).</HB>

b) *Relação entre aspectos semânticos* – As sentenças de um par podem apresentar eventos de diferentes aspectos semânticos (pontuais ou habituais), como a queda de um avião em certa localidade e o fato desse evento estar relacionado à ocorrência habitual desse tipo de acidente naquele mesmo local. Em (5), em S1 há um aspecto pontual, já que o evento narrado ocorreu apenas uma vez, enquanto a informação veiculada em S2 é de aspecto habitual, recuperado pela expressão "*são frequentes*".

(5)

S1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um portavoz das Nações Unidas.

S2: <HB>Acidentes aéreos são frequentes no Congo,</HB> onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.

c) *Relação entre eventos similares não idênticos* – As sentenças de um par podem apresentar eventos semelhantes, mas jamais idênticos, tendo um intervalo temporal grande entre eles. Em (5), por exemplo, em que se narra sobre o acidente aéreo em Bukavu, em S1, e sobre a frequência de acidentes aéreos similares na região,

em S2, a distância temporal entre os eventos é considerável, tendo em vista que a informação veiculada em S2 ocorre após S1.

4.2.2 Sinalizadores de Follow-up

a) *Posterioridade entre eventos* – Há eventos que ocorrem em sucessão e são separados por um pequeno intervalo temporal. Em (6), S1 narra sobre uma jogada futebolística que ocorre "aos 27 minutos" do jogo, enquanto S2 apresenta outra jogada que acontece em seguida, "aos 31" minutos.

(6)

S1: Aos 27min, Kaká arrancou e chutou de fora da área.

S2: <FU>Kaká acertou um belíssimo chute de longe no ângulo aos 31 e fez 3 a 0.</FU>

b) *Previsão de eventos* – Nos pares de sentenças em que se notou este sinalizador, apresentaram-se eventos que ocorreram sequencialmente por meio da relação que se estabelece entre os tempos verbais das sentenças, de maneira que em S1 têm-se verbos no presente ("Lula *tem*...") e em S2, verbos flexionados no futuro do pretérito ("O presidente *teria...* "), como demonstrado em (7). Percebeu-se que esse tipo de sinalizador ocorre em *clusters* cujos assuntos principais são política, desastres naturais e esporte, já que é relevante informar sobre a possibilidade de um evento futuro.

(7)

S1: De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL).

S2: <FU>O presidente teria 53% das intenções de voto contra 30% de Heloísa.</FU>

c) *Efetivação de evento projetado* – Há casos em que se identificou a relação entre as sentenças do par por meio de previsões/possibilidades que se realizam/concretizam, como a possível

indicação de Solange Vieira a um cargo comissionado, em S1, e a efetivação dessa indicação, em S2, como demonstra-se em (8). Tal como a *Relação hipotética entre eventos*, este tipo de sinalizador é comum em *clusters* que abordam textos sobre política e esporte.

(8)

S1: O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac).

S2: <FU>O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac).</FU>

d) *Prolongamento do mesmo evento* – Há casos em que o evento descrito na segunda sentença é apenas a extensão do mesmo narrado na primeira. Em (9), narra-se sobre o acidente aéreo ocorrido no aeroporto de Congonhas; na primeira sentença, apresentam-se o plano de voo e alguns detalhes sobre o acidente, enquanto na segunda, a informação complementar à primeira centra-se em continuar narrando sobre o plano de voo e a possível causa do acidente.

(9)

S1: Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o voo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista.

S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, <FU>e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.</FU>

4.2.3 Sinalizadores de Elaboration

a) *Foco argumentativo distinto* – Há casos no *corpus* em que as sentenças do par apresentam focos argumentativos diferentes. O par de sentenças em (10) narra sobre uma reforma ocorrida em uma das pistas no aeroporto de Congonhas; a primeira sentença aborda o fato de não haver atraso nos voos internacionais, enquanto a segunda aponta que ocorreram atrasos entre partidas e chegadas de voos. Nesses casos, as informações não foram tidas como contraditórias, já que a relação não se constrói sob o questionamento das informações propositivas, mas sob apontamentos narrados por algum agente na primeira sentença.

(10)

S1: Nenhuma partida ou chegada internacional, segundo os painéis da Infraero, estavam fora do horário, o que não ocorria com os voos domésticos.

S2: <ELAB>As informações da Infraero não batem com as do painel das companhias aéreas, são 20 partidas atrasadas e 24 pousos atrasados.</ELAB>

b) *Informação adicional* – Há casos em que a complementaridade ocorre sob a narração de uma informação adicional na segunda sentença não prevista na primeira. O par de sentenças em (11) narra sobre a indicação nominal para ocupar o cargo de diretor da Agência Nacional de Aviação Civil; na primeira sentença, apresenta-se o fato de indicar uma economista (no caso, Solange Vieira – informação recuperável apenas a partir da leitura dos textos-fonte do *cluster*), enquanto a complementaridade na segunda se dá pela adição da informação da duração do mandato do cargo.

(11)

S1: Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da agência reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo.

S2: <ELAB>Como os diretores de agências têm mandato de cinco anos</ELAB>, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa.

4.2.4 Sinalizadores não discriminativos de relações CST de complementaridade

Como dito, há sinalizadores que auxiliam na recuperação da informação complementar entre as sentenças de um par, mas que pela ocorrência em mais de uma relação CST não foi possível classificá-los como discriminativos. Por conta do espaço dispensado neste texto, escolheu-se explanar a *categoria* dos sinalizadores presentes na Tabela 1. No entanto, ressalta-se que os sinalizadores *genéricos* e *específicos* estão detalhados em Souza (2019).

a) *Sinalizador do tipo Estrutural* – Este tipo de sinalizador permite que se possa recuperar a informação complementar somente a partir da leitura prévia dos textos que compõem o *cluster*, pois somente após esse procedimento, identificando as sentenças do par em seus respectivos textos-fonte que é possível constatar a complementaridade.

b) *Sinalizadores do tipo Referenciação* – Este tipo de sinalizador auxilia a identificação da informação complementar por meio da recuperação de algum referente na primeira sentença, em relação à segunda do par. Identificou-se que essa referenciação se deu por meio de Anáforas nominal e lexical.

c) *Sinalizadores do tipo Morfológico* – Estes sinalizadores recuperam a informação complementar por meio de marcas de tempo nos verbos (como "estavam na rua", em S1, e "está fazendo a vistoria", em S2), expressões nominais (como "novo bombardeio"), verbos de elocução (como "disseram" e "comunicaram") e diferenças numéricas que estabelecem adição de informação.

d) *Sinalizadores do tipo Sintático* – Estes sinalizadores identificam a complementaridade com base em adjuntos adverbiais ("como também", por exemplo), discurso reportado em que se marca a fonte da informação (como "os rebeldes afirmaram que..."), deslocamento de tema-rema entre as sentenças do par e orações aditivas, explicativas, objetivas direta e orações reduzidas.

e) *Sinalizadores do tipo Semântico* – Este tipo de sinalizador recupera a informação complementar por meio de palavras do mesmo campo semântico (como a relação que há entre "ataques", em S1, e "ameaça", "bombardeio" e "guerra", em S2), relações semânticas de causa-efeito, hiponímia e parte-todo, expressões temporais (como "Olimpíadas de Pequim") e itens lexicais que denotam acréscimo (como "acrescentou").

f) *Sinalizadores do tipo Pragmático* – Por fim, estes sinalizadores auxiliam na recuperação da complementaridade por meio de detalhamento ou conhecimento de mundo (como a relação que há entre "Companhia de Engenharia de Tráfego" e "São Paulo").

4.3 Organização tipológica dos sinalizadores de complementaridade

Com base no estudo realizado, foi possível organizar tipologicamente os sinalizadores descritos. A categorização foi feita posteriormente à análise, após a contabilização da ocorrência de cada um. Observou-se que havia regularidade entre os sinalizadores, permitindo propor categorias, que os organizassem em *genéricos* e *específicos*, resultando na organização demonstrada na Tabela 1.

Como já demonstrado, há sinalizadores específicos de cada relação CST de complementaridade, os quais desempenham papel essencial na caracterização das relações. Ademais, há sinalizadores que auxiliam na recuperação da complementaridade informacional, mas não são capazes de discriminar as relações entre si.

Com relação aos sinalizadores não específicos, observou-se que eles ocorreram ora entre dois tipos de relação CST (temporal e atemporal), ora entre duas relações do mesmo tipo. No primeiro caso, é possível cogitar que as fronteiras entre as relações não estavam bem claras para os anotadores do *corpus*, o que pode ser resultado da junção de rótulos que aconteceu para a adaptação do modelo CST para o português. O segundo caso indica que o sinalizador diferencia o tipo, mas não a relação, dando indícios que a descrição deve ser aprimorada, observando a correlação entre os sinalizadores para caracterização do tipo e da relação CST.

TABELA 1 – Tipologia de sinalizadores de complementaridade no *corpus* de estudo no CSTNews

Tipologia			Relação CST de complementaridade			Total
			Elaboration	*Follow-up*	*Historical Background*	
Categoria	Sinal genérico	Sinal Específico				
Referenciação	**Anáfora**	Anáfora Associativa	50	31	0	81
		Anáfora Nominal	132	89	20	241
--------	**Estrutural**	Leitura do Cluster	48	60	14	122
Morfológico	**Classe de palavras**	Numeral	11	35	2	48
		Expressão Nominal	2	15	0	17
		Expressão Preposicional	7	0	17	24
	Temporal	Tempo verbal	12	134	3	149
	Verbos de elocução	Verbos de elocução	26	51	0	77
Sintático	**Período simples**	Adjunto adverbial	31	40	2	73
		Expressão superlativa	0	0	26	26
	Período composto	Discurso reportado	67	52	0	119
		Oração aditiva	26	2	0	28
		Oração explicativa	37	5	7	49
		Oração objetiva direta	22	7	0	29
		Oração reduzida	12	3	0	15
	Deslocamento	Tema-Rema	108	1	2	111
Semântico	**Campo semântico**	Campo semântico	29	34	0	63
	Relações semânticas	Causa-Efeito	12	23	0	35
		Hiponímia	16	4	0	20
		Parte-Todo	42	15	0	57
	Temporal	Expressão temporal	4	109	57	170
	Sentido de acréscimo	Semântica Lexical	27	42	8	77
Pragmático	**Sobre o evento**	Detalhe	103	59	0	162
		Posterior	0	92	0	92
		Previsão	0	17	0	17
		Prolongamento	0	57	0	57
		Projeção	0	18	0	18
		Similaridade	0	0	39	39
	Argumentação	Foco argumentativo	17	0	0	17
	Suplementação	Informação adicional	52	0	0	52
	Aspectualidade	Fato pontual	0	0	38	38
		Frequência	0	0	38	38
	Conhecimento adicional	Conhecimento de Mundo	5	28	14	47

Fonte: Elaboração própria.

Com relação à ocorrência dos sinalizadores *genéricos* percebe-se que os sinalizadores do tipo *pragmático* são mais frequentes no *corpus* (577 ocorrências), seguido dos tipos *sintático* (450), *semântico* (422), *anafórico* (322), *morfológico* (315) e *estrutural* (122). Até então, os sinalizadores identificados por Souza (2015) estavam restritos à distinção do tipo de complementaridade, a saber, temporal e atemporal, compreendidos apenas nos tipos *morfológico* (tempo verbal) e *semântico* (expressão temporal).

Entretanto, após a análise do *corpus* de estudo, concluiu-se, como previsto, que a complementaridade informacional via modelo CST é mais bem compreendida por meio de sinalizadores do tipo *pragmático*. Alguns desses sinalizadores, como demonstrado, são capazes de caracterizar cada uma das relações CST de complementaridade, bem como auxiliar na recuperação da informação adicional entre as duas sentenças de um par. A não consideração desse tipo de sinalizador, é uma possível causa para que houvesse equívocos quanto à distinção das relações *Follow-up* e *Elaboration* em Souza (2015).

Outra conclusão possível de se apontar é uma possível reconsideração sobre a classificação da complementaridade proposta por Maziero (2012) e Maziero, Jorge e Pardo (2010). Os autores distinguem os tipos de complementaridade entre temporal e atemporal, considerando a presença ou a ausência de informação adicional baseada em aspectos temporais. No entanto, como observado na Tabela 1, a informação temporal é capturada somente por dois de sinalizadores específicos. Nesse sentido; todos os outros sinalizadores deveriam recuperar a complementaridade atemporal (logo, a relação *Elaboration*). No entanto, isso é inverdade, como é possível concluir, pois os do tipo *pragmático*, por exemplo, ocorrem mais frequentemente nas relações do tipo temporal. Assim, é possível que, futuramente, um novo estudo sobre a complementaridade informacional resulte em uma nova classificação do fenômeno via modelo CST, a qual não seja baseada no aspecto temporal presente ou ausente nas sentenças de um par, mas na informação pragmática veiculada pelas sentenças.

5 Considerações finais

Neste trabalho, aprofundou-se a descrição do fenômeno da complementaridade que ocorre em conjuntos de textos jornalísticos

que abordam um mesmo assunto. Especificamente, com base em estudo de *corpus*, estendeu-se a descrição já realizada por Souza (2015) que, ao se basear apenas em atributos restritos a informações temporais nas sentenças de um par, já havia obtido resultados bastantes satisfatórios.

Diferentemente das relações RST que são propositais, no modelo CST a informação complementar não intencional por parte dos autores dos textos, mas ocasionada a partir da anotação semântica das relações previstas no modelo. Nesse sentido, mais que determinar quais sinalizadores auxiliam na recuperação da complementaridade, enquanto fenômeno linguístico, eles delimitam o ponto de vista dos anotadores do *corpus* CSTNews. Embora essa consideração seja irrefutável, cabe destacar que a concordância medida entre os anotadores proporciona margem de confiança nos dados, já que seguiram uma metodologia de anotação semelhante à proposta por Hovy e Lavid (2010). Essa questão pode ser confirmada quanto à regularidade da ocorrência dos sinalizadores nas relações CST de complementaridade, o que determina que são características ora do fenômeno, ora das relações que o traduzem.

Com relação à frequência, percebem-se sinalizadores que tiveram baixa ocorrência nos pares de sentença. Em termos de descrição linguística, eles podem ser essenciais para a compreensão da manifestação da complementaridade em determinados contextos, como é o caso de *expressões nominais*, em *Follow-up*. Em termos de aplicação computacional, é possível que sinalizadores como esse sejam descartados das análises, por não serem considerados bastante robustos em relação aos outros.

Ao que se refere ao gênero textual, não é possível ser assertivo quanto à caracterização da complementaridade em relação a esse fator. Como descrito, o *corpus* de estudo é composto apenas por textos jornalísticos e, preliminarmente, tende-se a apontar que os sinalizadores são típicos desse gênero. Entretanto, será necessário, em trabalhos futuros, construir-se um *corpus* multidocumento de treinamento com variação de gênero textual, a fim de verificar a ocorrência dos sinalizadores propostos neste trabalho anotando novos textos. Isso permitirá determinar se os sinalizadores são restritos ao gênero textual, como também corroborar se são característicos do fenômeno da complementaridade.

Como proposto por Hovy e Lavid (2010), um dos próximos passos que deve ser seguido na anotação de *corpus* é a proposição de métodos automáticos para essa tarefa. Entretanto, é necessário destacar que o

estado da arte ainda não permite que os mesmos resultados apresentados aqui sejam alcançados, por serem derivados da análise manual do *corpus* de estudo. Além disso, os *parsers* disponíveis atualmente em PLN podem ser capazes de identificar os sinalizadores dos tipos *sintático*, *morfológico*, *referenciação*; entretanto, ainda não processam com alto desempenho os sinalizadores específicos dos tipos *pragmático* e *semântico*, os quais são mais proeminentes na complementaridade, além de serem discriminativos das relações CST entre si.

Por fim, como parte dos trabalhos futuros, pretende-se observar a ocorrência combinada dos sinalizadores de complementaridade com abordagens de AM. Além disso, pretende-se verificar se os sinalizadores aqui delimitados também têm potencial para subsidiar o aprimoramento da descrição de outras relações CST, como aquelas que traduzem a redundância e a contradição.

Agradecimentos

Em tempos em que a ciência é atacada, seus investimentos são cada vez mais limitados e professores são desvalorizados, é importante destacar o auxílio financeiro empenhado nesta pesquisa pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a orientação atenciosa da Profa. Dra. Ariani Di Felippo ao projeto de doutoramento, do qual se deriva este artigo: certamente o destino deste estudo poderia ter sido outro sem o financiamento e a dedicação de minha orientadora. Muito obrigado!

Referências

ALEIXO, P.; PARDO, T. A. S. CSTNews: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory). São Carlos: USP; UFSCar; UNESP, 2008. (Série Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional - NILC)

BIBER, D. Representatividade em planejamento de *corpus*. Tradução de Paula Marcolin. *Cadernos de Tradução*, Porto Alegre, v. 1, n. 30, p. 11-45, 2012.

CARDOSO, P. C. F.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. M. R.; DI-FELIPPO, A.; RINO, L. H. M.; NUNES, M. G. V.; PARDO, T. A. S. CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. *In:* RST BRAZILIAN MEETING, 3rd., 2011, Cuiabá. *Proceedings* […]. Cuiabá: SBC, 2011. p. 88-105.

CASELI, H. M. O uso de corpora paralelos para a criação de um tradutor automático estatístico. *In:* VIANA, V.; TAGNIN, S. E. O. *Corpora na Tradução*. São Paulo: HUB Editorial, 2015. p. 243-267.

DAS, D.; TABOADA, M. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, [*S.l.*], v. 52, n. 1, p. 149-184, 2018. DOI: https://doi.org/10.1007/s10579-017-9383-x

DAS, D.; TABOADA, M.; MCFETRIDGE, P. *RST Signalling Corpus*. Philadelphia: Linguistic Data Consortium, 2015.

DAVIES, M.; KIM, J. The Advantages and Challenges of 'Big Data': Insights from the 14 Billion Word iWeb Corpus. *Linguistic Research*, [*S.l.*], v. 36, p. 1-34, 2019. DOI: https://doi.org/10.17250/khisli.36.1.201903.001

DIAS, M. S.; GARAY, A. Y. B.; CHUMAN, C.; BARROS, C. D.; MAZIERO, E. G.; NOBREGA, F. A. A.; SOUZA, J. W. C.; CABEZUDO, M. A. S.; DELEGE, M.; JORGE, M. L. R. C.; SILVA, N. L.; CARDOSO, P. C. F.; BALAGE FILHO, P. P.; CONDORI, R. E. L.; MARCASSO, V.; DI-FELIPPO, A.; NUNES, M. D. G. V.; PARDO, T. A. S. Enriquecendo o *corpus* CSTNews: a criação de novos sumários multidocumento. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE – PROPOR, 2014, São Carlos. *Proceedings*… São Carlos: SBC, 2014. p. 239-243.

DI-FELIPPO, A.; SOUZA, J. W. C. O projeto do *corpus* para a construção de uma wordnet terminológica. In: PINTO, M. V.; SHEPHERD, T. M. G.; SARDINHA, T. B. (org.). *Caminhos da Linguística de Corpus*. Campinas: Mercado de Letras, 2012. p. 225-245.

HALL, M. *et al*. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, [*S.l.*], v. 11, n. 1, p. 10-18, 2009. doi: https://doi.org/10.1145/1656274.1656278

HOVY, E.; LAVID, J. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, [*S.l.*], v. 22, n. 1, p. 13-36, 2010.

KUHN, D.; ABARCA, E.; NUNES, M. G. Corpus Nilc de português escrito no Brasil. São Carlos: São Carlos: USP; UFSCar; UNESP, 2000. (Série Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional - NILC)

MANN, W. C.; THOMPSON, S. A. *Rhetorical Structure Theory*: A theory of Text Organization. Marina del Rey, CA: University of Southern California, Information Sciences Institute, 1987.

MAZIERO, E. G. *Identificação automática de relações multidocumento*. 2012. 118f. Tese (Doutorado em Ciências da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2012.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying Multi-Document Relations. *In:* INTERNATIONAL WORKSHOP ON NATURAL LANGUAGE PROCESSING AND COGNITIVE SCIENCE, 2010, Funchal. *Proceedings* [...]. Funchal: Polytechnic Institute of Setúbal, 2010. p. 60-69.

PEDRO, W. G.; VALE, O. A. ComentCorpus: o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um corpus opinativo. *In:* FINATTO, M. J. B.; REBECHI, T.; SARMENTO, S; BOCORNY, A.E.P. (org.). *Linguística de corpus*: perspectivas. Porto Alegre: Instituto de Letras da Universidade Federal do Rio Grande do Sul, 2018. p. 19-40.

RADEV, D. R. A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. *In:* SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1ST., 2000, Hong Kong. *Proceedings*... Hong Kong: Association for Computational Linguistics, 2000. p. 74-83. DOI: https://doi.org/10.3115/1117736.1117745

RODRIGUES, R. *Contribuições para um léxico-gramática das construções locativas do espanhol*. 2019. 174f. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2019.

SARDINHA, T. B. *Linguística de corpus.* Barueri: Editora Manole, 2004.

SARDINHA, T. B. Linguística de *corpus*: histórico e problemática. *Delta: Documentação de Estudos em Linguística Teórica e Aplicada*, São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005

SOUZA, J. W. C. *Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento*. 2019. 117f. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2019.

SOUZA, J. W. C. *Descrição linguística da complementaridade para a sumarização automática multidocumento*. 2015. 105f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2015.

SOUZA, J. W. C.; DI FELIPPO, A. Caracterização linguística da complementaridade: subsídios para Sumarização Automática Multidocumento. *ALFA: Revista de Linguística*, São Paulo, v. 62, n.1, p. 125-150, 2018. DOI: https://doi.org/10.1590/1981-5794-1804-6

TABOADA, M.; DAS, D. Annotation upon Annotation: Adding Signalling Information to a *Corpus* of Discourse Relations. *Dialogue Discourse*, [*S.l.*], v. 4, n. 2, p. 249-281, 2013. DOI: https://doi.org/10.5087/dad.2013.211

TAGNIN, S. E a Linguística de Corpus vai desbravando novos horizontes. *In*: FINATTO, M. J. B.; REBECHI, T.; SARMENTO, S; BOCORNY, A. E. P. (org.). *Linguística de corpus*: perspectivas. Porto Alegre: Instituto de Letras da Universidade Federal do Rio Grande do Sul, 2018. p. 11-15.

TAGNIN, S. E. O.; BEVILACQUA, C. Corpora *na Terminologia*. São Paulo: HUB Editorial, 2015.

VIANA, V.; TAGNIN, S. E. O. Corpora *no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2011.

ZHANG, Z.; GOLDENSHON, S. B.; RADEV, D. R. Towards CST-Enhanced Sumarization. *In:* NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-2002), 18th., 2002, Edmonton. *Proceedings* […]. Edmonton: AAAI, 2002. p. 439-445.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1089-1135, 2021

Pragmática de Corpus: o que é e onde estamos

Corpus Pragmatics: what it is and where we are now

Giovani Santos
Mary Immaculate College, University of Limerick, Limerick / Irlanda
giovani.santos@mic.ul.ie
http://orcid.org/0000-0003-4116-5613

Mateus Miranda
Mary Immaculate College, University of Limerick, Limerick / Irlanda
mateus.desouza@mic.ul.ie
http://orcid.org/0000-0003-2575-8769

Resumo: Este trabalho objetiva apresentar um novo campo que emergiu a partir da intersecção entre a Linguística de Corpus e a Pragmática: a Pragmática de Corpus. Para tanto, através de uma revisão da literatura como ponto de partida, traçamos um panorama que abarca a origem, os aspectos teórico-metodológicos, e os desafios da nova área. Ademais, introduzimos as abordagens forma-função e função-forma, dois modelos investigativos que integram a disciplina. Finalmente, por meio de um estudo de caso, a fim de ilustrar um dos possíveis percursos de análise, investigamos o marcador pragmático *kind of* por meio da filtragem, método que compõe a abordagem forma-função, no discurso oral de brasileiros universitários. Os subcorpora que subsidiam a pesquisa são o *Spoken Corpus of Brazilian Portuguese and L2-English* (SCoPE²) e o *Brazilian Spoken English Learner Corpus* (BraSEL). Os resultados apontam que quando usado pragmaticamente, mesmo em contextos linguísticos distintos, *kind of* ocorre em seus três domínios funcionais (atitudinal, interpessoal, textual) e como parte constituinte de marcadores de linguagem vaga.

Palavras-chave: pragmática de *corpus*; forma-função; função-forma; *kind of*.

Abstract: This work aims to present a new field which has emerged from the intersection between Corpus Linguistics and Pragmatics: Corpus Pragmatics. To do so, through a literature review as a starting point, we offer an overview that encompasses the origin,

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1089-1135

the theoretical and methodological aspects, and the challenges of the new field. In addition, we introduce the form-to-function and function-to-form approaches, two investigative models which integrate the discipline. Finally, by means of a case study in order to illustrate one of the possible analytical routes, we investigate the pragmatic marker *kind of* by employing sifting, a method which comprises the form-to-function approach, in the oral discourse of Brazilian university students. The subcorpora which support the research are the *Spoken Corpus of Brazilian Portuguese and L2-English* (SCoPE²) and the *Brazilian Spoken English Learner Corpus* (BraSEL). The results show that when used pragmatically, even in different linguistic contexts, *kind of* occurs in its three functional domains (attitudinal, interpersonal, textual) and as a constituent part of vague language markers.

Keywords: corpus pragmatics; form-to-function; function to form; kind of.

Submetido em 08 de outubro de 2020

Aceito em 14 de dezembro de 2020

1 Introdução

A linguagem para Firth (1957), segundo Sinclair (2004, p. 103), está atrelada ao contexto que integra fatores como a ação verbal e sofre influência das pessoas, coisas e eventos. Com base em postulados como os de Firth, a Linguística de Corpus (doravante LC), a qual conhecemos hoje, foi desenvolvida por neo-firthianos como Sinclair, que contribuíram para sua expansão, desenvolvendo estudos por meio da observação da linguagem em seu contexto real (McENERY; HARDIE, 2012). Tomando a definição de Sinclair (2005, p. 16), um corpus "é uma coleção de textos em formato eletrônico, selecionados de acordo com critérios externos para representar, o melhor possível, uma língua ou sua variação como fonte de dados para a pesquisa linguística".[1] Nos anos 1960, o *Brown Corpus*, em formato eletrônico e com um milhão de palavras, foi um marco na área, inspirando, posteriormente, a compilação de outros. Hoje, corpora com milhões de palavras não são incomuns, como o *Corpus of Contemporary American English* (COCA), o Corpus do Português NOW, o *British*

[1] Nossa tradução para: "[A corpus] is a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research."

National Corpus (BNC), entre outros.[2] A LC veio a se consolidar nos anos 1980, mas, se em uma perspectiva mundial podemos considerá-la uma disciplina jovem em relação a outras áreas da linguística, as primeiras publicações no Brasil só aconteceriam no final dos anos 1990 e início dos anos 2000 (BERBER SARDINHA, 1999, 2000).

Da mesma forma, Gomes de Matos observava em 1982 que os estudos da Pragmática ainda não estavam inclusos nas ementas dos cursos de pós-graduação no Brasil, e que as pesquisas nesse campo ainda eram incipientes (RAJAGOPALAN, 1999), apesar de o termo pragmática, em seu uso moderno, ser atribuído ao filósofo Charles Morris na década de 1930 (LEVINSON, 1983). Para Rajagopalan (1999, p. 323), a própria indefinição do termo pragmática era um fator dificultador para a impopularidade da área no país e também afetava pesquisadores pelo mundo. Em uma publicação mais recente, Rajagopalan (2016, p. 203) argumenta que os pragmaticistas ainda divergem no que tange às temáticas e prioridades de investigações da área e afirma que há uma enorme carência de obras sobre a Pragmática em língua portuguesa.

Destacamos ainda a interdisciplinaridade da LC que possibilita, por meio de corpora, investigações que contemplam o léxico, a sintaxe, o discurso, o ensino e a produção de materiais para a sala de aula de línguas, os estudos de registros e gêneros, os estudos literários e tradução, a sociolinguística, dentre outros (cf. O'KEEFFE; McCARTHY, 2010). Em oposição à crítica de que o marasmo se instalou no campo dos estudos linguísticos, Rajagopalan (2006, p. 160) faz menção à LC e sua interdisciplinaridade, com foco na lexicografia e na sintaxe, como uma área revolucionária que "questiona um dos postulados da linguística tradicional (teórica), com base em uma perspectiva de ordem eminentemente prática". Nesse sentido, o autor destaca que a lexicografia, vista como um exercício da semântica intencional, e norteada pelo racionalismo, não abre espaço para variações dialetais e geográficas de sentidos, ou para novas acepções da língua, o que contrasta com o interesse da LC em registrar os diferentes usos de cada item lexical em contextos variados.

Como vimos, tanto a LC quanto a Pragmática podem ser consideradas disciplinas relativamente novas e, de fato, foram por muito

[2] Cf. Tagnin (2010) para uma lista de corpora nas línguas alemã, espanhola, francesa, inglesa, italiana e portuguesa.

tempo exclusivas devido às suas diferenças metodológicas (ROMERO-TRILLO, 2008b). No entanto, recentemente, a partir de seu caráter interdisciplinar, a LC se fundiu com a Pragmática em uma nova subárea que vem a ser denominada Pragmática de Corpus e que pode caracterizar-se como um avanço para os estudos da linguagem.

Considerando o panorama da LC e da Pragmática no Brasil, e no espírito desta chamada para discorrer sobre as conquistas e desafios da LC, este trabalho se propõe a apresentar a Pragmática de Corpus (uma conquista mútua da LC e da Pragmática) e seu estado da arte. Para tanto, a estrutura deste artigo está disposta da seguinte forma: após esta introdução, perpassamos as Seções 2 e 3 que abordam, respectivamente, a LC e a Pragmática. O resultado da fusão entre as duas disciplinas, a Pragmática de Corpus, é descrito na Seção 4 a partir de seu percurso teórico e metodológico. Tal seção é ainda subdividida em três partes em que discorremos sobre as limitações e desenvolvimentos da nova área, além de focarmos em suas duas abordagens de análise: forma-função e função-forma. Na seção 5, complementamos, a fim de exemplificação, o trajeto das seções anteriores com um estudo de caso através de corpora orais. Por fim, tecemos na última seção as considerações finais.

2 Linguística de Corpus

Estudos no campo da LC têm aumentado de maneira consistente e substancial nos últimos 30 anos, e os resultados confirmam que a LC é um meio eficiente para se fazer análises da linguagem em uma vasta gama de contextos linguísticos (cf. BERBER SARDINHA, 2000; McCARTHY; O'KEEFFE, 2010 para uma perspectiva histórica).

Segundo McCarthy e O'Keeffe (2014, p. 271), "[a] evidência estatística e contextual que pode ser obtida através do uso de um software [de LC] nos permite fazer interpretações confiáveis das intenções comunicativas dos falantes e escritores."[3] Esta combinação estatística e contextual permite que pesquisadores analisem seus dados em termos qualitativos e quantitativos, através, por exemplo, do estudo empírico de listas de frequência e de concordância, reduzindo significativamente o risco de se introduzir quaisquer inclinações pré-concebidas.

[3] Nossa tradução para: "[the] statistical and contextual evidence the [CL] software can provide us with enables us to make more reliable interpretations of speakers' and writers' communicative purposes."

Validade e confiabilidade são, de fato, características representativas dos resultados obtidos pela LC, uma vez que as ferramentas principais aplicadas numa pesquisa baseada em corpus são computacionais, o que se opõe a possíveis limitações humanas. Em outras palavras, enquanto a análise manual é suscetível ao erro com relação às mais simples ocorrências de palavras e padrões linguísticos, abordagens baseadas em corpus oferecem ao analista a oportunidade de processar, de maneira meticulosa, grandes quantidades de dados. Tais processos não apenas têm o potencial de recuperar cada ocorrência de um item em questão, como também destacam estruturas complexas de padronização linguística e associações de palavras.

Desta forma, imparcialidade, precisão, processabilidade, e o acesso à linguagem natural (autêntica) em grande escala são benefícios essenciais que a LC oferece aos pesquisadores. Por conseguinte, muito mais informação e conhecimento sobre a forma como a linguagem opera pode ser revelado e descoberto atualmente, principalmente quando comparado com uma época anterior à LC. Biber *et al.* (1998, p. 5) observam que, uma vez propriamente utilizado, um corpus bem arquitetado e compilado pode fornecer maior variedade de informações sobre a língua em seu uso real. De fato, questões linguísticas que preliminarmente pareciam quase impossíveis de serem abordadas, ou que ainda não haviam sido sequer consideradas, foram evidenciadas e continuam a ser investigadas com a ajuda das ferramentas da LC e da disponibilidade de muitos corpora gratuitamente acessíveis.

Em um relato sobre o processo analítico e técnico de um grande e influente projeto baseado em corpus, realizado no início dos anos 1980, Sinclair (1991, p. 2) refere-se ao fato de que como língua mais descrita no mundo, eram mínimas as chances de que a língua inglesa apresentasse quaisquer tipos de novas evidências reveladoras que, por séculos, ainda não haviam sido observadas. O projeto foi, inicialmente, desenvolvido para ajudar na construção de um dicionário inovador, nomeado *COBUILD Dictionary* (SINCLAIR *et al.,* 1987). Tal projeto necessitou do desenho e desenvolvimento de um grandioso corpus da língua inglesa, como também de novas técnicas para auxiliar na observação e análise da linguagem em uso. Conhecido pelo mesmo nome do dicionário, o projeto não só lançou uma nova abordagem para a construção de dicionários, como também foi fundamental para a introdução de uma nova perspectiva para a descrição da linguagem.

Uma das mais importantes contribuições que resultaram do projeto COBUILD segue na forma do trabalho de Sinclair, mais notavelmente em Sinclair (1991), no qual o princípio idiomático é proposto e delineado. Tal princípio estabelece que os usuários de uma língua nem sempre selecionam suas palavras de maneira em que espaços sintáticos são preenchidos com escolhas abertas. Ao invés disso, falantes e escritores são mais propensos a formar o seu discurso e textos usando porções de linguagem semiconstruídas que estão disponíveis no sistema linguístico e que variam em diferentes níveis de flexibilidade. Além da questão de a linguagem autêntica mostrar-se através de construções pré-fabricadas, ou padrões linguísticos, outra valiosa inferência resultante do princípio idiomático, e importante para a observação e descrição da língua, é o conceito de unidades de significado. Nas palavras de Sinclair (2003, p. 3), "[s]e estudarmos casos de uso real, descobrimos que palavras e frases adjacentes ajudam muito na determinação do significado."[4] Ou seja, o significado de muitas palavras, se não da maioria, é determinado pela sua interação e correlação com outras palavras (cotexto), sugerindo que as palavras se selecionam e que tal performance possui uma relação clara com o significado.

O projeto COBUILD documenta um exemplo prático e significativo do quão favorável a LC pode ser ao estudo da linguagem natural empregada em contextos reais. Desde então, os métodos e as técnicas da LC têm sido aprimorados, desenvolvidos, e de grande influência, além de aplicados a muitas outras subáreas da linguística, incluindo os campos da linguagem falada e da Pragmática.

É evidente que os estudos do discurso oral têm se beneficiado grandemente com os avanços da LC nos últimos anos (CAINES *et al*., 2016; LOVE, 2020). Nesse sentido, podemos afirmar que outra contribuição considerável da LC foi o seu auxílio na investigação e exposição de características importantes da língua oral que, do contrário, não teriam sido facilmente acessadas e descobertas – especialmente com o uso de dados fabricados. Embora ainda seja uma tarefa desafiadora, devido a questões de disponibilidade de participantes e do trabalho transcritório, a construção de corpora orais tem sido facilitada consideravelmente pelos avanços tecnológicos. Consequentemente, pesquisadores hoje têm

[4] Nossa tradução para: "[i]f we study instances of usage, we find that the surrounding words and phrases help a lot in determining the meaning."

maior facilidade de acesso a dados linguísticos autênticos do discurso oral e, portanto, conseguem sistematicamente extrair dados confiáveis para a investigação da singularidade das interações desta modalidade. Como Sinclair (1991, p. 4) adequadamente aponta, "a capacidade de se examinar grandes corpora de maneira sistemática nos permite acesso a evidência de uma qualidade antes não disponível".[5] Tal afirmação é particularmente verdadeira com respeito aos corpora de fala, uma vez que eles oferecem ao pesquisador características distintivas em linguagem natural que representam interações reais entre interlocutores.

Nos últimos anos, tem havido uma crescente conscientização em relação à importância do corpus de fala para os estudos linguísticos, uma vez que este tem o potencial de revelar e destacar características linguísticas que são peculiares deste contexto comunicativo. Ademais, apesar de tanto a modalidade escrita quanto a oral proporcionarem ricos contextos linguísticos para análises de fenômenos pragmáticos, é na interação oral que se encontra um terreno ainda mais fértil para o surgimento destes fenômenos, visto que esta requer que os interlocutores negociem a construção de conteúdo e sentido em tempo real. Nesse contexto, é plausível argumentar que a LC é um complemento vantajoso tanto para o estudo da língua oral quanto da pragmática. Certamente a LC tem sido acolhida como um complemento valioso para muitas áreas dos estudos linguísticos, e uma área que parece ter encontrado um aliado perfeito na LC é exatamente a do campo da Pragmática, que, juntas, provocaram o surgimento de um campo de estudo relativamente recente: a Pragmática de Corpus (doravante PC). Antes de introduzirmos tal fusão, contudo, faz-se necessária a apresentação de um breve panorama sobre a Pragmática.

3 Pragmática

Como mencionado na Seção 1, definir pragmática não é tarefa fácil. Em seu nível mais fundamental, podemos definir a pragmática como o estudo da língua em uso real, e que considera as relações entre contexto de uso e sentido intencionado. O termo pragmática, como conhecemos hoje, é atribuído a Morris (1938), a partir da obra *Fundamentos de uma*

[5] Nossa tradução para: "the ability to examine large text corpora in a systematic manner allows access to a quality of evidence that has not been available before."

teoria dos signos que dividiu a linguagem em três planos: o sintático, o semântico e o pragmático. Por um lado, Morris define a área da linguística que lida com o contexto e aponta seu vínculo com outras áreas (*e.g.* sintaxe e semântica), mas, por outro, apresenta uma definição limitada de contexto, ignorando fatores importantes como as relações sociais e situacionais (CULPEPER; HAUGH, 2014). Para Culpeper e Haugh (2014, p. 6), apesar de indicarem que uma divisão anglo-americana e continental da pragmática não deve ser enfatizada, Morris propõe uma visão micro do contexto, pois parte de uma visão anglo-americana, que também é o fundamento para outros trabalhos, como as implicaturas conversacionais de Grice (1975) e a Teoria da Relevância de Sperber e Wilson (1995).

A outra visão da pragmática, denominada continental, relaciona-se para além da linguística, com outras áreas cognitivas, sociais e culturais, através da linguagem em uso e do comportamento humano em contextos sociais (CULPEPER; HAUGH, 2014). Dessa forma, tendo em vista o exposto acima e o escopo deste trabalho, não se pretende aqui explorar a definição de pragmática de forma exaustiva. No entanto, a definição de Yule (1996) nos parece ir de encontro ao propósito da PC.[6] Segundo o autor:

> A pragmática é o estudo da relação entre as formas linguísticas e os usuários dessas formas. [...] A vantagem de estudar a linguagem por meio da pragmática é que se pode falar sobre as intenções das pessoas, suas suposições, seus propósitos ou objetivos, e os tipos de ações (por exemplo, pedidos) que realizam quando falam[7]. (YULE, 1996, p. 4).

Ademais, Yule (1996, p. 4) nota que uma das vantagens de considerar a pragmática como objeto de estudo é a possibilidade de investigar o que ele denomina *conceitos humanos*, que incluem os

---

[6] Destacamos que o silêncio, a prosódia e a linguagem corporal também são fenômenos pragmáticos que auxiliam na interpretação do significado e, especialmente em corpora multimodais, são também relevantes para os propósitos da PC.

[7] Nossa tradução para: "Pragmatics is the study of the relationship between linguistic forms and the users of those forms. [...] The advantage of studying language via pragmatics is that one can talk about people's intended meanings, their assumptions, their purposes or goals, and the kinds of actions (for example, request) that they are performing when they speak".

significados, as suposições, os propósitos, os objetivos e as ações dos indivíduos, mesmo apontando que as análises de tais conceitos são extremamente difíceis de serem conduzidas. Todavia, ao discorrer sobre a relação entre a LC e a Pragmática, Raso (2016) reforça que por meio da contribuição empírica da LC, todos os temas da Pragmática são passíveis de investigação. Dessa forma, Rühlemann e Clancy (2018) confirmam que se a pragmática geralmente lida com pequenas quantidades de textos em determinados contextos, através da LC pode-se aplicar tais análises a um volume maior de dados. Nesse sentido, a seção que segue tem como objetivo apresentar a origem da PC através dos princípios que a norteiam, destacando suas limitações, desenvolvimentos e as duas abordagens que a compõem: forma-função e função-forma.

4 Pragmática de Corpus

A PC pode ser definida como o campo linguístico de investigação da linguagem autêntica e em uso real com o auxílio de corpora, com vista à interpretação contextual da linguagem escrita ou falada. Tal campo linguístico dá-se pela intersecção entre os campos da LC e da Pragmática (RÜHLEMANN; AIJMER, 2015) e, embora o termo em si seja de cunhagem recente, esta é uma junção que tem evoluído consideravelmente na última década (cf. ROMERO-TRILLO, 2008b para uma introdução à trajetória científica que uniu as duas áreas de conhecimento).

Estudos de fenômenos pragmáticos baseados em corpora têm sido realizados desde os anos 90 (JUCKER; TAAVITSAINEN, 2014), com o interesse em tal abordagem aumentando gradativamente através dos anos e afirmado, posteriormente, com a publicação de um volume dedicado à LC no '*Journal of Pragmatics'* em 2004, com a edição da Conferência da IPrA (*International Pragmatics Association*) em 2007, também com foco na LC, e com a edição da Conferência ICAME (*International Computer Archive of Modern and Medieval English*) com foco em pragmática e discurso em 2008. Contudo, foi com a influente publicação de Romero-Trillo (2008a) que a atenção se voltou para o fato de que há um relacionamento de interesses mútuos entre a LC e a Pragmática, o que poderia ser proveitosamente explorado quando estas duas disciplinas eram fundidas. *Pragmatics and Corpus Linguistics: a mutualistic entente* (ROMERO-TRILLO, 2008a) é o primeiro livro

que agrega pesquisadores que combinam as metodologias de ambas as disciplinas para a investigação de diversas questões linguísticas. Desde então, a produção acadêmica na área da PC tem aumentado exponencialmente, e firmado a disciplina como uma abordagem eficiente para a investigação e compreensão do uso de diversos recursos linguísticos em contextos reais. Tal afirmação é evidenciada pela abundante literatura disponível, na qual inclui uma série anual de livros lançada em 2013 (*Yearbook of Corpus Linguistics and Pragmatics*), livros de contribuições organizadas (AIJMER; RÜHLEMANN, 2015; JUCKER *et al*., 2009; TAAVITSAINEN *et al*., 2014), um guia de estudo e pesquisa (RÜHLEMANN, 2019), e um periódico recentemente lançado e exclusivamente dedicado à PC (*Corpus Pragmatics*).

No que concerne aos tipos de estudos que são abordados pela PC, Aijmer e Rühlemann (2015) demonstram que essa disciplina não se limita à investigação de apenas alguns fenômenos pragmáticos como, por exemplo, os marcadores pragmáticos – provavelmente o fenômeno mais estudado na PC – mas também abrange temas centrais da Pragmática, como relevância, dêixis, processabilidade e atos da fala. O livro foi organizado para ser uma referência, assim como uma contribuição para o crescimento da disciplina, e é dividido em seis áreas nucleares da Pragmática: atos da fala; princípios pragmáticos; marcadores pragmáticos; avaliação; referência; e tomada de turno. Por sua vez, Clancy e O'Keeffe (2015) oferecem uma discussão crítica sobre os temas que são mais investigados dentro da PC, sendo eles: atos da fala; marcadores pragmáticos; linguagem, poder e ideologia; a organização de discurso; dêixis. Além de apresentarem uma revisão da literatura e exemplificação de estudos, os autores destacam que a PC é uma área que está "madura e carregada de oportunidades de pesquisa"[8] (CLANCY; O'KEEFFE, 2015, p. 236)

Embora a LC e a Pragmática sejam hoje reconhecidas como mutuamente favoráveis (CLANCY; O'KEEFFE, 2015; ROMERO-TRILLO, 2008b), Romero-Trillo (2008b) observa que estes campos representaram, por muitos anos, duas linhas de pensamento paralelas, porém mutuamente exclusivas e excludentes. Isto porque a LC, de um lado, trabalha com grandes quantidades de textos e adota uma abordagem mecânica, matemática e mais estrita; enquanto, por outro lado, o campo da

[8] Nossa tradução para: "[…] the area is ripe with research opportunities."

Pragmática trabalha com estudos em menor escala (um pequeno número de textos), abordando a língua de uma perspectiva contextual que permite uma visão mais flexível, inferencial e interpretativa da linguagem em uso através de diferentes métodos (cf. JUCKER *et al.*, 2018). Em outras palavras, enquanto a LC é essencialmente quantitativa (embora permita também espaço para uma análise qualitativa) e inicialmente preocupada com número de frequência e medidas estatísticas, a Pragmática é fundamentalmente qualitativa e preocupada com a interpretação do significado no contexto de uso (RÜHLEMANN; AIJMER, 2015; RÜHLEMANN; CLANCY, 2018).

Contudo, pode-se dizer que, apesar de metodologicamente distintas e opostas, e, portanto, aparentemente impossíveis de se relacionar, a LC e a Pragmática compartilham ao menos duas características que permitiram a exploração de um terreno em comum, assim corroborando a fusão de tais disciplinas. A primeira característica diz respeito ao fato de que, historicamente, ambos os campos acolheram outras metodologias e quadros teóricos para substanciar e fortalecer suas próprias abordagens. O'Keeffe *et al*. (2011, p. 20) constatam que a Pragmática, como um quadro teórico de investigação da produção linguística e interação, não é em si uma metodologia, sendo então combinada muitas vezes com uma gama de métodos em pesquisas focadas em fenômenos pragmáticos. Em contraste, Rühlemann e Clancy (2018, p. 242) observam que apesar de os corpora (particularmente os megacorpora) terem sido originalmente compilados com propósitos lexicográficos em mente, seu crescimento e desenvolvimento os levaram à combinação de muitos quadros teóricos linguísticos com uma metodologia de LC. Em concordância com esta visão, para Biber *et al*. (1998, p. 9), uma investigação baseada em corpus não deveria ser considerada como uma abordagem exclusiva e distinta, mas como uma abordagem que complementa outras mais tradicionais. Por sua vez, a segunda característica compartilhada é a de que ambas as disciplinas têm como pressuposto básico o entendimento de que forma e função não estão intrinsecamente correlacionadas. Em outras palavras, o significado e a função de uma forma linguística são recuperados e interpretados através de sua relação com o cotexto (LC) e com o contexto (Pragmática).

O relacionamento complementar entre a LC e a Pragmática é essencialmente simbiótico. Isto porque a Pragmática oferece meios para a interpretação contextual da riqueza de descobertas resultantes

dos métodos da LC. Por outro lado, enquanto interessada em dados reais e no significado criado na interação comunicativa, a Pragmática também se beneficia da LC através da disponibilidade de uma grande quantidade de produção linguística de ocorrência natural. Além do mais, a metodologia da LC tem o potencial de recuperar com sucesso itens pragmáticos específicos e/ou destacar padrões pragmáticos que seriam, de outra forma, difíceis de iluminar sem o auxílio dos instrumentos da LC (O'KEEFFE *et al*., 2011).

Portanto, por essas razões, podemos argumentar que a LC e a Pragmática têm uma tendência natural a se gravitarem. Como assinalado por Romero-Trillo (2008b, p. 5, itálico dos autores), "a linguística de corpus e a pragmática são *duas versões do mesmo fenômeno*: a mecânica – material – (estudos baseados em corpus), e a sua interpretação e explicação (pragmática)."[9]

Como um quadro teórico-metodológico, a PC une os princípios fundamentais e as metodologias das áreas do conhecimento que a deram origem. Segundo Rühlemann e Aijmer (2015, p. 12), a PC tem uma metodologia de leitura integrada (FIGURA 1), no sentido de que integra a metodologia de análise tipicamente qualitativa e horizontal da pragmática com a metodologia predominantemente quantitativa e vertical da LC.

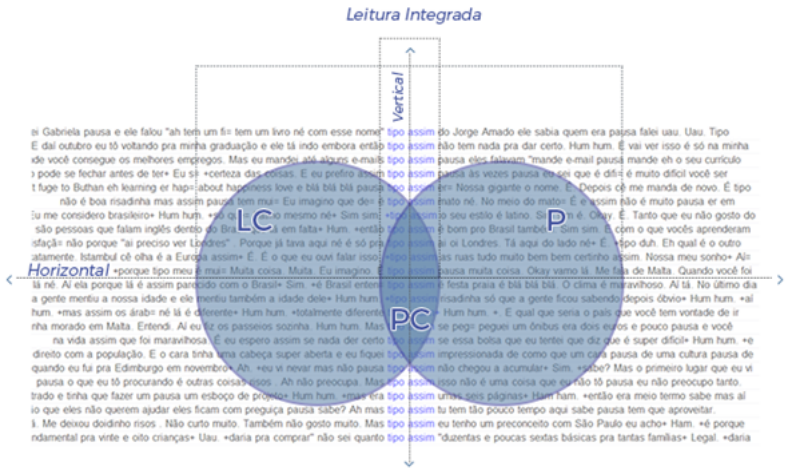
Em relação à metodologia vertical, os autores referem-se ao nódulo (ou item de busca), que toma a posição central e vertical nas listas de concordância, e sua relação colocacional com as palavras à sua direita e esquerda. A leitura vertical do corpus oferece ao analista uma hipótese inicial com base na posição do item investigado e seu ambiente linguístico, como também na preferência lexical (ou padrões de colocação e coligação) do mesmo. Ademais, é durante esse processo de leitura que aquele de posse de um corpus simples (um corpus que não foi etiquetado) filtra os dados para que usos pragmáticos de um item específico sejam separados daqueles não pragmáticos.

Em contrapartida, a metodologia de leitura horizontal permite uma investigação mais aprofundada e minuciosa pela qual o analista leva em consideração contextos maiores do que apenas as palavras e frases imediatamente ao redor do nódulo. Assim, não apenas o cotexto imediato

[9] Nossa tradução para: "corpus linguistics and pragmatics are two versions of the same phenomenon: the mechanics – the subject-matter – (corpus studies), and its interpretation and explanation (pragmatics)."

é levado em consideração, como também a leitura do contexto expandido, para além da linha visualizada no concordanciador de uma dada busca.

FIGURA 1 – Metodologia integrada de leitura da Pragmática de Corpus



Fonte: Adaptada de Rühlemann e Clancy (2018, p. 245)

A integração das leituras vertical e horizontal na PC cria um processo analítico de dupla direcionalidade que tem um grande potencial para a investigação de material linguístico autêntico. Em síntese, esta sinergia mutualista entre a LC e a Pragmática faz da PC um quadro teórico-metodológico significativo para a análise fidedigna e empírica da linguagem usada e desenvolvida (*e.g.* desenvolvimento de L2) em contextos reais. Dentro deste quadro teórico-metodológico, o analista pode conduzir análises estatísticas enquanto, concomitantemente, realiza uma interpretação detalhada dos dados com referência a informações contextuais importantes que podem referir-se não apenas ao cotexto, como também à situação na qual a comunicação ocorre e ao conhecimento prévio entre interlocutores.

4.1. Limitações, oportunidades e desenvolvimento da Pragmática de Corpus

Não é sem limitações e desafios que a PC, como uma disciplina emergente, posiciona-se como uma subárea da Linguística. Romero-Trillo (2008b), já no princípio, notara que uma das desvantagens em conduzir estudos pragmáticos com uma metodologia de LC é que, muitas

vezes, os corpora eram desprovidos de informações de contextualização, como o status socioeconômico e cultural dos participantes e o tipo de relacionamento e contexto situacional das interações comunicativas. Tal incapacidade, contudo, já está solucionada com a adição de metadados durante o processo de compilação e construção dos corpora. O acesso ao contexto dá-se ainda com maior precisão em casos de corpora menores,[10] em que o analista e o compilador são um e o mesmo, o que permite uma perspectiva interna de análise, sustentando as condições de interpretação de dados. Este relacionamento próximo entre o analista e os dados, como também entre linguagem e contexto, faz dos pequenos corpora a opção perfeita para os estudos pragmáticos. De fato, pesquisadores que têm se debruçado sobre os estudos e discussões que permeiam o desenvolvimento de um modelo científico para a PC defendem e promovem o uso de pequenos corpora para a investigação de fenômenos pragmáticos (CLANCY; O'KEEFFE, 2015; RÜHLEMANN; CLANCY, 2018; VAUGHAN; CLANCY, 2013).

Isso não significa, contudo, que estudos baseados em PC podem apenas ser realizados quando pequenos corpora são utilizados. Grandes e megacorpora podem ser, e já foram, utilizados em pesquisas de PC (cf. VAUGHAN; CLANCY, 2013; O'KEEFFE *et al*., 2020 para a descrição de alguns exemplos e soluções). Entretanto, quando se trata de análise qualitativa, é comum que os pesquisadores tenham de reduzir a amostra a fim de avaliar e analisar as ocorrências linguísticas com a perspectiva pragmática.

O'Keeffe (2018) aponta para um outro desafio que, por sua vez, afeta diretamente um dos temas mais espinhosos no desenvolvimento de um modelo de PC: a questão da forma e função. Vejamos que tal desafio se dá pela oposição metodológica entre a LC e a Pragmática. De um lado, temos a LC que possui uma metodologia de análise ascendente, ou seja, de baixo para cima (*bottom-up*) a partir da forma. Para isso,

[10] O conceito de pequeno corpus é abstrato, uma vez que ainda não há uma concordância entre os linguistas de corpus a esse respeito. Pode-se dizer que um pequeno corpus é um que contenha menos que um milhão de palavras. Contudo, no âmbito da Pragmática de Corpus, um pequeno corpus que permita etiquetamento manual, processamento e análise de dados, varia entre 50 e 500 mil palavras. Há ainda estudos pragmáticos de corpus que utilizam corpora menores que 50 mil palavras, como Vaughan e Clancy (2013) e McAllister (2015).

a investigação ocorre a partir da observação do corpus por meio de percursos metodológicos específicos, como a extração de listas de palavras classificadas por ordem de frequência, listas de palavras-chave com base em um corpus de referência, agrupamentos lexicais e linhas de concordância. Nesse sentido, "os estudos pragmáticos baseados em corpora são geralmente baseados na forma e começam mapeando palavras ou construções em uma série de funções'[11] (AIJMER, 2018, p. 555). No entanto, existem aspectos negativos desta abordagem para a análise pragmática, atrelados à diversidade funcional e ambiguidade (O'KEEFFE, 2018). Isto porque em uma abordagem que toma a forma como o ponto de partida para, então, conduzir uma análise das funções de dado item linguístico (abordagem forma-função), nos deparamos com o desafio de filtrar as funções pragmáticas e não pragmáticas que este performa no discurso. Durante o processo de filtragem, há ainda a possibilidade de os resultados apresentarem ocorrências ambíguas, como também aquelas que demonstram uma diversa seleção de funções pragmáticas para o mesmo item.

Por outro lado, temos a Pragmática com uma metodologia tradicionalmente descendente, ou seja, de cima para baixo (*top-down*) a partir da função. Para tanto, a investigação inicia-se por uma determinada função (*e.g.* atos de fala) e caminha em direção às formas que a desempenham. Este modelo de abordagem efetua-se através de instrumentos mais diretos, partindo da coleta de produção elicitada, em contexto controlado, e por meio de técnicas como entrevistas, testes de complementação discursiva (TCDs), *role-plays*, diários, entre outros. Como ponderam O'Keeffe *et al*. (2020), é inegável que o campo da Pragmática possui abordagens tradicionais de investigação já estabelecidas e, consequentemente, a transposição da metodologia da LC para a pragmática apresenta desafios. Isto porque uma função pragmática pode apresentar inúmeras formas linguísticas. Ou seja, em uma abordagem que toma a função como ponto de partida para, então, recuperar as formas linguísticas que a desempenham (abordagem função-forma), o pesquisador enfrentará a dificuldade de não extrair todas as formas de uma função, à exceção de casos em que o corpus é manual e pragmaticamente anotado.

[11] Nossa tradução para: "Corpus-based pragmatic studies are generally form-based and they start by mapping words or constructions onto a range of functions."

Podemos, assim, considerar que as observações descritas acima são limitações metodológicas em ambas as áreas. Nesse panorama, como já antecipado, dentro de um modelo para a PC, precisamos encontrar um equilíbrio entre as duas abordagens, de maneira que o melhor que cada uma tem a oferecer para os estudos pragmáticos baseados em corpora possa ser utilizado na investigação de quaisquer perguntas norteadoras a serem abordadas dentro de um quadro teórico-metodológico de PC (O'KEEFFE *et al*., 2020).

Adiante, apresentamos ambas as abordagens para os estudos da PC, forma-função e função-forma, bem como seus modelos metodológicos que, junto ao modelo de leitura integrada de Aijmer e Rühlemann (2015) previamente apresentado, constroem o quadro metodológico da nova disciplina.

4.1.1 Abordagem forma-função

A abordagem forma-função na PC é herança da LC, e tem servido bem aos estudos pragmáticos com base em corpora, cujos interesses são o mapeamento das funções pragmáticas desempenhadas por específicas formas linguísticas. Dentro de um modelo metodológico de PC, a abordagem forma-função é ideal para a investigação de fenômenos pragmáticos cujas formas são bem definidas e menos variáveis do que, por exemplo, os atos da fala. Fenômenos pragmáticos que se adaptam bem a esta abordagem incluem dêixis, marcadores pragmáticos (abrangendo todas as subcategorias deste grupo, como marcadores discursivos, marcadores de hesitação, tokens de resposta, mitigadores, etc.) e itens lexicogramaticais de vagueza.

O início da PC deu-se, majoritariamente, com a utilização da abordagem forma-função e, ainda hoje, estudos pragmáticos com base em corpora são geralmente abordados tendo a forma como o ponto de partida (AIJMER, 2018; JUCKER 2013). Isto se justifica pelo fato de a LC oferecer uma metodologia de busca e análise linguística rigorosa e já bem definida, portanto, mais proeminente e dominante, como também pelo fato de a LC encontrar certas dificuldades quando a abordagem inicia-se de maneira inversa, i.e. função-forma (Seção 4.1.2 abaixo). Aijmer (2018, p. 555) argumenta, contudo, que uma grande vantagem em conduzir uma análise forma-função dentro de um modelo de PC é a de que uma dada forma sob investigação pode ser estudada em detalhe e precisão, levando-se em consideração aspectos como frequência,

posição sintática, prosódia semântica, colocação e coligação. Ademais, o fato de que múltiplas funções pragmáticas de uma determinada forma linguística podem ser evidenciadas, avaliadas e categorizadas, faz da PC uma ferramenta que oferece uma nova e significativa perspectiva de investigação aos estudos de fenômenos pragmáticos.

Neste contexto, O'Keeffe *et al*. (2020) apresentam quatro modelos de estratégias investigativas que compõem a abordagem forma-função dentro do quadro metodológico da PC. Os modelos foram adaptados daqueles propostos por Ädel e Reppen (2008) para os estudos baseados em corpus de maneira geral. São eles: (a) *busca direta* (*one-to-one searching*), (b) *amostragem* (*sampling*), (c) *filtragem* (*sifting*), e (d) *listagem baseada em frequência* (*frequency-based listing*). Como é apontado pelos autores na proposta original, e reiterado pelos autores da recente adaptação para a PC, estes modelos mesclam-se com frequência, ou seja, uma pesquisa pode implementar apenas um modelo como também uma combinação de dois ou mais modelos.

A **busca direta** é aquela na qual o pesquisador insere uma forma para busca e recupera cem por cento de ocorrências relevantes. Contudo, tal abordagem só é possível uma vez que o corpus tenha sido previamente etiquetado. Por exemplo, consideremos um pesquisador que se interesse pelo vocativo *querido*(*a*) (*e.g. querido, não faça isso*). Para a recuperação total de ocorrências relevantes, o corpus necessita de anotação pragmática para que ocorrências do adjetivo ou substantivo *querido*(*a*) (*e.g. amigo querido; aproveitaram para cantar parabéns para o querido*) sejam automaticamente excluídas da busca e, assim, o pesquisador tenha em mãos apenas uma lista de ocorrências do fenômeno de seu interesse. Com todas as ocorrências pragmáticas da forma *querido*(*a*) recuperadas e isoladas automaticamente, o pesquisador segue para a análise qualitativa e contextual das suas funções.

Por sua vez, o modelo de **amostragem** diz respeito àquela abordagem em que o pesquisador criteriosamente seleciona um ou mais itens de busca que representam um certo fenômeno pragmático. Ädel e Reppen (2008, p. 3) alertam para o fato de que há uma desvantagem na utilização deste modelo investigativo: apenas um subconjunto de casos de determinado fenômeno é investigado, e não o fenômeno em sua totalidade. Ainda assim, a partir de uma cuidadosa avaliação e desenho de um quadro de amostragem, o emprego de itens de busca representativos tem grande potencial para a recuperação de ocorrências relevantes. Os

itens de busca representativos de um fenômeno pragmático podem ou não ser linguísticos. No caso de atos de fala, por exemplo, itens de busca que podem representar o agradecimento incluem dispositivos indicadores de força ilocutória (DIFIs) como (*muito*) *obrigado*(*a*), *eu agradeço*, *estou* (*muito*) *agradecido*(*a*) e *Deus lhe pague*. Por outro lado, itens de busca não linguísticos variam de informações extralinguísticas anotadas no corpus ao uso de códigos de transcrição (cf. VAUGHAN, 2008, para um estudo em que *laughs* e *laughter* são usados como itens de busca para o resgate de ocorrências de humor; CLANCY; MCCARTHY, 2015, para um estudo em que etiquetas de falantes, *e.g.* <$1>, são utilizadas como itens de busca para resgatar casos em que *if* e *when* são empregados como iniciadores de turno). Estudos que lançam mão deste modelo investigativo certamente fazem também o uso da estratégia de filtragem, uma vez que o relacionamento de uma forma linguística com uma função pragmática específica não é garantido.

Filtragem é o modelo de estratégia investigativa no qual o pesquisador necessita de uma leitura manual das ocorrências obtidas em uma busca para que, desta maneira, a separação entre ocorrências relevantes e não relevantes seja manualmente efetuada. Uma vez que uma primeira leitura superficial de exclusão é concluída, o pesquisador pode seguir para uma segunda leitura, mais detalhada e minuciosa, das ocorrências relevantes para a análise, determinação e descrição das funções de dada forma. A filtragem é especialmente útil para estudos que utilizam corpora não etiquetados, iniciando a análise a partir da forma. É também empregada em estudos que se baseiam em amostras, principalmente aqueles que empregam itens de busca não linguísticos. Contudo, estudos que não se baseiam em amostra, mas cujo foco está em determinada e específica forma linguística a fim de mapear suas funções pragmáticas, também se enquadram bem dentro deste modelo. Por exemplo, consideremos um estudo em que a análise foca nas funções que são performadas pelo marcador pragmático *tipo* na língua portuguesa oral do Brasil. A fim de analisar contextualmente o marcador pragmático e descrever as suas funções no discurso, o pesquisador precisa, primeiramente, 'separar o joio do trigo'; ou seja, eliminar ocorrências nas quais *tipo* não atua em uma função pragmática (*e.g. eu prefiro outro tipo de comida*).

Por fim, temos a **listagem baseada em frequência**. Neste modelo investigativo, que mais se assemelha a uma metodologia típica

de LC, o pesquisador inicia a pesquisa a partir de listas de frequência. Ou seja, listas de frequência, sejam elas de um corpus singular ou de comparações entre corpora, são analisadas com o objetivo de identificar formas salientes (palavras ou colocações) que possam indicar fenômenos pragmáticos. Neste sentido, a pesquisa e seus itens de busca são delimitados pelo corpus e seu contexto discursivo em particular e, de acordo com Ädel e Reppen (2008, p. 3), esta é uma estratégia efetiva para destacar padrões linguísticos de um banco de dados específico. O'Keeffe *et al*. (2020, p.53) afirmam que variações em frequência, distribuição e padrões linguísticos, são comumente justificadas através de conclusões pragmáticas em estudos com base em listas de frequência. De fato, itens linguísticos de natureza pragmática são omnipresentes no discurso oral, e mesmo abordagens guiadas por corpus (diferentes das baseadas em corpus) têm uma alta probabilidade de apresentar resultados que apontam para o sistema pragmático da linguagem em uso (cf. O'KEEFFE *et al*., 2007, capítulos 2 e 8).

Um exemplo de estudo guiado por corpus é o de Santos (2019). Neste estudo, o pesquisador usa uma amostra do *Spoken Corpus of Brazilian Portuguese and L2-English* (SCoPE²) (SANTOS, 2020) e compara uma lista das 15 palavras mais frequentes com uma mesma lista feita a partir do *Limerick Corpus of Irish English* (LCIE) (FARR *et al*., 2004), a fim de determinar diferenças e similaridades entre nativos do inglês irlandês e brasileiros universitários na Irlanda. Nesta etapa, *like* foi identificada como uma palavra merecedora de maior atenção, estando colocada na décima quarta posição da lista SCoPE², enquanto na terceira posição na lista LCIE. Uma segunda etapa de listagem incluiu uma lista de palavras-chave provenientes do SCoPE² quando este corpus foi contrastado e comparado ao LCIE como o corpus referência para o estudo. Neste momento, *like* ocupou a posição de terceiro lugar, confirmando sua saliência no SCoPE². Após uma filtragem manual de ocorrências pragmáticas e não pragmáticas, Santos (2019) pôde confirmar um total de 85% de ocorrências de *like* funcionando como um marcador pragmático. Em sua etapa qualitativa, o estudo demonstra que *like* apresenta as mesmas sete funções em ambos os corpora (mitigador, dispositivo de aproximação, exemplificador, marcador de hesitação, marcador de foco e dispositivo de discurso relatado), com diferenças concernentes apenas à quantidade de ocorrências em cada função, como também à posição sintática do marcador pragmático. Este estudo exemplifica e demonstra

a adequabilidade da estratégia de listagem baseada em frequência como um ponto de partida inicial para a investigação de formas linguísticas que se destacam em corpora de contextos linguísticos específicos devido às suas funções pragmáticas.

É inegável que a abordagem forma-função, com seus quatro modelos investigativos apresentados aqui, oferece ferramentas eficazes para o estudo de uma grande variedade de fenômenos pragmáticos baseados em corpora. Contudo, há também fenômenos que são mais efetivamente abordados dentro de um quadro metodológico reverso (função-forma), devido à dificuldade de se recuperar e identificar toda a abrangência de formas que os caracterizam. Esta abordagem reversa, embora mais complexa e justificavelmente mais trabalhosa, tem ganhado a atenção dos pesquisadores de PC nos últimos anos, e estes (especialmente os da escola europeia) têm se debruçado sobre o assunto a fim de desenvolverem um modelo metodológico inclusivo para a PC; ou seja, que inclua ambas as abordagens. Abaixo, apresentamos uma proposta de modelos investigativos que constituem a abordagem função-forma.

4.1.2 Abordagem função-forma

A abordagem função-forma espelha-se na metodologia tradicional pragmática na qual o ponto de partida acontece através de uma determinada função e investiga as formas que emergem a partir de contextos controlados por mecanismos de elicitação. Se, por um lado, a LC mostra-se uma perfeita aliada à Pragmática pela abordagem forma-função, como observamos na seção anterior, por outro, um dos desafios metodológicos da PC é identificar, em corpora de estudo, fenômenos pela abordagem função-forma que são variáveis e difíceis de ser identificados. Isso se dá porque as buscas pelas funções não são realizadas de forma automática e precisam ser bem delimitadas.

A metodologia pragmática partindo da função ainda é incipiente no campo da PC. Isso se justifica pela dificuldade em extrair as ocorrências de dada função, uma vez que para a maioria dos fenômenos pragmáticos, a correlação forma e função é inexistente (RÜHLEMANN; AIJMER, 2015). Ao mesmo tempo em que temos a técnica tradicional de elicitação de dados com controle do contexto situacional dos estudos pragmáticos a partir da função, acessamos uma menor quantidade de

dados. Em contrapartida, temos uma grande variedade de formas nos corpora disponíveis e pouca riqueza contextual, a menos que o corpus seja compilado pelo próprio pesquisador (O'KEEFFE, 2018).

É nesse desafio para equilibrar função, contexto, forma e dados linguísticos em grandes quantidades, que apresentamos e discutimos a seguir os modelos de estratégias investigativas, a partir de O'Keeffe *et al*. (2020), que compõem a abordagem função-forma dentro do quadro metodológico da PC. São eles: (a) *busca direta* (*one-to-one searching*), (b) *amostragem, busca e filtragem* (*sampling, searching and sifting*), (c) *sementes* (*seeds*), e (d) *mapeamento pragmático indireto* (originalmente nomeado pelos autores como *solutions for larger corpora*, mas aqui apresentado com a terminologia adaptada).

A **busca direta** apresenta maiores dificuldades na abordagem função forma**.** Tomando como base os atos de fala em um corpus, McAllister (2015, p. 29) adverte-nos que neste tipo de busca o pesquisador limita-se às formas linguísticas que considera atuar de forma pragmática, não sendo suficiente para destacar as formas de tal fenômeno em sua totalidade, o que descarta, principalmente, as formas não presumidas como tendo valor pragmático. Desta forma, a recuperação de todas as ocorrências relevantes da busca direta nessa perspectiva, que parte da função, só é possível por meio de corpora pragmaticamente anotados, considerados os 'cálices sagrados' para a área da PC (O'KEEFFE *et al*., 2020).

A construção de corpora orais não é tarefa fácil, e isso se reflete no grande número de corpora escritos que, presumivelmente, ainda dominam a área da LC. Isso sucede porque a compilação envolve uma série de procedimentos, como o desenho do corpus, a disponibilidade dos informantes, o processo de gravação, a elaboração dos critérios de transcrição, o processo transcritório em si, entre outros. Nesse sentido, o número de corpora orais anotados pragmaticamente é ainda menor, mas se torna necessário dependendo do tópico a ser investigado. Apesar de a anotação de corpora ser um processo manual maçante, esta realidade está gradativamente mudando nos últimos anos com o advento de novas ferramentas e sistemas de anotação que, posteriormente, nos permitem a extração automática de um determinado fenômeno pragmático, como o *Dialogue Annotation and Research Tool* (DART) (WEISSER, 2015). Segundo Weisser (2019, p. 131), a ferramenta possibilita pesquisas com grandes corpora não apenas para investigações de aspectos sintáticos,

mas também pragmáticos, como DIFIs, atos de fala, entre outras características da interação.

No Brasil, destacamos o C-ORAL-BRASIL (RASO; MELLO, 2012), corpus de fala espontânea informal do português brasileiro que surgiu a partir do C-ORAL-ROM (CRESTI; MONEGLIA, 2005), por sua vez composto por línguas românicas europeias (espanhol, francês, italiano e português europeu). A estrutura do C-ORAL-BRASIL permite o estudo da fala sob uma perspectiva pragmática devido ao seu design e ampla variedade situacional (RASO, 2012). Para isso, o corpus é segmentado em enunciados (atos de fala) e em unidades tonais, respectivamente com barra dupla (//) e barra simples (/), e símbolos para indicação de enunciados interrompidos (+) e retratações ([/n]). Ademais, as transcrições e o material acústico são alinhados. Todo o processo de segmentação do C-ORAL-BRASIL foi validado para a redução de desacordos entre os membros do projeto. Como exposto, devido à anotação de um corpus grande ser desafiadora, uma saída para o estudo de caráter pragmático que os pesquisadores têm encontrado é a anotação de corpora menores que possibilitam uma anotação mais precisa. Ainda assim, como aponta O'Keeffe (2018), o trabalho requer critérios de validação.

Exemplos de corpora internacionais anotados incluem o subcorpus do *Michigan Corpus of Academic Spoken English* (MICASE), manualmente anotado para atos de fala (MAYNARD; LEICHER, 2007) e o *Systems of Pragmatic annotation in ICE-Ireland* (SPICE-Ireland) Corpus, um projeto desenvolvido a partir do subcorpus do *International Corpus of English-Ireland* (ICE-Ireland) (KALLEN; KIRK, 2012), e são considerados avanços para a área. O SPICE-Ireland é anotado pragmática e prosodicamente e possui aproximadamente 600,000 palavras. Dessa forma, por exemplo, o status do ato de fala de cada enunciado foi identificado com etiquetas, tendo Searle (1976) como base: representativas <rep> ... </rep>; diretivas <dir> ... </dir>; comissivas <com> ... </com>; expressivas <exp> ... </exp> e declarativas <decl> ... </decl>. Além disso, outras marcações foram usadas, como (*) para marcadores discursivos, (+) para indicar discursos relatados, (%) em final de unidade entonacionais, entre outras combinações. Portanto, num modelo de busca direta, dentro da abordagem função-forma, o uso de um corpus pragmaticamente anotado como o SPICE-Ireland possibilita a extração total de formas que representam um determinado fenômeno pragmático. Ou seja, um pesquisador interessado em atos da fala com

foco, por exemplo, em expressivas do inglês irlandês, tem a possibilidade de estudar todas as formas desta função naquela variedade linguística com o auxílio das etiquetas que foram previamente anotadas no corpus. Em outras palavras, todas as formas de uma determinada função pragmática são recuperadas através do uso de itens de busca com base em etiquetas e códigos anotados.

Corpora pequenos, apesar de não serem ideias para certos estudos como os lexicais e fraseológicos, podem ser adequados a outros (KOESTER, 2010). Nesse sentido, especialmente a partir de perguntas de pesquisa e contextos bem delimitados, conseguimos anotar e extrair fenômenos pragmáticos específicos para análise. Já considerando a abordagem função-forma para investigar corpora maiores, precisaremos combinar modelos investigativos a fim de delimitar os dados.

A **amostragem**, **busca e filtragem** é a combinação de modelos investigativos, aplicada a corpora maiores, que nos possibilita uma análise pragmática partindo da função. A transformação de um corpus grande em uma amostra menor é ideal para a PC. Com o objetivo de tornar determinado corpus mais facilmente manejável e a investigação completa de determinada função pragmática, este modelo inicia-se através da escolha de uma amostragem aleatória de determinado corpus. A partir desta amostra, estabelecemos os parâmetros para identificar os itens e conduzimos uma filtragem manual, ao mesmo tempo em que conduzimos o processo de anotação de todas as ocorrências relevantes nos dados. Dessa forma, podemos, através da anotação por uma leitura qualitativa, categorizar todos os fenômenos restantes. Um dos principais benefícios deste modelo investigativo, é que, se o pesquisador ainda tiver acesso aos metadados, este terá em mãos um número maior de ocorrências significativas do que um estudo tradicional pragmático possibilitaria, além das informações contextuais detalhadas das situações.

Por sua vez, a busca a partir de **sementes**, isto é, resultados de pesquisas anteriores, é outro método que pode ser aplicado a corpora maiores. Os resultados provenientes de pesquisas pragmáticas, através dos variados instrumentos de coleta ao longo dos anos, constituem um rico banco de dados neste modelo de investigação. Dessa forma, o pesquisador não desconsidera o que foi produzido e tem as sementes como ponto de partida para análise das funções pragmáticas em corpora. Considerando as pesquisas pragmáticas com TCDs, por exemplo, podemos considerar resultados de estudos como os de Beebe *et al*. (1990), que investigaram

a transferência pragmática em recusas de inglês como segunda língua em um grupo de japoneses nativos de japonês, japoneses falantes de inglês como segunda língua e americanos nativos de inglês. Os atos de fala de recusa foram analisados em pedidos, convites, ofertas e sugestões, considerando ainda variáveis sociais. Segundo os autores (p. 57), se um participante, na recusa de um convite para um jantar realizado por amigos, proferir um ato de recusa como '*I'm sorry, I have theater tickets that night. Maybe I could come by later for a drink*', teremos uma fórmula como [*expression of regret*] [*excuse*] [*offer of alternative*]. Assim, neste modelo de análise da função-forma, podemos considerar formas específicas e fórmulas oferecidas pelo estudo de Beebe *et al.* (1990) como sementes para a busca em outros corpora, considerando ainda contextos situacionais similares, obtendo um conjunto de dados comparáveis. Tem-se, nesse caso, modelos de formas existentes para extrair fenômenos específicos em corpora e gerar um subcorpus comparável pelas mesmas condições (O'KEEFFE, 2018).

4.1.2.1 Mapeamento pragmático indireto na função-forma

Como vimos até aqui, as formas de refinamento na abordagem função-forma exigem que o pesquisador busque mecanismos diferentes para extrair dados relevantes. Outros métodos que podem ser usados como forma de investigação, como apontado por O'Keeffe *et al*. (2020), consistem no uso de (a) *DIFIs*, (b) *busca por léxico ou características gramaticais associadas a um ato de fala*, e (c) *busca por expressões metacomunicativas*. Destacamos que estes métodos são apresentados pelos autores como constituintes de um conjunto de ações a serem aplicadas como soluções metodológicas para o uso de corpora de maior escala. Contudo, consideramos tais modelos investigativos como parte de um mapeamento pragmático indireto que consiste em mecanismos de buscas mediadas por itens que objetivam abranger determinadas funções pragmáticas ao máximo. Como apresentado anteriormente, soluções metodológicas para grandes corpora que não são pragmaticamente anotados relacionam-se, frequentemente, com estratégias como amostragem e filtragem, que, por sua vez, são invariavelmente aplicadas aos modelos apresentados dentro do escopo de mapeamento pragmático indireto, assim justificando uma adaptação de terminologia para o último modelo que constitui a abordagem função-forma.

Nesse sentido, a partir do refinamento de dados por meio do uso dos DIFIs, considerando o ato de fala expressivo na língua inglesa, tomamos como exemplo o ato de desculpar-se e sua realização pelo DIFI *sorry*. Dessa forma, teremos várias ocorrências para esta busca, mas observamos que o uso de DIFIs na PC está relacionado à precisão e busca limitados. Isso se dá porque quando buscamos por *sorry*, deparamo-nos com ocorrências não relacionadas ao ato de desculpar-se, com o item de busca aparecendo, por exemplo, como atributivo (JUCKER, 2013). Da mesma forma, a busca é limitada, pois nem todos os pedidos de desculpas incluem *sorry,* e tal DIFI pode não ser representativo deste ato. Assim, para um mapeamento mais abrangente, precisaremos incluir outros DIFIs como *my mistake*, *pardon* ou *excuse me*. No entanto, o processo de filtragem neste processo é indispensável.

Como extensão da busca por DIFIs, consideramos o uso do léxico ou expressões gramaticais associadas a determinados atos de fala (JUCKER, 2013). Em um estudo sobre elogios através de diferentes culturas, Wolfson (1981, p. 122) mostra que estes são altamente convencionalizados no inglês americano em estruturas sintáticas como *NP [is/looks] (really) ADJ* ou *I (really) [like/love] NP*. Destacamos aqui a importância de termos um corpus etiquetado morfossintaticamente que ajudará nesse tipo de busca. A não extração de todas as formas, principalmente as mais incomuns, é uma limitação, além de não ser uma metodologia direta, já que precisamos transformar expressões em fórmulas específicas (JUCKER, 2013).

Por fim, expressões metacomunicativas são 'palavras e frases que podem ser usadas para falar sobre aspectos da comunicação, no sentido de que nomeiam um determinado ato de fala, como elogio, saudação, insulto ou agradecimento, ou um tipo específico de comportamento, como a polidez e a impolidez[12] (JUCKER *et al.*, 2012). Portanto, podemos examinar como as pessoas concebem e avaliam o uso da linguagem (HAUGH, 2018). No entanto, a análise da metalinguagem ou expressões metacomunicativas necessita do suporte de megacorpora para a obtenção de ocorrências significativas, e, ainda que o acesso a evidências de

[12] Nossa tradução para: "[metacommunicative expressions are] words and phrases that can be used to talk about aspects of communication, in the sense that they name a particular speech act, such as compliment, greet, insult or thank, or a particular type of behaviour, such as polite or impolite"

determinados elementos a partir de como as pessoas os avaliam seja indireto, temos em mãos dados importantes de natureza etnográfica (JUCKER, 2013). Nesse sentido, o estudo de Culpeper (2009) demonstra isso ao investigar a metalinguagem da impolidez por acadêmicos e não acadêmicos no *Oxford English Corpus* com aproximadamente 2 bilhões de palavras. O autor teve como ponto de partida uma lista de rótulos usados na literatura sobre polidez e impolidez, como *impolite(ness)*, *rude(ness)*, *aggravation*, *aggravated/aggravating, language/facework* (*aggravated impoliteness*), *aggressive facework*, *face attack* e *verbal aggression*. Culpeper revela que *rude* é frequente na linguagem de não acadêmicos e, na maioria das vezes, ocorre em contextos públicos, além de que entre os sujeitos tidos como *rude*, estão porteiros, garçons/garçonetes, equipe. Curiosamente, as pessoas também fazem menção à 'Yorker' e 'French', reforçando estereótipos de lugares.

Discorremos, nas duas últimas seções, os oito modelos investigativos de análise que integram as abordagens forma-função e função-forma, essenciais para as investigações de fenômenos pragmáticos em corpora. Na próxima seção do artigo, escolhemos um modelo investigativo, a **filtragem**, pela abordagem forma-função, para ilustrar em um estudo de caso como a análise do marcador pragmático *kind of* pode ser conduzida.

5 Estudo de caso: *kind of* no inglês de brasileiros universitários

Segundo Jucker *et al.* (2018, p. ix), PC é uma das três abordagens metodológicas empíricas para a análise de fenômenos pragmáticos (sendo as outras duas a pragmática experimental e a pragmática observacional). Ademais, dentre os fenômenos pragmáticos, os marcadores pragmáticos (MPs) são considerados como umas das áreas-chave da pesquisa pragmática de corpus (CLANCY; O'KEEFFE, 2015). MPs são uma extensa e eclética classe de itens linguísticos que se manifestam em formas de palavras (*e.g. like* no inglês, e *tipo* no português) ou fórmulas linguísticas padronizadas (*e.g. you know* e *you know what I mean* no inglês, e *olha só* e *e por aí vai* no português), muitas vezes também manifestando-se em formas que possam não ser consideradas como palavras (*e.g.* interjeições: *ah*, *oh*; tokens responsivos: *mmhm*/*hum hum*). Estes itens linguísticos marcam as atitudes e posicionamentos dos falantes, e atuam de maneira interpessoal no ato da conversação, além

de funcionarem como um auxílio na organização estrutural do discurso (cf. BRINTON, 1996; CARTER; McCARTHY, 2006).

Assim, em consideração à natureza pervasiva dos MPs na linguagem, como também atentos à sua importância pragmática no processo de interpretação de sentidos na comunicação, apresentamos agora um estudo de caso comparativo, no qual analisamos o MP *kind of* no inglês falado por brasileiros universitários no Brasil e na Irlanda. O estudo de caso apresentado tem objetivo duplo: (a) exemplificar o modelo da PC em um estudo que considera um fenômeno pragmático com base em corpora, e (b) contribuir com o campo de estudos que visa a melhor compreensão do uso de MPs em língua adicional/estrangeira (L2). Apesar de a produção acadêmica com foco em MPs na língua materna (L1) não ser escassa, muito ainda precisa ser estudado e compreendido com respeito ao uso de MPs na L2. Em sua maioria, os estudos de MPs na L2 tendem a comparar a produção linguística do não-nativo com aquela do nativo. Embora tal comparação seja positiva, por destacar características particulares da L2, alertamos para o fato de que tal comparação não deve atrelar-se à perspectiva de linguagem deficiente com a qual muitos destes estudos são relacionados quando descrevem a linguagem do não-nativo que não atinge uma norma nativa (cf. PRODROMOU, 2008). Neste sentido, o presente estudo não só aborda L2 de uma perspectiva de competência, como também compara L2 com L2, a fim de investigar possíveis diferenças no uso do MP *kind of* entre dois grupos distintos de brasileiros universitários falantes do inglês.

O sentido nuclear de *kind of/sort of* é o de aproximação e imprecisão. Ao fazerem uso deste MP, os falantes indicam aos seus interlocutores que o material linguístico que o sucede deve ser interpretado de maneira vaga e imprecisa. Ou seja, *kind of* (como também sua versão mais usada no inglês britânico, *sort of*) sinaliza uma informação de imprecisão, e convida o interlocutor a aproximar e ajustar o material linguístico para que a interpretação intencionada seja feita. Estudos que investigaram *kind of*/*sort of* em L1 destacam que estes MPs operam nos três principais domínios funcionais da classe: atitudinal, interpessoal e textual (cf. AIJMER, 2002; KIRK, 2015). No domínio atitudinal (ou de posicionamento), o falante marca que há uma discrepância entre o que este tem em mente e o material linguístico falado. Neste domínio, *kind of/sort of* apresenta um status especial quando comparado com a maioria dos MPs, uma vez que sua presença afeta diretamente o sentido do conteúdo proposicional (AIJMER, 2002). No domínio interpessoal

(ou interativo), *kind of/sort of* sinaliza que o material linguístico marcado deve ser interpretado de maneira não assertiva e final. Em outras palavras, neste domínio, o falante leva em consideração a maneira que sua escolha de palavras será interpretada pelo seu interlocutor e, portanto, *kind of/ sort of* é usado para mitigar a força ilocutória do material linguístico que o sucede. Finalmente, no domínio textual, o MP é utilizado para sinalizar uma necessidade de ajustamento por parte do próprio falante com relação à estrutura do seu discurso. Neste domínio, *kind of/sort of* auxilia o falante com a reestruturação semântica e sintática do conteúdo comunicado, indicando também, a nível interpessoal, que o falante deseja manter o turno e continuar seu discurso.

No que diz respeito aos estudos investigativos do MP *kind of/sort of* em L2, destacamos três que lançaram mão do *Louvain International Database of Spoken English Interlanguage* (LINDSEI) a fim de apontar diferenças no uso deste MP entre nativos e não-nativos. O primeiro, Gilquin (2008), investiga marcadores de hesitação no discurso de aprendizes em nível avançado de inglês com língua materna francesa. Para tanto, a autora compara o LINDSEI-FR com o *Louvain Corpus of Native English Conversation* (LOCNEC). Dentre os marcadores de hesitação analisados, a autora inclui uma subcategoria de MPs, da qual *kind of/sort of* é pertencente. Os resultados de sua pesquisa demonstram que este coorte de aprendizes faz grande uso de pausas, sejam elas preenchidas ou silenciosas, mas não conseguem explorar a variedade de MPs que desempenham a mesma função. A autora também sugere, com base em seus resultados, que essa má representação de MPs pode ser devido a uma superdependência do MP *well*, que apresentou grande frequência no LINDSEI-FR. Com referência ao MP *kind of/sort of*, a autora destaca que *sort of* é usado, funcionalmente, de maneira diferente quando comparado com nativos, uma vez que estes aprendizes fazem um maior uso da sua função textual devido a limitações de vocabulário. Apesar de a maioria dos exemplos apresentados pela autora serem discutivelmente casos de hesitação, no sentido de reformulação e ganho de tempo para recuperação cognitiva de vocabulário, o uso de *kind of/ sort of* nesta função por não-nativos é previsto. Tal função já foi evidenciada no discurso nativo do inglês irlandês (KIRK, 2015), e do inglês britânico (AIJMER, 2002), sendo uma estratégia conversacional crucial. Sendo assim, é de se esperar que não-nativos, especialmente aprendizes, façam uso da função textual do MP *kind of/sort of* em sua busca por formulação linguística adequada em uma L2.

Buysse (2010), por sua vez, analisa o inglês falado por aprendizes avançados e nativos da língua holandesa. O autor investiga o uso de uma seleção de alguns dos MPs mais comuns da língua inglesa: *so*, *well*, *you know*, *like*, *kind of*/*sort of*, e *I mean*. Este é um estudo quantitativo e apresenta resultados de frequências de ocorrências de cada um dos MPs investigados de maneira comparativa entre os corpora LINDSEI-DU e LOCNEC. Os resultados apresentados mostram que os aprendizes do LINDSEI-DU raramente fazem uso de MPs com funções interpessoais (*you know*, *like*, *kind of*/*sort of*, *I mean*), enquanto aqueles com funções textuais são usados em demasia. Apesar da importância funcional dos MPs de funções interpessoais, o autor argumenta que um motivo para sua baixa frequência no LINDSEI-DU, quando comparado ao LOCNEC, é o fato de tais MPs serem relacionados à linguagem informal e, muitas vezes, serem estigmatizados.

Finalmente, Miranda (2020) investiga o uso de marcadores de linguagem vaga por brasileiros aprendizes de inglês em nível avançado e americanos nativos da língua inglesa. Comparando os corpora LINDSEI-BR e SBCSAE (*Santa Barbara Corpus of Spoken American English*), o autor faz uma análise quantitativa de vários marcadores de linguagem vaga, destacando o MP *kind of/sort of* por ser o mais frequente em ambos os corpora. Contudo, e interessantemente, apesar de os aprendizes brasileiros não fazerem uso da forma *sort of*, enquanto os falantes americanos o fazem com ambas as formas, LINDSEI-BR apresenta uma frequência maior do que no SBCSAE quando somados *kind of* e *sort of* juntos. Na etapa qualitativa de seu estudo, Miranda (2020) identifica as funções pragmáticas exercidas pelo MP *kind of* em ambos os corpora, e nota que a função mais usada pelos aprendizes brasileiros é atitudinal, marcando inexatidão do material linguístico que o sucede, seguida da função interpessoal de mitigação. No SBCSAE, contudo, a mitigação marcada pelo *kind of* é mais frequente que no LINDSEI-BR, e casos de *kind of* marcando vocabulário técnico ou complexo, como também linguagem vulgar, são encontrados apenas no SBCSAE.

O que estes estudos sobre o uso de MPs na linguagem do aprendiz revelam é que, apesar de algumas limitações, MPs são também presentes no discurso oral do aprendiz e têm um papel importante na L2. Por equiparem falantes de L2 em um nível interpessoal quando usados com sucesso, e por limitarem a contribuição destes falantes em conversação quando ausentes, MPs são indispensáveis para uma comunicação bem

sucedida. Considerando este panorama apresentado, o presente estudo contribui para o corpo de estudos em MPs na L2 ao analisar *kind of* no discurso oral de brasileiros falantes de inglês em dois contextos diversos.

5.1 Corpora e metodologia

Os dados sob investigação no presente estudo são provenientes de dois corpora: o SCoPE² e o *Brazilian Spoken English Learner Corpus* (BraSEL), e apenas um subcorpus de cada é utilizado. O SCoPE² e o BraSEL foram desenvolvidos para projetos de doutorado com o objetivo de analisar a pragmática do inglês falado por brasileiros, porém em contextos linguísticos distintos. O primeiro é um corpus bilíngue composto pelo inglês falado por universitários na Irlanda e pelo português nativo destes mesmos participantes (cf. SANTOS, 2020 para uma descrição detalhada do processo de desenho, compilação e transcrição do corpus). O material linguístico do SCoPE² representa a linguagem informal em uso real, e foi coletado através de encontros informais de bate-papo entre o pesquisador e os participantes. É importante ressaltar que este não é um corpus de aprendizes, mas sim de inglês como língua adicional, uma vez que todos os participantes terminaram os seus cursos de língua inglesa, obtiveram com sucesso um certificado internacional de proficiência em nível avançado ou proficiente, e se comunicam eficientemente dentro de um ambiente internacional com ambos nativos e não-nativos da língua inglesa a nível pessoal, profissional e cultural. Nesse sentido, os participantes do SCoPE² são considerados *successful users of English* (SUEs; PRODROMOU, 2008).

Em contrapartida, o BraSEL é um corpus que está em fase de compilação, e é composto por material linguístico coletado de brasileiros universitários no Brasil que, seguindo a tabela de níveis de proficiência do Quadro Europeu Comum de Referência para as Línguas (CEFR, no acrônimo em inglês), classificam-se entre os níveis A1-C2. Este é um corpus de aprendizes de inglês como língua estrangeira e os dados são coletados através de um encontro entre um entrevistador e o participante para uma conversa informal que envolve três atividades comunicativas distintas. Na primeira parte, os aprendizes têm a oportunidade de brevemente apresentarem-se, enquanto a segunda parte decorre através de uma conversa informal sobre a vida em geral dos aprendizes (*e.g.* interesses e hobbies). Por fim, a terceira e última etapa é baseada em uma série de imagens provocantes, das quais o aprendiz deve selecionar uma e

descrevê-la, também respondendo perguntas sobre as imagens. Tabela 1, abaixo, resume as características de cada subcorpus utilizado neste estudo.

TABELA 1 – Características dos subcorpora provenientes do SCoPE² e do BraSEL Corpus

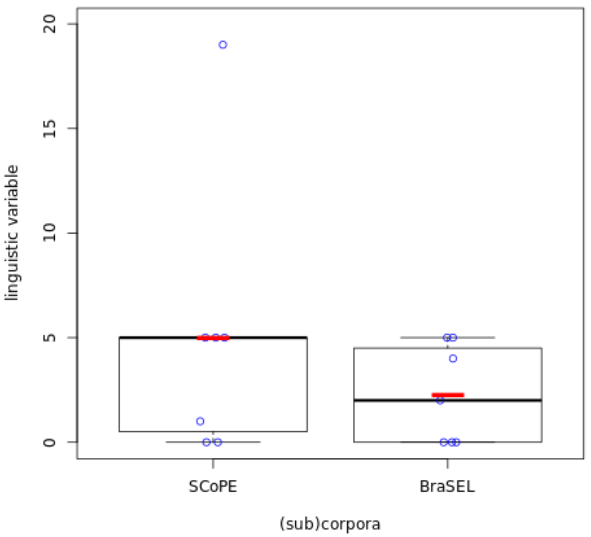
	SCoPE²	BraSEL
Número de participantes	7	7
Gênero	5 mulheres, 2 homens	5 mulheres, 2 homens
Nacionalidade	Brasileiros	Brasileiros
Contexto de L2	Universitários na Irlanda	Universitários no Brasil
Nível de proficiência	C1-C2 CEFR; SUEs (Prodromou, 2008)	B2 CEFR
Tipo de dados	Gravações de áudio de bate-papos informais	Gravações de áudio de tarefas comunicativas informais
Tipo de interação	Um a um (pesquisador e participantes)	Um a um (entrevistador e participantes)
Número de tokens	43,596	20,890

A metodologia de análise aplicada no presente estudo refere-se a um dos modelos de estratégias investigativas da abordagem de PC forma-função: a filtragem (cf. Seção 4.1.1). O método de análise também se beneficia do modelo de leitura integrada proposto por Rühlemann e Aijmer (2015) e descrito na Seção 4. Assim sendo, primeiramente o número total de ocorrências do MP *kind of* em cada subcorpus foi contabilizado e normalizado a fim de comparação quantitativa. Apenas as ocorrências referentes aos participantes foram consideradas, uma vez que as participações do pesquisador e do entrevistador foram excluídas. As ocorrências foram, então, manualmente verificadas e filtradas para que casos não pragmáticos de *kind of* fossem também excluídos. Na etapa qualitativa, cotexto e contexto foram analisados a fim de informar e auxiliar na determinação de funções pragmáticas desempenhadas pelo MP no discurso oral de L2 dos brasileiros universitários participantes de ambos os corpora. Finalmente, cada caso pragmático de *kind of* foi alocado em um dos três domínios funcionais do MP (atitudinal, interpessoal e textual) a motivo de comparação de preferência funcional demonstrada por cada grupo de brasileiros ao fazerem uso desta forma linguística.

5.2 Resultados

Apesar de *kind of* e *sort of* performarem as mesmas funções pragmáticas como MPs, apenas o primeiro é utilizado pelos brasileiros nas amostras analisadas. Considerando que *kind of* é mais comumente relacionado com o inglês americano, e *sort of* com o inglês britânico, a ausência do segundo pode ser justificada pelo fato de brasileiros terem mais contato com o inglês americano, seja através de instrução formal em sistemas de ensino de língua inglesa, ou através de mídias de entretenimento. Quando o número de ocorrências é quantitativamente comparado entre SCoPE² e BraSEL, *kind of* ocorre 16 vezes no BraSEL (frequência normalizada de 11.6 por cada 10,000 palavras) e 35 vezes no SCoPE² (13 por cada 10,000 palavras). Contudo, como apresentado na Figura 2 abaixo, a dispersão do MP em ambos os subcorpora não é nivelada. No SCoPE², dois participantes não usam o MP de maneira alguma, enquanto um participante sozinho faz uso do MP 19 vezes (cada participante é identificado por um círculo na figura, ou *boxplot*, abaixo). No BraSEL, por sua vez, três participantes não fazem uso do MP, enquanto outros dois fazem uso do MP 5 vezes, e outro o utiliza 4 vezes, não apresentando nenhum caso fora da curva (ou *outlier*; BREZINA, 2018).

FIGURA 2 – Dispersão de *kind of* nos subsorpora SCoPE² e BraSEL



Fonte: Boxplot gerado pela plataforma Lancaster Stats Tools online (BREZINA, 2018)

Estes resultados demonstram que, embora a frequência do uso de *kind of* entre os dois subcorpora analisados seja similar em sua superfície, uma investigação mais aprofundada de dispersão do item linguístico em ambos os subcorpora revela que nem todos os participantes fazem uso deste MP, e que um *outlier* em particular no SCoPE² apresenta uma possível maior dependência de *kind of* em seu discurso oral de inglês L2. Na próxima subseção, apresentamos resultados qualitativos a fim de descrever e exemplificar as funções performadas por *kind of* nos subcorpora analisados, bem como sugerir uma possível explicação para o caso fora da curva apresentado no SCoPE².

5.2.1 Resultados qualitativos

Como observado acima, o MP *kind of* atua nos três domínios funcionais da classe: o atitudinal (de posicionamento), o interpessoal (interativo) e o textual (organização do discurso). Contudo, *kind of* foi também identificado como atuando pragmaticamente em fórmulas padronizadas que constituem MPs de linguagem vaga (*e.g*. marcadores de categorias vagas como *and this kind of things*, *or this kind of stuff*, etc.), e também em casos onde o falante usa *kind of* coocorrendo com *things* para marcar vagueza no discurso. Estes dois casos, que não se classificam como MPs *kind of* independentes, são aqui classificados como *Outros*. Exemplos de todas as funções citadas são apresentados a seguir. Contudo, Tabelas 2 e 3, abaixo, primeiramente expõem a distribuição funcional de *kind of* no SCoPE² e no BraSEL, respectivamente.

TABELA 2 – Distribuição funcional de *kind of* no SCoPE²

	Atitudinal	Interpessoal	Textual	Outros	TOTAL
SUE_1	1	2	1	1	5
SUE_2	0	0	0	0	**0**
SUE_3	8	0	1	10	**19**
SUE_4	1	0	0	0	**1**
SUE_5	5	0	0	0	**5**
SUE_6	0	0	0	0	**0**
SUE_7	2	3	0	0	**5**
TOTAL	17	5	2	11	35

TABELA 3 – Distribuição funcional de *kind of* no BraSEL

	Atitudinal	Interpessoal	Textual	Outros	TOTAL
Aprendiz_1	1	0	1	0	**2**
Aprendiz_2	0	0	0	0	0
Aprendiz_3	0	0	0	0	**0**
Aprendiz_4	1	3	1	0	**5**
Aprendiz_5	2	3	0	0	**5**
Aprendiz_6	4	0	0	0	**4**
Aprendiz_7	0	0	0	0	**0**
TOTAL	8	6	2	0	16

Como visto nas Tabelas 2 e 3 acima, todas as funções pragmáticas de *kind of* são encontradas em ambos os subcorpora, apesar de não utilizadas na mesma frequência por todos os brasileiros participantes ou, às vezes, sequer usadas por alguns, com exceção apenas de *kind of* marcando linguagem vaga na estrutura interna da oração ou constituindo MPs de categorias vagas no BraSEL.

No domínio funcional atitudinal, o falante marca o material linguístico que sucede *kind of* como impreciso e inexato, e sinaliza que tal material deve ser ajustado de maneira a aproximar-se o máximo possível do sentido intencionado. Desta forma, o MP pode ser usado para (a) ajustar uma diferença entre um pensamento e uma representação linguística; (b) marcar palavras técnicas, jargões, uso metafórico da linguagem e coloquialismos do falante; (c) como um aproximador de numerais; e (d) para compensar a ausência de vocabulário no repertório linguístico de um falante (AIJMER, 2002). Nos subcorpora analisados, a função mais usada do *kind of* atitudinal é (a), seguida por (d); a função (b) não está presente nos subcorpora, enquanto a função (c) é usada apenas uma vez. É importante ressaltar que ambas as funções (a) e (d) caminham lado a lado e, muitas vezes, há uma linha tênue entre *kind of* marcando a necessidade de ajuste entre um pensamento e sua representação linguística e o mesmo marcando explicitamente a ausência de vocabulário no repertório linguístico de um falante, especialmente em contextos de L2. Considere os exemplos abaixo extraídos do SCoPE² e do BraSEL, que representam as funções (a) e (d). Note que, como destacado nas Tabelas 2 e 3 acima, participantes SUE fazem parte do SCoPE², enquanto participantes aprendizes fazem parte do BraSEL:

(1) SUE_3: *I went to Whistler. It´s* ***kind of*** *Campos do Jordão in Brazil.*

(2) Aprendiz_6: *Actually, I have started one book at vacation <nv> laugh </nv> but when when, em, our classes returned I* ***kind of*** *abandoned it.*

(3) SUE_3: *... students who live like Sligo <$E> pause </$E> they receive two amounts <$E> pause </$E> one for food <$E> pause </$E> three hundred+*

Pesquisador: *Mmhm.*

SUE_3: *+in a card that is* ***kind of*** *uhm vale-alimentação+*

Pesquisador: *Okay.*

SUE_3: *+and another one is the money to use with things, four hundred.*

(4) Aprendiz_5: *... I thought it was* ***kind of*** *a a lash, I think that's the name.*

Entrevistador: *A leash.*

Aprendiz_5: *A leash yes now yes uh-huh.*

Exemplos (1) e (2), acima, são casos de marcação de necessidade de ajuste lexical pelo MP *kind of*. No trecho (1), SUE_3 faz uma referência a uma cidade brasileira conhecida por ambos os participantes da conversa, e convida o seu interlocutor a inferir características reconhecidas de Campos do Jordão para a conceptualização mental da cidade de Whistler, no Canadá. Em outras palavras, *kind of* não só marca Campos do Jordão como uma representação linguística aproximada do que SUE_3 tem em mente, como também ajuda na interpretação de sentido da informação oferecida. Em (2), similarmente, Aprendiz_6 sinaliza que o verbo *abandon* pode não ser exatamente a melhor opção para o sentido intencionado, convidando, assim, o interlocutor a participar de um processo de inferência para a interpretação do sentido proposto.

Em contrapartida, exemplos (3) e (4) demonstram *kind of* marcando uma carência de vocabulário. Em (3), SUE_3 mais uma vez marca a troca de idiomas com *kind of*, mas, desta vez, claramente demonstrando que o termo em inglês para *vale-alimentação* não faz parte de seu repertório linguístico. Similarmente, Aprendiz_5 demonstra

dificuldade com a palavra *leash* e marca a necessidade de aproximação interpretativa por parte de seu interlocutor com o MP *kind of*.

No que diz respeito ao domínio funcional interpessoal, *kind of* também é utilizado para sinalizar uma discrepância entre um pensamento e uma representação linguística. Contudo, neste domínio funcional, o MP atua como um mitigador, suavizando a força ilocutória do material linguístico que o sucede. Na maioria dos casos analisados neste estudo, *kind of* interpessoal é usado para minimizar o impacto da escolha de palavras que são usadas, tendo em consideração a maneira como o interlocutor pode interpretá-las. Exemplo (5), abaixo, demonstra tal função interpessoal de *kind of*, onde Aprendiz_4 fala sobre a forma com a qual moradores de BH encaram um determinado bairro como um lugar perigoso. Aprendiz_4 usa a expressão *common sense*, mas a suaviza por não saber se a opinião é compartilhada pelo entrevistador.

(5) Aprendiz_4: *... how people talk about wh= what what stuff happens in BH. Everybody just mention oh the north area it's dangerous, don't don't go there. It's* ***kind of*** *common sense.*

Por sua vez, no domínio funcional textual, *kind of* auxilia o falante com a construção do discurso, no sentido de autocorreção por parte do falante que sinaliza que há uma discrepância entre a estrutura ou vocábulo usado e o que este tem em mente. Nesta função, *kind of* também é interativo, uma vez que sinaliza que o falante deseja manter o turno e completar sua fala. Apenas quatro ocorrências desta função foram encontradas nos subcorpora analisados, e o trecho (6), abaixo, exemplifica esta função:

(6) Aprendiz_4: *Because it, like, when you live in BH you, the society* ***kinda*** *mm not the society but how people talk about wh= what what stuff happens in BH.*

No trecho (6), que precede o trecho (5), Aprendiz_4 reestrutura a sua fala, e o faz com o auxílio de diversos marcadores de hesitação e autocorreção, como *like*, *mm*, e palavras cortadas e repetidas (*wh*= e *what*), incluindo a versão mais informal de *kind of*, *kinda*.

Finalmente, onze casos de *kind of* categorizados como *Outros* foram identificados, todos no SCoPE², e dos quais dez são utilizados por

um SUE apenas. Trechos (7) e (8) exemplificam duas subcategorias que compõem *Outros*:

(7) SUE_3: *... actually, not now but before during the the year <$E> pause </$E> uhm every Tuesday a woman came to see if everything was clean* ***this kind of things****.*

(8) SUE_3: *And because we know we have* ***this kind of things*** *in Brazil and I travel a lot with my parents.*

No exemplo (7), *kind of* é componente de uma fórmula fixa que compõe uma subclasse de MPs, denominada marcadores de categorias vagas. A fórmula ocorre sempre como uma etiqueta de acréscimo ao final de uma contribuição linguística e marca aproximação e necessidade de ajuste por parte do interlocutor de uma categoria inteira. Em (7), SUE_3 fala sobre ter uma funcionária que, semanalmente, vai à hospedagem estudantil para fazer a limpeza. Desta maneira, *this kind of things* é retroativo e convida o interlocutor a inferir os tipos de coisas que se classificariam em uma categoria vaga de ações que envolvem limpeza. Em outras palavras, SUE_3 não precisa listar uma extensiva lista de atividades, uma vez que tal lista pode ser inferida pelo interlocutor através de uma interpretação de *this kind of things* no contexto em que este é usado. Diferentemente, exemplo (8) demonstra a mesma forma linguística, mas usada de maneira distinta. Em (8), *this kind of things* não é um acréscimo extra, mas parte interna da oração. Juntos, *kind of* e *things* atuam para implicar um sentido de linguagem vaga, e convidam o interlocutor a interpretar tal forma com referência ao contexto previamente explicitado. No caso, SUE_3 fala sobre o fato de haver no Brasil muitos destinos de viagem com paisagens, justificando seu interesse por viajar para lugares urbanos que destacam a arquitetura na Europa. Ao usar *this kind of things* dentro do contexto apresentado, o interlocutor é capaz de inferir o sentido intencionado pelo uso de tal linguagem vaga. Interessantemente, SUE_3 é a participante do subcorpus SCoPE² que apresentou menor performance linguística, apesar de seu certificado de proficiência C1. E é exatamente esta participante que se mostra como um *outlier* (fora da curva da média) na contagem de frequência de *kind of*. Das 19 ocorrências verificadas no discurso da SUE_3, 18 foram casos de uso do MP para indicar discrepância de vocabulário (atitudinal) ou para evitar a provisão explícita de vocabulário específico (*Outros*), o que

indica que SUE_3 pode manter uma dependência, ou preferência, por tais funções de *kind of* para compensar possíveis insuficiências linguísticas.

Este estudo confirma o valor pragmático do MP *kind of* em conversação e aponta para o fato de que, mesmo em contextos linguísticos distintos, os dois grupos de brasileiros fazem uso das funções de *kind of* em seus três domínios funcionais. Contudo, nem todos os participantes de ambos os subcorpora utilizam *kind of* pragmaticamente. Seria, contudo, equivocado concluir que os participantes que não fazem uso do MP *kind of* sofram de uma deficiência pragmática em seus processos comunicativos, uma vez que estes participantes podem lançar mão de outras formas que performam as mesmas funções pragmáticas apresentadas neste estudo. De fato, em outro estudo de caso preliminar, Santos (2019) nota que os participantes sob investigação, também de uma amostra do SCoPE², apoiam-se significativamente no MP *like* e em suas funções textuais de reestruturação do discurso.

6 Considerações finais

As considerações finais aqui apresentadas se remetem ao título deste trabalho. Propusemos, como objetivo principal, expor o estado da arte da Pragmática de Corpus, ponto de contato que se originou a partir de duas áreas que, apesar de lidarem com investigações sobre a linguagem em uso, possuem metodologias distintas – a Linguística de Corpus e a Pragmática. Ao fazer isso, inicialmente, contextualizamos a Linguística de Corpus e a Pragmática. Em seguida, apresentamos o percurso da Pragmática de Corpus a partir de seu contexto histórico, discursando sobre sua metodologia de dupla direcionalidade, além de suas limitações e desenvolvimentos. Nesta mesma senda, introduzimos duas abordagens profícuas da nova área, forma-função e função-forma. Finalizamos, então, com um estudo de caso ilustrativo sobre o marcador pragmático *kind of* em dois corpora orais, aplicando o método de filtragem pela abordagem forma-função.

É evidente que diferentes áreas da linguística buscam incorporar, com o passar dos anos, exemplos autênticos da linguagem às suas análises. Vimos que a Linguística de Corpus, por meio de seu desenvolvimento nas últimas décadas e rigor metodológico, dispõe de grande potencial para se comunicar com outros campos do conhecimento, considerando o que estes tradicionalmente já estabeleceram, introduzindo formas

inovadoras e eficazes para a manipulação de dados linguísticos reais em grandes ou pequenas quantidades. É assim que reiteramos o poder analítico da Pragmática de Corpus para a investigação e avanço dos estudos da linguagem, na qual duas disciplinas são beneficiadas em dependência mútua. Ressaltamos que, devido ao seu caráter jovem, a disciplina ainda está em desenvolvimento, principalmente no que diz respeito ao seu modelo metodológico. Sendo assim, novos estudos continuam a ser conduzidos, juntamente com o aperfeiçoamento de técnicas de análise e de anotação pragmática, e, por conseguinte, cada vez mais facilmente tomando a função como ponto de partida. Dessa forma, o que apresentamos aqui é base para o desenvolvimento futuro desta área emergente e promissora.

Agradecimentos

Agradecemos à Prof. Dra. Anne O'Keeffe, do Departamento de Língua Inglesa e Literatura da Mary Immaculate College, Irlanda, pelo contínuo apoio e pela disponibilidade e confiança em nos providenciar a literatura necessária para a escrita deste artigo.
Também agradecemos aos pareceristas por suas avaliações, sugestões e comentários que muito contribuíram para a versão final deste trabalho.

Contribuição dos autores

O artigo foi desenvolvido conjuntamente pelos autores, uma vez que ambos estão inseridos no contexto atual do desenvolvimento da Pragmática de Corpus por parte da escola europeia. Giovani Santos formulou o desenho geral do artigo, e Mateus Miranda auxiliou no afunilamento da proposta. As seções foram divididas entre os autores para serem redigidas, mas se beneficiaram da contribuição de ambos os autores. A análise dos dados também foi feita pelos dois autores de forma conjunta.

Referências

ÄDEL, A.; REPPEN, R. The Challenges of Different Settings: An Overview. *In:* ÄDEL, A.; REPPEN, R. (org.). *Corpora and Discourse*: The Challenges of Different Settings. Amsterdam; Philadelphia: John Benjamins, 2008. p. 1-6. DOI: https://doi.org/10.1075/scl.31

AIJMER, K. *English Discourse Particles*: Evidence from a Corpus. Amsterdam/Philadelphia: John Benjamins, 2002. DOI: https://doi.org/10.1075/scl.10

AIJMER, K. Corpus Pragmatics: from Form to Function. *In:* JUCKER, A. H.; SCHNEIDER, K. P.; BUBLITZ, W. (org.). *Methods in Pragmatics*. Berlin; Boston: De Gruyter Mouton, 2018. p. 555-586. DOI: https://doi.org/10.1515/9783110424928-022

AIJMER, K.; RÜHLEMANN, C. (org.). *Corpus Pragmatics*: A Handbook. Cambridge: Cambridge University Press, 2015. DOI: https://doi.org/10.1017/CBO9781139057493

BEEBE, L.; TAKAHASHI, T.; ULISS-WELTZ, R. Pragmatic Transfer in ESL Refusals. *In:* SCARCELLA, R. C.; ANDERSON, E.; KRASHEN, S. D. (org.). *Developing Communicative Competence in a Second Language*. New York: Newbury, 1990. p. 55-73.

BERBER SARDINHA, T. Beginning Portuguese Corpus Linguistics: Exploring a Corpus to Teach Portuguese as a Foreign Language. *DELTA*, São Paulo, v. 15, n. 2, p. 289-299, 1999. DOI: https://doi.org/10.1590/S0102-44501999000200003

BERBER SARDINHA, T. Linguística de corpus: histórico e problemática. *DELTA*, São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005

BIBER, D.; CONRAD, S.; REPPEN, R. *Corpus Linguistics*: Investigating Language Structure in Use. Cambridge: Cambridge University Press, 1998. DOI: https://doi.org/10.1017/CBO9780511804489

BREZINA, V. *Statistics in Corpus Linguistics*: A Practical Guide. Cambridge: Cambridge University Press, 2018. DOI: https://doi.org/10.1017/9781316410899

BRINTON, L. *Pragmatic Markers in English*: Grammaticalization and Discourse Functions. Berlin: Mouton de Gruyter, 1996. DOI: https://doi.org/10.1515/9783110907582

BUYSSE, L. Discourse Markers in the English of Flemish University Students. *In:* WITCZAK-PLISIECKA, I. (org.). *Pragmatic Perspectives on Language and Linguistics*: Speech Actions in Theory and Applied Studies. Newcastle-upon-Tyne: Cambridge Scholars, 2010. p. 461-484.

CAINES, A.; McCARTHY, M.; O'KEEFFE, A. Spoken Language Corpora and Pedagogical Applications. *In:* FARR, F.; MURRAY, L. (org.). *The Routledge Handbook of Language Learning and Technology*. Abingdon: Routledge, 2016. p. 348-361.

CARTER, R.; McCARTHY, M. *Cambridge Grammar of English*: A Comprehensive Guide. Cambridge: Cambridge University Press, 2006.

CLANCY, B.; McCARTHY, M. Co-Constructed Turn-Taking. *In:* AIJMER, K.; RÜHLEMANN, C. (org.). *Corpus Pragmatics*: A Handbook. Cambridge: Cambridge University Press, 2015. p. 430-453. DOI: https://doi.org/10.1017/CBO9781139057493.023

CLANCY, B.; O'KEEFFE, A. Pragmatics. *In:* BIBER, D.; REPPEN, R. (org.). *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015. p. 235-251. DOI: https://doi.org/10.1017/CBO9781139764377.014

CRESTI, E.; MONEGLIA, M. (org.). *C-ORAL-ROM*: Integrated Reference Corpora for Spoken Romance Languages. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2005. DOI: https://doi.org/10.1075/scl.15

CULPEPER, J. The Metalanguage of Impoliteness: Explorations in the Oxford English Corpus. *In:* BAKER, P. (org.). *Contemporary Corpus Linguistics*. London: Continuum, 2009. p. 64-86.

CULPEPER, J.; HAUGH, M. *Pragmatics and the English Language*. Palgrave Macmillan: Basingstoke, 2014. DOI: https://doi.org/10.1007/978-1-137-39391-3

FARR, F.; BRÓNA, M; O'KEEFFE, A. The Limerick Corpus of Irish English: Design, Description and Application. *Teanga*, Dublin, v. 21, p. 5-29, 2004. DOI: https://doi.org/10.35903/teanga.v21i0.172

FIRTH, J. R. *Papers in Linguistics 1934-1951*. London: Oxford University Press, 1957.

GILQUIN, G. Hesitation Markers among EFL Learners: Pragmatic Deficiency or Difference? *In:* ROMERO-TRILLO, J. (org.). *Corpus Linguistics and Pragmatics:* A Mutualistic Entente. Berlin: De Gruyter Mouton, 2008. p. 119-149.

GOMES DE MATOS, F. Pós-graduação em lingüística no Brasil: orientações curriculares e output (dissertações). *Boletim da ABRALIN*, Recife, v. 3, p. 81-87, 1982.

GRICE, H. P. Logic and Conversation. *In:* COLE, P.; MORGAN, J. J. P. (org.). *Syntax and Semantics 3:* Speech Acts. New York: Academic Press. 1975. p. 41-58. DOI: https://doi.org/10.1163/9789004368811_003

HAUGH, M. Corpus-based metapragmatics, *In:* JUCKER, A. H.; SCHNEIDER, K. P.; BUBLITZ, W. (org.). *Methods in Pragmatics*. Berlin; Boston: De Gruyter Mouton, 2018. p. 619-644. DOI: https://doi.org/10.1515/9783110424928-023

JUCKER, A. H. Corpus Pragmatics. *In:* ÖSTMAN, J. O.; VERSCHUEREN, J. (org.). *Handbook of Pragmatics*. Amsterdam: John Benjamins, 2013. p. 1-18. DOI: https://doi.org/10.1075/hop.17.cor3

JUCKER, A.; TAAVITSAINEN, I.; SCHNEIDER, G. Semantic Corpus Trawling: Expressions of Courtesy and Politeness in the Helsinki Corpus. *In:* SUHR, C.; TAAVITSAINEN, I. (org.). *Developing Corpus Methodology for Historical Pragmatics.* Studies in Variation, Contacts and Change in English. Helsinki: Helsingin Yliopisto, 2012. v. 11. [s.p.]. Disponível em: http://www.helsinki.fi/varieng/series/volumes/11/jucker_taavitsainen_schneider/. Acesso em: 20 set. 2020.

JUCKER, A. H.; SCHNEIDER, K. P.; BUBLITZ, W. (org.). *Methods in Pragmatics*. Berlin; Boston: De Gruyter Mouton, 2018. DOI: https://doi.org/10.1515/9783110424928

JUCKER, A. H.; SCHREIER, D.; HUNDT, M. (org.). *Corpora*: Pragmatics and Discourse. Amsterdam; New York: Rodopi, 2009. DOI: https://doi.org/10.1163/9789042029101

JUCKER, A. H.; TAAVITSAINEN, I. Diachronic Corpus Pragmatics: Intersections and Interactions. *In:* TAAVITSAINEN, I.; JUCKER, A. H.; TUOMINEN, J. (org.). *Diachronic Corpus Pragmatics*. Amsterdam; Philadelphia: John Benjamins, 2014. p. 3-26. DOI: https://doi.org/10.1075/pbns.243.03juc

KALLEN, J.; KIRK, J. *SPICE-Ireland: A User's Guide*. Belfast: Cló Ollscoil na Banríona, 2012.

KIRK, J. M. Kind of and Sort of: Pragmatic Discourse Markers in the SPICE-Ireland Corpus. *In:* AMADOR-MORENO, C. P.; McCAFFERTY, K.; VAUGHAN, E. (org.). *Pragmatic Markers in Irish English*. Amsterdam; Philadelphia: John Benjamins, 2015. p. 88-113. DOI: https://doi.org/10.1075/pbns.258.04kir

KOESTER, A. Building Small Specialised Corpora. *In:* O'KEEFFE, A.; McCARTHY (org.). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2010. p. 66-79. DOI: https://doi.org/10.4324/9780203856949-6

LEVINSON, S. *Pragmatics*. Cambridge: Cambridge University Press, 1983.

LOVE, R. *Overcoming Challenges in Corpus Construction:* The Spoken British National Corpus 2014. New York: Routledge, 2020. DOI: https://doi.org/10.4324/9780429429811

MAYNARD, C.; LEICHER, S. Pragmatic Annotation of an Academic Spoken Corpus for Pedagogical Purposes. *In:* FITZPATRICK, E. (org.). *Corpus Linguistics beyond the Word:* Corpus Research from Phrase to Discourse. Amsterdam: Rodopi, 2007. p. 107-116. DOI: https://doi.org/10.1163/9789401203845_008

McALLISTER, P. G. Speech Acts: A Synchronic Perspective. *In:* AIJMER, K.; RÜHLEMANN, C. (org.). *Corpus Pragmatics*: A Handbook. Cambridge: Cambridge University Press, 2015. p. 29-51. DOI: https://doi.org/10.1017/CBO9781139057493.003

McCARTHY, M.; O'KEEFFE, A. Historical Perspective: What Are Corpora and How Have They Evolved? *In:* O'KEEFFE, A; McCARTHY, M. (org.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, 2010. p. 3-13. DOI: https://doi.org/10.4324/9780203856949-1

McCARTHY, M.; O'KEEFFE, A. Spoken Grammar. *In:* CELCE-MURCIA, M.; BRINTON, D. M.; SNOW, M. A. (org.). *Teaching English as a Second or Foreign Language*. 4. ed. Boston: National Geographic Learning, 2014. p. 271-287.

McENERY, T.; HARDIE, A. *Corpus Linguistics:* Method, Theory and Practice. Cambridge: Cambridge University Press, 2012. DOI: https://doi.org/10.1017/CBO9780511981395

MIRANDA, M. 'Dreams Seem Kind of Utopic': Vague Category Markers in a Learner Corpus. *Revista Intercâmbio*, São Paulo, v. 44, p. 84-107, 2020.

MORRIS, C. W. *Foundations of the Theory of Signs*. Chicago: University of Chicago Press, 1938.

O'KEEFFE, A. Corpus-Based Function-to-Form Approaches. *In:* JUCKER, A. H.; SCHNEIDER, K. P.; BUBLITZ, W. (org.). *Methods in Pragmatics*. Berlin; Boston: De Gruyter Mouton, 2018. p. 587-618. DOI: https://doi.org/10.1515/9783110424928-023

O'KEEFFE, A.; CLANCY, B.; ADOLPHS, S. *Introducing Pragmatics in Use*. London: Routledge, 2011.

O'KEEFFE, A.; CLANCY, B.; ADOLPHS, S. *Introducing Pragmatics in Use*. 2. ed. rev. e aum. Abingdon; New York: Routledge, 2020. DOI: https://doi.org/10.4324/9780429342950

O'KEEFFE, A.; McCARTHY, M.; CARTER, R. *From Corpus to Classroom:* Language Use and Language Teaching. Cambridge: Cambridge University Press, 2007. DOI: https://doi.org/10.1017/CBO9780511497650

O'KEEFFE, A; McCARTHY, M. (org.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, 2010.

PRODROMOU, L. *English as a Lingua Franca*: A Corpus-Based Analysis. London: Continuum, 2008. DOI: https://doi.org/10.1093/elt/ccn064

RAJAGOPALAN, K. Os caminhos da pragmática no Brasil. *DELTA*, São Paulo, v. 15, número especial, p. 323-338, 1999. DOI: https://doi.org/10.1590/S0102-44501999000300013

RAJAGOPALAN, K. Repensar o papel da linguística aplicada. *In:* LOPES, L. P. M. (org.). *Por uma linguística aplicada indisciplinar*. São Paulo: Parábola, 2006. p. 149-168.

RAJAGOPALAN, K. Pragmática. *In:* MOLLICA, M. C.; FERRAREZI, C. J. (org.). *Sociolinguística e sociolinguísticas*. São Paulo: Contexto, 2016. p. 197-204.

RASO, T. O corpus C-ORAL-BRASIL. *In:* RASO, T.; MELLO, H. (orgs.). *C-ORAL-BRASIL I:* Corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012. p. 55-89.

RASO, T. Aspectos sociais e pragmáticos da linguística de corpora. *In:* MOLLICA, M. C.; FERRAREZI, C. J. (org.). *Sociolinguística e sociolinguísticas*. São Paulo: Contexto, 2016. p. 205-216.

RASO, T.; MELLO, H. (org.). *C-ORAL-BRASIL I:* corpus de referência do português brasileiro falado informal. Belo Horizonte: Editora UFMG, 2012.

ROMERO-TRILLO, J. (org.). *Corpus Linguistics and Pragmatics:* A Mutualistic Entente. Berlin: Walter de Gruyter, 2008a. DOI: https://doi.org/10.1515/9783110199024

ROMERO-TRILLO, J. Introduction. *In:* _____. (org.). *Corpus Linguistics and Pragmatics:* A Mutualistic Entente. Berlin: De Gruyter Mouton, 2008b. p. 1-10. DOI: https://doi.org/10.1515/9783110199024

RÜHLEMANN, C. *Corpus Linguistics for Pragmatics:* A Guide for Research. Abingdon; New York: Routledge, 2019. DOI: https://doi.org/10.4324/9780429451072

RÜHLEMANN, C.; AIJMER, K. Corpus Pragmatics: laying the foundations. *In:* AIJMER, K.; RÜHLEMANN, C. (org.). *Corpus Pragmatics*: a handbook. Cambridge: Cambridge University Press, 2015. p. 1-26. DOI: https://doi.org/10.1017/CBO9781139057493.001

RÜHLEMANN, C.; CLANCY, B. Corpus Linguistics and Pragmatics. *In:* ILIE, C.; NORRICK, N. R. (org.). *Pragmatics and Its Interfaces*. Amsterdam; Philadelphia: John Benjamins, 2018. p. 241-266. DOI: https://doi.org/10.4324/9780429451072

SANTOS, G. Second Language Pragmatics: A Corpus-Based Study of the Pragmatic Marker Like. *Letrônica*, Porto Alegre, v. 12, n. 4, p. 1-16, 2019. DOI: https://doi.org/10.15448/1984-4301.2019.4.34002

SANTOS, G. Designing and Building SCoPE²: A Spoken Corpus of Brazilian Portuguese and L2-English. *Research in Corpus Linguistics*, Murcia, v. 8, n. 1, p. 49-64, 2020. DOI: https://doi.org/10.32714/ricl.08.01.04

SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

SINCLAIR, J. *Reading Concordances*. London: Pearson; Longman, 2003.

SINCLAIR, J. *Trust the Text:* Language, Corpus and Discourse. Abingdon: Routledge, 2004.

SINCLAIR, J. Corpus and Text: Basic Principles. *In:* WYNNE, M. (org.) *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford: Oxbow Books, 2005. p. 1-16.

SINCLAIR, J.; FOX, G.; BULLEN, S.; MANNING, E. *Collins-COBUILD English Language Dictionary*. London: Collins, 1987.

SPERBER, D.; WILSON, D. *Relevance*: Communication and Cognition. Oxford: Blackwell, 1995.

TAAVITSAINEN, I.; JUCKER, A. H.; TUOMINEN, J. (org.). *Diachronic Corpus Pragmatics*. Amsterdam; Philadelphia: John Benjamins, 2014. DOI: https://doi.org/10.1075/pbns.243

TAGNIN, S. Corpora on-line. *In:* VIANA, V.; TAGNIN, S. E. O. (org.). *Corpora no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2010. p. 354-361.

VAUGHAN, E. "Got a Date or Something?": An Analysis of the Role of Humour and Laughter in the Workplace Meetings of English Language Teachers. *In:* ÄDEL, A.; REPPEN, R. (org.). *Corpora and Discourse*: The Challenges of Different Settings. Amsterdam; Philadelphia: John Benjamins, 2008. p. 95-115. DOI: https://doi.org/10.1075/scl.31.07vau

VAUGHAN, E.; CLANCY, B. Small Corpora and Pragmatics. *In:* ROMERO-TRILLO, J. (org.). *Yearbook of Corpus Linguistics and Pragmatics 2013:* New Domains and Methodologies. Dordrecht: Springer, 2013. p. 53-73. DOI: https://doi.org/10.1007/978-94-007-6250-3_4

WEISSER, M. Speech Act Annotation. *In:* AIJMER, K.; RÜHLEMANN, C (org.). *Corpus Pragmatics:* A Handbook. Cambridge: Cambridge University Press, 2015. p. 84-113. DOI: https://doi.org/10.1017/CBO9781139057493.005

WEISSER, M. The DART Annotation Scheme: Form, Applicability & Application. *Studia Neophilologica,* [*S.l.*], v. 91, p. 131-153, 2019. DOI: https://doi.org/10.1080/00393274.2019.1616218

WOLFSON, N. Compliments in Cross-Cultural Perspective. *TESOL Quarterly*, [*S.l.*], v. 15, p. 117-124, 1981. DOI: https://doi.org/10.2307/3586403

YULE, G. *Pragmatics*. Oxford: Oxford University Press, 1996.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1137-1189, 2021

Linguística de Corpus aplicada à Semântica de Frames: investigando conceptualizações pró-escolha no debate da Sugestão Legislativa nº. 15/2014

Corpus Linguistics applied to Frame Semantics: investigating pro-choice conceptualizations in SUG nº. 15/2014's debate

Aline Nardes dos Santos
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, Rio Grande do Sul / Brasil
aline.nardes@gmail.com
https://orcid.org/0000-0002-9302-484X

Rove Chishman
Universidade do Vale do Rio dos Sinos (Unisinos), São Leopoldo, Rio Grande do Sul / Brasil
rove@unisinos.br
https://orcid.org/0000-0003-2287-5548

Resumo: Este artigo vincula-se a uma tese doutoral cujo objetivo foi compreender, por meio da identificação de diferentes instanciações de *frames* semânticos, as redes de significado que (re)enquadram os direitos humanos e reprodutivos das mulheres no contexto das audiências públicas que debateram a Sugestão Legislativa (SUG) nº 15/2014 – tal proposta visou a regular o aborto nas primeiras doze semanas de gestação (SANTOS, 2020). Especificamente, o texto trata de alguns desdobramentos analíticos possibilitados pela integração da ferramenta de análise qualitativa NVivo ao recurso Sketch Engine, tendo em vista a necessidade de segmentação do *corpus* em unidades temáticas para posterior processamento dos dados no concordanciador. De modo a abordar tal percurso, o artigo discute a identificação de *frames* no discurso dos defensores da proposta da SUG nº 15, cujas escolhas lexicais refletem a conceptualização do abortamento como questão de saúde pública e de justiça social. Como resultados, o artigo destaca que o uso integrado de diferentes ferramentas de análise empírica permite uma descrição baseada em *corpus* que evidencia a dimensão multifacetada do

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1137-1189

frame semântico – uma estrutura sociocognitiva que se constrói nos entrelaçamentos entre léxico, discurso e cognição.

Palavras-chave: Linguística de *Corpus*; Semântica de Frames; Sugestão Legislativa n.º 15/2014; direitos reprodutivos.

Abstract: This article relates to a Ph.D. thesis which aimed at comprehending, throughout the identification of different semantic frame instantiations, the meaning networks that (re)frame women's human and reproductive rights in the context of the public hearings that discussed the SUG no. 15/2014 – such a proposal intended to regulate abortion in the first twelve weeks of pregnancy, in Brazil (SANTOS, 2020). Specifically, the text presents some analytical developments made available by the integration of the qualitative analysis tool NVivo to the Sketch Engine tool, considering the need of a corpus segmentation into thematic units for a later processing of these data in a concordancer. In order to discuss this process, the article describes the identification of frames within the discourse of the ones that advocate for the SUG proposal, whose lexical choices reflect the conceptualization of abortion as a public heath matter, as well as a social justice one. Concerning the results, the article emphasizes that the integrated usage of different tools devoted to empirical analysis allows a corpus-based description that reveals the multifaceted dimension of a semantic frame – a socio-cognitive structure that is built in the interconnections between lexicon, discourse and cognition.

Keywords: Corpus Linguistics; Frame Semantics; SUG no. 15/2014; reproductive rights.

Recebido em 10 de outubro de 2020

Aceito em 18 de novembro de 2020

1 Introdução

Os caminhos da Linguística de Corpus (LC) e da Semântica de Frames (FILLMORE, 1982, 1985) têm seus pressupostos epistemológicos entrecruzados desde os primórdios da teoria fillmoriana. Afinal, a Semântica de Frames, muito antes de integrar oficialmente o escopo da Linguística Cognitiva, constituiu-se como proposta que visava a compreender estruturas sintáticas por meio de "requisitos contextuais"[1] (FILLMORE, 1975, p. 130) que evidenciam o *continuum* entre léxico, sintaxe e semântica. Em tal percurso, a teoria do autor rompeu com postulados gerativistas que relegavam o léxico ao "asilo dos fora da

[1] "contextual requirements".

lei" da Linguística (SALOMÃO, 2006, p. 7), delineando uma primeira versão do conceito de *frame* como um "sistema de *escolhas linguísticas*"[2] (FILLMORE, 1975, p. 124, grifo nosso), aspecto que situa a teoria, desde suas primeiras versões, no campo dos Modelos Baseados no Uso, que partem do princípio de que "eventos de uso são a fonte de todas as unidades linguísticas" (LANGACKER, 2008, p. 220). Além disso, a criação da FrameNet – um recurso lexicográfico-computacional baseado em *frames* – a partir de evidências provenientes da LC como metodologia consolidou a Semântica de Frames como modelo empírico de análise sistemática de dados linguísticos.

Em suas trajetórias convergentes, tanto a Linguística de Corpus quanto a teoria da Semântica de Frames têm sido incorporadas à proposta de novas interfaces que buscam dar conta da complexidade de fenômenos linguísticos na contemporaneidade, os quais têm como gênese os usos feitos pelo "sujeito interativo" (MIRANDA, 2001, p. 59) ao construir significados "no curso de sua interação comunicativa" (SALOMÃO, 1997, p. 26). Nessa direção, estudos como os de Vereza (2016a, 2016b) reivindicam o estabelecimento de um *continuum* entre cognição e discurso, no contexto de uma "virada cognitivo-discursiva" (VEREZA, 2016a, p. 22) de análises realizadas nesse contexto.

É seguindo tal perspectiva que buscamos investigar as redes de significado construídas em audiências públicas da Sugestão Legislativa (SUG) n.º 15, compreendendo *frames* semânticos como construtos cognitivo-discursivos que podem ser perspectivados de maneiras diferentes, consoante os propósitos comunicativos dos falantes (MIRANDA, 2001; SALOMÃO, 2009; TOMASELLO, 1999, 2008). Tal empreendimento também implicou levar em conta as dimensões macro e microcontextuais (HANKS, 2008) do *corpus*, o que nos levou a integrar uma ferramenta de análise qualitativa de dados – o NVivo – ao Sketch Engine (SE), tendo em vista a necessidade de segmentação dos dados em unidades temáticas para posterior processamento dos dados no concordanciador.

Considerando tais aspectos, este artigo visa a discutir os desdobramentos analíticos de tal proposta investigativa, com vistas a evidenciar a necessidade de novos entrelaçamentos metodológicos entre Linguística de Corpus e Semântica de Frames, visando a dar conta de descrições que integrem diferentes perspectivas sobre os mesmos dados

[2] "system of linguistic choices".

empíricos. Para isso, o texto se estrutura da seguinte forma: na seção 2, delineamos a noção cognitivo-discursiva de *frame* semântico, salientando o caráter multifacetado desse construto no âmbito dos estudos linguísticos. Na seção 3, contextualizamos o *corpus* de estudo da SUG nº 15, composto de transcrições de audiências públicas que debateram a Sugestão. Na seção 4, com o objetivo de elucidar as possibilidades analíticas propiciadas por tal proposta, discutimos a identificação de *frames* no discurso dos defensores da proposta da SUG nº 15, cujas escolhas lexicais refletem a conceptualização do abortamento como questão de saúde pública e de justiça social. Por fim, na seção 5, trazemos as considerações finais.

2 Referencial teórico: *frames* fillmorianos e(m) discurso

Nos primeiros textos de Charles Fillmore acerca da noção de *frame* semântico, menciona-se a pertinência dessa estrutura em sua dimensão mais interacional, embora ele mesmo categorize algumas de suas contribuições como meras notas sugestivas (FILLMORE, 1975). Trata-se de reflexões esparsas, publicadas entre as décadas de 1970 e 1980, fase na qual seus artigos ainda especulavam o possível alcance da Semântica de Frames quanto a suas contribuições para estudos linguísticos.

Por exemplo, no artigo "Frame Semantics and the Nature of Language", Fillmore (1976) aborda a relevância do processo de *framing* para a compreensão do funcionamento da linguagem humana, considerando que *frames* são sempre ativados "[...] na percepção, no pensamento e na *comunicação*" (FILLMORE, 1976, p. 20, grifo nosso).[3] Nesse texto, o autor situa a Semântica de Frames como abordagem contextualista do significado, considerando que as teorias então vigentes necessitavam incluir

> [...] uma atenção à *importância das funções sociais da linguagem*, uma preocupação com a natureza da produção da fala e com processos de compreensão, bem como um interesse *nas relações entre o que um falante diz e o contexto no qual ele diz isso*. (FILLMORE, 1976, p. 23, grifo nosso).[4]

[3] "[…] in perceiving, thinking, and communicating".

[4] "[…] an awareness of the importance of the social functions of language, a concern with the nature of the speech production and comprehension processes, and an interest in the relationships between what a speaker says and the context in which he says it."

No que se refere ao termo *contexto*, em seu sentido amplo, valemo-nos dos apontamentos de Morato (2010) acerca da noção de *frame* como mais um construto teórico que visa a dar conta de como os falantes disseminam e partilham de significados a partir de sua experiência de mundo, para além de aspectos relativos ao contexto verbal de produção, ou cotexto. Nesse sentido, tendo em vista que o *frame* não é apenas uma estrutura de conhecimento dissociada da interação, mas que também diz respeito às práticas sociais envolvidas, é pertinente categorizá-lo como estrutura de expectativa, considerando que, com base em sua interação com a realidade, "o sujeito organiza o conhecimento sobre o mundo e usa esse conhecimento para prever interpretações e relações referentes a novas informações, eventos e experiências" (TANNEN, 1979, p. 138-139).[5] Tal dimensão da estrutura do *frame* a torna "[...] dinâmica, uma vez que é continuamente confrontada com a experiência e revista" (MIRANDA, 1999, p. 82), bem como "[...] construída e modelada em situações de interação social" (DUQUE, 2015, p. 40).

É importante pontuar que os *frames*, em sua dimensão sociocultural, são sempre, em alguma medida, resultado de experiências que se moldam e se reconstroem na interação. Como observam Koch, Morato e Bentes (2011, p. 82), é necessário considerar a natureza sociocultural de tais construtos:

> A noção de contexto, como a de situação social, enquadre ou *frame*, tem a ver com estruturas de expectativa, isto é, não se trata de algo concebido *a priori* e nem de forma independente quanto a nossas experiências socioculturais; pelo contrário, dependem dos atos de significação e, portanto, das práticas mediadas largamente por linguagem.

Nessa perspectiva, a teoria da cognição social humana desenvolvida por Tomasello (1999, 2003, 2008) tem verificado empiricamente os fundamentos epistemológicos que sustentam o conceito contemporâneo de cognição defendido pela Linguística Cognitiva, o qual enfatiza a faceta mais sociointeracional de estruturas como os *frames* semânticos: uma cognição social, pautada na capacidade humana para o engajamento e para o reconhecimento das intencionalidades do outro no curso da comunicação

[5] "[…] based on one's experience of the world in a given culture (or combination of cultures), one organizes knowledge about the world and uses this knowledge to predict interpretations and relationships regarding new information, events, and experiences."

(BOOTH, 2016; MIRANDA, 2001; SALOMÃO, 2006). Dessa maneira, estudos como os do autor revelam que nossa cognição é primordialmente social, calcada no reconhecimento de intencionalidades e no engajamento em situações de atenção conjunta, de modo que a ontogênese humana tem como primeiro grande desdobramento o reconhecimento do outro como agente intencional – aspecto que leva o sujeito a, desde muito antes de falar, "[...] tentar manipular os estados intencionais e mentais do outro para vários fins cooperativos e competitivos"[6] (TOMASELLO, 2003, p. 12). Nesse sentido, os *frames* e os símbolos linguísticos que os evocam não são somente aprendidos socialmente, como também são perspectivados pelo sujeito, "[...] dependendo de seus propósitos comunicativos [...]"[7] (TOMASELLO, 2003, p. 12). É esse aspecto sociocultural do contexto e do uso da língua que Geeraerts, Kristiansen e Peirsman (2010, p. 3) consideram como crucial a análises semântico-cognitivas na contemporaneidade, dado que estruturas conceptuais, dentre elas os *frames*, são manipuladas por meio de processos socioculturalmente situados de cognição.

No cenário brasileiro, destacamos a inserção da Hipótese Sociocognitiva da Linguagem (MIRANDA, 2001; SALOMÃO, 1997) nesse contexto de preconização de uma Linguística Cognitiva mais social. Tal posicionamento valoriza o papel da interação no processo de construção de significados, como explica Salomão (1997, p. 26): "A hipótese que [...] adotamos advoga ser a significação *uma construção mental produzida pelos sujeitos cognitivos no curso de sua interação comunicativa*." Complementando esse aspecto, Silva (2015, p. 67) elenca três grandes dimensões a serem abrangidas em estudos sociocognitivos:

> [...] (i) as interações socioculturais e o modo como elas afetam o discurso; (ii) os processos cognitivos de interação discursiva; e (iii) a relação entre as dimensões conceptuais, as dimensões interacionais e as dimensões socioculturais da linguagem *em uso* (grifo nosso).

É com tal perspectiva que descrevemos, na seção a seguir, o *corpus* da SUG n.º 15, as ferramentas utilizadas para explorá-lo e as etapas analíticas empregadas para investigar conceptualizações pró-escolha nesse contexto.

---

[6] "[…] to attempt to manipulate one another's intentional and mental states for various cooperative and competitive purposes."

[7] "[…] depending on her communicative goal".

3 Percurso metodológico

Conforme abordamos na seção anterior, a noção de *frame* semântico é central ao nosso escopo analítico, bem como a sua integração aos chamados *modelos baseados no uso* (BYBEE, 2012; LANGACKER, 2008; TOMASELLO, 2003), que defendem "[...] a natureza dialética da relação entre o uso da língua e seu sistema. [...] De acordo com essa visão, é possível adquirir conhecimentos sobre o sistema linguístico por meio da análise de eventos de uso que o instanciam"[8] (GEERAERTS; KRISTIANSEN; PEIRSMAN, 2010, p. 4). É a partir de tal postulado que *frames* semânticos são descritos com base na Linguística de Corpus como metodologia que propicia a identificação de "*formas linguísticas*" que ativam "*estruturas cognitivas* – os *frames*" (FILLMORE; BAKER, 2010, p. 314).

Com base nesses pressupostos, a próxima subseção descreve nosso *corpus* de estudo e apresenta as ferramentas computacionais que utilizamos para explorá-lo – nomeadamente, o NVivo e o Sketch Engine. Na segunda subseção, delineamos as etapas de análise que são empregadas para atingirmos o objetivo proposto.

3.1 Contextualizando o *corpus*: as transcrições das audiências públicas da SUG nº 15/2014

Em setembro de 2014, no portal e-Cidadania, foi criada a ideia legislativa nº 29.984, que deu origem à SUG nº 15. A proposta central foi assim redigida: "Regular a interrupção voluntária de gravidez, nas primeiras doze semanas de gestação, pelo Sistema Único de Saúde" (BRASIL, 2014). A motivação inicial dessa proposta teve como foco principal o abortamento clandestino como questão de saúde pública, tendo em vista a mortalidade de mulheres pobres em procedimentos de aborto inseguro. Essa ideia legislativa de iniciativa popular, ao receber o apoio de 20 mil manifestações individuais, passou a ser debatida no âmbito do Senado, por meio de audiências públicas que foram registradas em vídeo e cujas transcrições estão disponíveis *online*. Assim, este estudo se restringiu a essas transcrições, disponibilizadas no formato de atas

[8] "[…] the dialectic nature of the relation between language use and the language system. […] According to this view, one can gain insight into the language system by analyzing the usage events that instantiate it."

de reunião, que debateram a Sugestão Legislativa nº 15 entre maio de 2015 e abril de 2016.

Os cinco arquivos – um para cada audiência – foram disponibilizados no formato RTF (Rich Text Format), compatível com todas as versões do Microsoft Word e com editores mais simples, como o WordPad. A imagem a seguir exibe os arquivos baixados, mantendo sua nomenclatura original:

FIGURA 1 – Formato original das transcrições da SUG 15

Fonte: Elaborada pelas autoras.

Ao todo, o *corpus* na íntegra tem 169,253 *tokens* e 140,428 *types*, totalizando cerca de 230 páginas.[9] Considerando a extensão média do material, que pode ser considerado um *corpus* pequeno (BERBER SARDINHA, 2000), foi possível fazer uma leitura integral dos dados antes de começar a manipulá-los – tarefa que não é possível quando a extensão considerável do *corpus* permite apenas sua manipulação por meio de ferramentas digitais. Ao encontro disso, Koester (2010, p. 67) elenca as seguintes vantagens de se trabalhar com *corpora* pequenos em âmbitos como esse: "[...] eles permitem um elo mais forte entre o *corpus* e os contextos nos quais os textos do *corpus* foram produzidos."[10]

Como se tratou de um estudo de caso – pois sua análise está restrita a um objeto específico, "[...] de maneira a permitir o seu conhecimento amplo e detalhado" (GIL, 2008, p. 58) –, consideramos necessária uma etapa de maior familiarização com o conteúdo do *corpus*, ainda que de forma preliminar ao exercício analítico propriamente dito. Nesse processo, separamos do material as falas dos senadores que presidiram as seções, os quais abriam as audiências lendo um texto que retomava

[9] Considerando uma lauda com fonte Arial 12, com espaçamento simples.

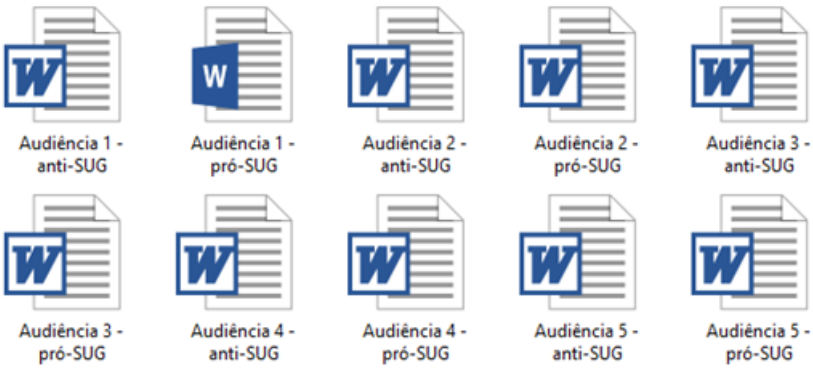
[10] "[…] they allow a much closer link between the corpus and the contexts in which the texts in the corpus were produced."

a pauta da SUG e apresentando os convidados. Visto que suas falas se reduziam, em sua maioria, a estabelecer o protocolo a ser seguido nas sessões, não as selecionamos para compor o *corpus* de estudo.

Essa etapa do processo de preparação do *corpus* a partir de uma leitura preliminar ainda incluiu o mapeamento dos participantes de acordo com sua categoria de participação: se eram painelistas convidados, que expuseram seu posicionamento ao longo de 15 minutos; ou se eram participantes que pediram direito de fala ao final das exposições finais, cujo tempo-limite para arguição era de três minutos – em alguns casos, houve a participação dos próprios painelistas na sessão final, geralmente para reiterar os principais pontos de sua apresentação. Inicialmente, separamos os convidados painelistas e os demais participantes das sessões nas grandes categorias "pró-SUG" e "anti-SUG" – ou seja, consideramos aqueles que defendem a regulação do aborto nas 12 primeiras semanas de gestação como pró-SUG; e aqueles que se opõem à proposta, como anti-SUG. Salientamos que, neste artigo, tratamos somente das conceptualizações dos defensores da SUG, considerando que esses resultados são pertinentes para discutirmos nosso percurso cognitivo-discursivo de identificação de *frames* com base em *corpus*.

A Figura 2 exibe a segunda segmentação realizada, a partir da qual realizamos uma terceira classificação, por participantes, com vistas a rastrear algumas características dos dados ao longo do processamento do *corpus*.

FIGURA 2 – Segmentação do *corpus* por participantes pró-SUG e anti-SUG



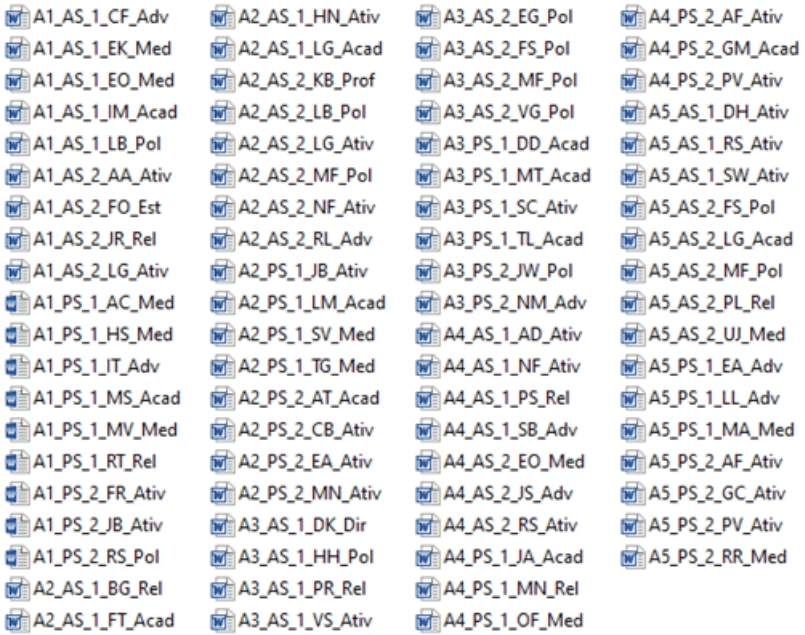
Fonte: Elaborada pelas autoras.

Como mostra a Figura 3, criamos um padrão de nomeação conforme os exemplos a seguir:

A1_PS_1_MV_Med – A3_AS_2_EG_Pol

Nesses dois casos, temos a seguinte notação: A1 e A3 = audiência pública 1 e audiência pública 3 (a numeração vai até 5); PS e AS = pró-SUG e anti-SUG; MV e EG = iniciais das respectivas participantes; e Med e Pol = iniciais do grupo socioprofissional que representam – médico(a) e político(a) – ao se manifestarem na respectiva audiência, conforme as credenciais incluídas nas atas e reproduzidas pelos próprios painelistas. Assim, trata-se dos papéis institucionais (LANGLOTZ, 2015) que os participantes desempenham nesse âmbito.

FIGURA 3 – Segmentação do *corpus* por audiência pública, participante, posicionamento e papel institucional



Fonte: Elaborada pelas autoras.

Ao todo, chegamos a um total de 78 participantes. Após esse mapeamento, considerando a primeira leitura do material, organizamos quadros com uso do Microsoft Excel, no qual sistematizamos os dados em cinco colunas, quais sejam: nome do participante; papel institucional; posicionamento em relação à SUG (contra ou a favor da proposta); resumo da sua exposição em uma frase; e palavras-chave coletadas ao longo da leitura, com o objetivo de mapear preliminarmente os temas abordados por cada sujeito.

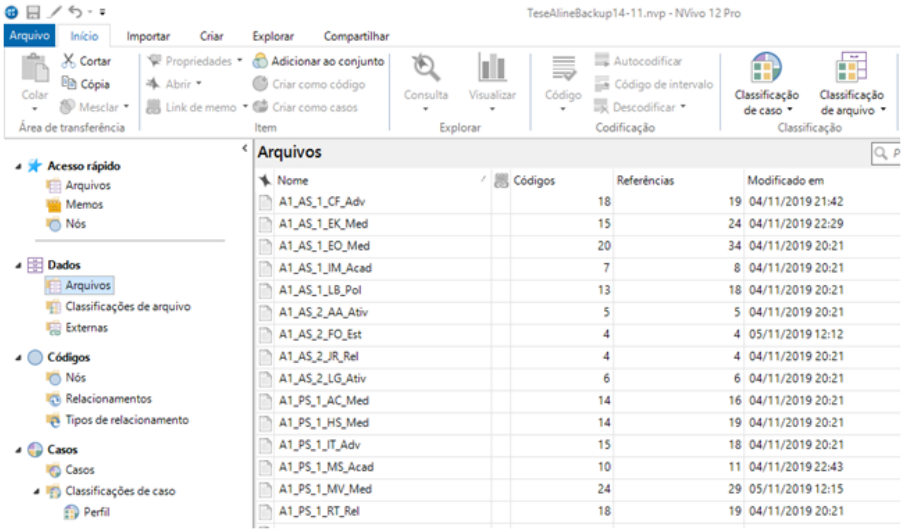
Após a preparação e a compilação das transcrições, pudemos avançar na manipulação dos dados por meio das ferramentas computacionais NVivo e Sketch Engine.

3.2 Ferramentas de análise empregadas: QSR NVivo e Sketch Engine

O QSR NVivo é uma ferramenta qualitativa de análise de dados que tem sido bastante utilizada nas áreas de Ciências Sociais e Humanas – principalmente na Educação – e de Ciências da Saúde (LAGE, 2011). Trata-se de um *software* que se fundamenta "[...] no princípio da codificação e armazenamento de textos em categorias específicas [...]" (GUIZZO; KRZIMINSKI; OLIVEIRA, 2003, p. 54), permitindo o cruzamento de diversos parâmetros que classificam os dados. A figura a seguir exibe a interface do programa, que precisa ser instalado no computador e exige a compra de uma licença.[11] A versão que utilizamos se chama NVivo 12 Pro. O *software* também dispõe de uma versão de testes, que pode ser utilizada gratuitamente por quinze dias – foi a partir de tal versão que analisamos a pertinência de sua utilização para a análise dos dados da SUG.

[11] Uma licença para estudantes foi adquirida pelo Grupo SemanTec, a qual tem validade de dois anos.

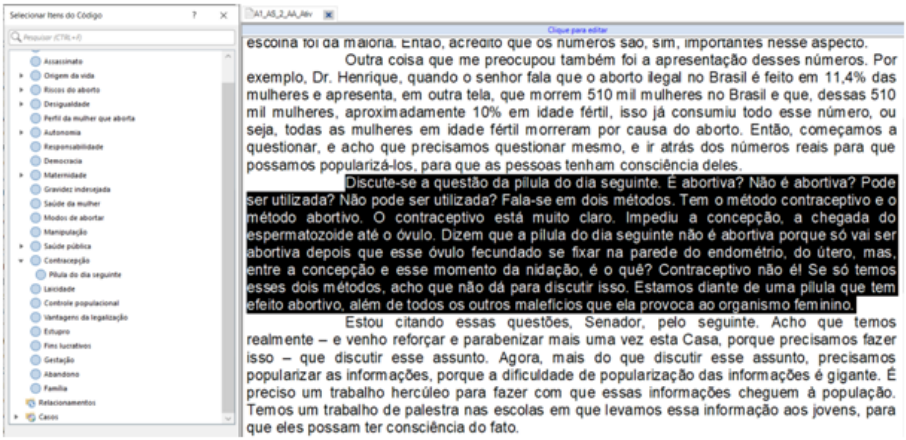
FIGURA 4 – Tela inicial do QSR NVivo



Fonte: Elaborada pelas autoras.

Ao se abrir um dos arquivos de transcrição, é possível marcar excertos do texto e classificá-los nos chamados *nós*, que "[...] representam uma categoria ou ideia abstrata [...]" criada pelo pesquisador (GUIZZO; KRZIMINSKI; OLIVEIRA, 2003, p. 57). No caso de nossa análise, tal recurso foi bastante útil para identificarmos os grandes temas abordados em cada participação. A figura a seguir mostra um caso em que assinalamos um excerto com o nó "Pílula do dia seguinte", dado que o participante manifesta seu posicionamento contrário ao uso da contracepção de emergência. Observamos que um mesmo trecho pode ser ligado a mais de um nó, conforme tais categorias vão sendo criadas ao longo da exploração dos dados.

FIGURA 5 – Exemplo de atribuição de um excerto do *corpus* a um nó

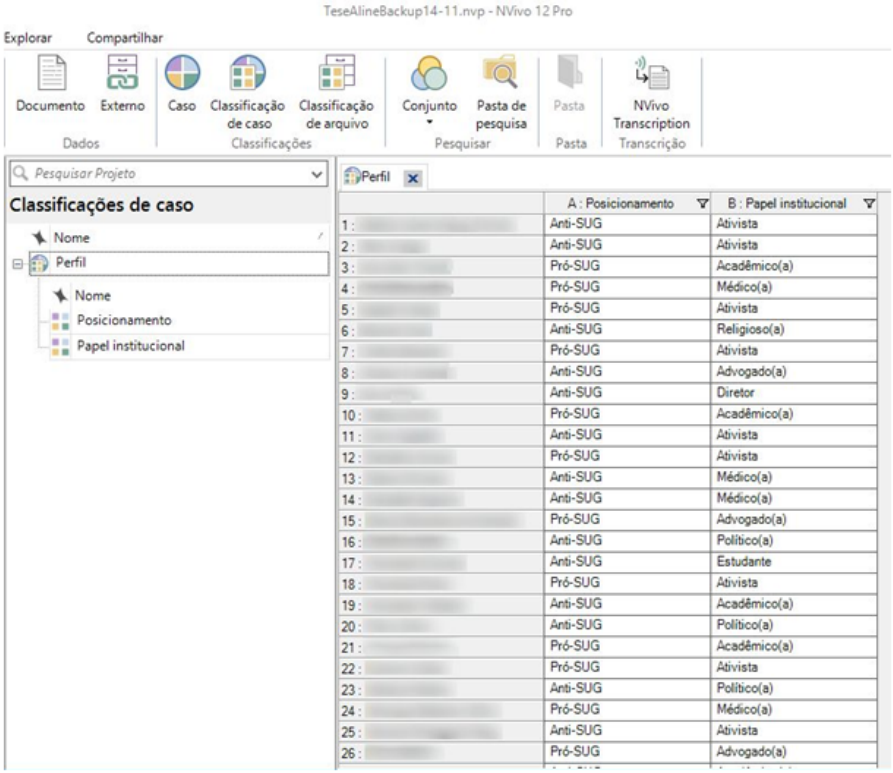


Fonte: Elaborada pelas autoras.

Salientamos que os próprios nós deram origem a pequenos *subcorpora* separados por temas, dentre os quais os maiores (com mais de vinte excertos) foram também processados na ferramenta de *corpus*, com vistas a explorarmos o léxico de forma mais ampla e confirmarmos possíveis evocadores de *frames*.

Os nós do NVivo podem ser cruzados com informações sobre o perfil de cada participante, as quais podem ser previamente cadastradas na ferramenta. Assim, para poder inter-relacionar diferentes informações sobre as audiências, registramos individualmente cada participante, associando cada perfil ao posicionamento (pró-SUG ou anti-SUG) e ao seu papel institucional, conforme o recurso Classificações de Caso, exibido na Figura 6 – o tipo de caso foi denominado Perfil, e suas categorias são Posicionamento e Papel Institucional. Esse procedimento visou a dar conta da noção de *frame* como construto interacional que é manipulado pelos falantes dependendo de seu ponto de vista, conforme discutimos na seção 2. Assim, foi crucial rastrear nos dados quem eram os sujeitos conceptualizadores e qual era o seu posicionamento acerca da proposta da SUG n.º 15.

FIGURA 6 – Recurso Classificações de Caso do NVivo

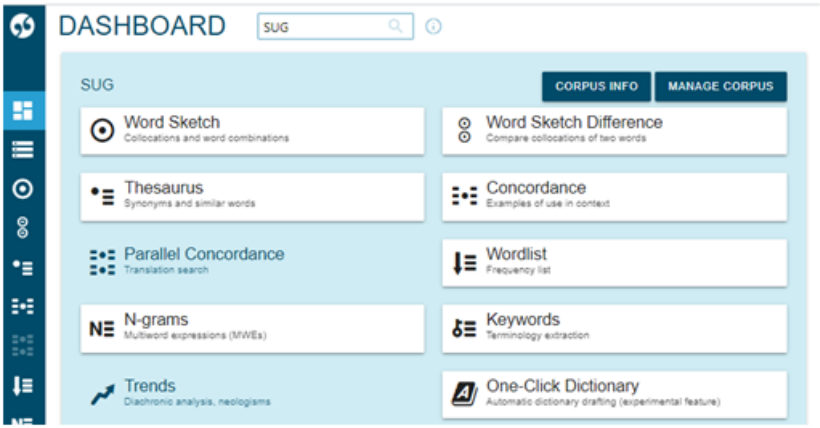


Fonte: Elaborada pelas autoras.

Os resultados da exploração com o uso do NVivo foram atrelados à ferramenta de *corpus* Sketch Engine, um *software* eficiente na manutenção de *corpus* que tem sido utilizado nas pesquisas linguísticas do grupo SemanTec, de que fazemos parte, mostrando-se um profícuo recurso para exploração do léxico em estudos que visam a identificar *frames* semânticos (CHISHMAN *et al.*, 2014, 2015, 2018; SANTOS; CHISHMAN, 2018. Assim como o NVivo, o SE exige a compra de uma licença. Em relação à forma de acesso, a diferença principal entre os recursos está no fato de que o Sketch Engine não necessita ser instalado; sua interface, exibida na Figura 7, está disponível *online*, por meio de inserção de login e senha no seu site.[12]

[12] Endereço para login: https://auth.sketchengine.eu/.

FIGURA 7 – Tela inicial do Sketch Engine

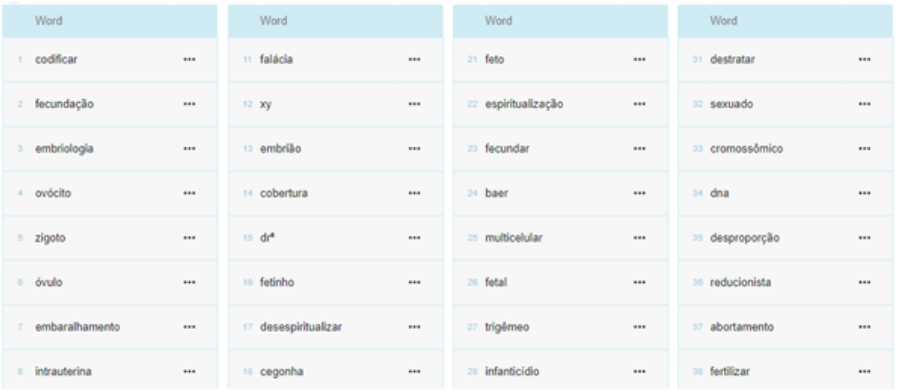


Fonte: Sketch Engine.

Para processamento dos dados, o *corpus* foi carregado para a ferramenta no formato docx. O recurso fez a compilação automática do material, utilizando o etiquetador Freeling, que atribui etiquetas sintáticas aos termos para facilitar buscas por combinatórias. À parte as falas protocolares, que foram excluídas dos dados processáveis, o tamanho do *corpus* é de 114.429 *tokens*. Especificamente, os recursos do programa que utilizamos são elencados na sequência:

a) *Keywords*: permite uma comparação entre as palavras mais frequentes do *corpus* de estudo em relação a um *corpus* maior, ou *corpus* de referência. Como resultado, obtém-se uma lista das palavras-chave do *corpus* de estudo, ou seja, aquelas que são estatisticamente mais proeminentes. Tal recurso foi usado especificamente para processar os *subcorpora* maiores de nós codificados por meio do NVivo.

FIGURA 8 – Lista parcial de palavras-chave do *subcorpus* do nó Origem da Vida



Fonte: Elaborada pelas autoras.

b) *Concordance*: permite a visualização da palavra pesquisada, ou palavra-nó, juntamente com o texto adjacente, ou cotexto. A palavra buscada aparece em destaque na chamada linha de concordância. Além disso, ao se clicar na palavra-nó, o Sketch Engine possibilita visualizar porções maiores de texto antes e após a concordância consultada. Tal recurso foi utilizado principalmente ao longo do processo de anotação semântica das sentenças que evocavam os *frames* analisados.

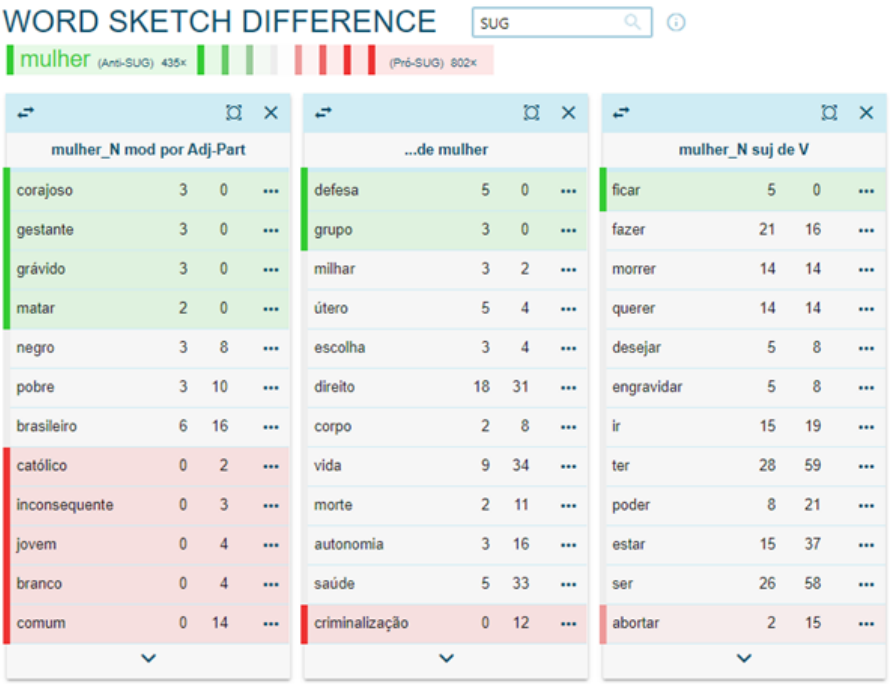
FIGURA 9 – Recorte da ferramenta Concordance para a palavra *mulher*

ios que manter a linha de coerência - é a defesa da	mulher	. </s><s> As grandes vítimas do aborto são duas: as m
</s><s> As grandes vítimas do aborto são duas: as	mulheres	, que estão aqui e me ouvem, e a vida que elas gestam
i e me ouvem, e a vida que elas gestam. </s><s> A	mulher	é vítima de aborto. </s><s> Ela é a grande vítima! </s>
passo seguinte, coerente com essa preservação da	mulher	, à luz do que acabei de lhes trazer sobre essa jovem a
rtalecer - já direi o que há aqui - toda a situação da	mulher	e a criar uma rede protetiva da mulher que se vê absol
situação da mulher e a criar uma rede protetiva da	mulher	que se vê absolutamente abandonada até por seus far
es forem ler, vão ver que se criam casas de apoio à	mulher	nessas situações, à maria abandonada. </s><s> Crian
s> Nós devemos criar casas de acolhida para essa	mulher	. </s><s> Esse é que é o ponto. </s><s> Temos de gar
> Nós não gostamos de abortar. </s><s> Nenhuma	mulher	quer abortar." Todas as minhas irmãs dizem isso. </s>
volvam e desenvolvam a vida que vocês carregam,	mulheres	!" Que você prescinda do machão, deste aí, para que e
ntalmente. </s><s> Esse é o ponto sobre o qual as	mulheres	têm de refletir, meu Deus do céu! </s><s> São mulhere
eres têm de refletir, meu Deus do céu! </s><s> São	mulheres	jovens, mulheres mais maduras, mulheres envelhecida
, meu Deus do céu! </s><s> São mulheres jovens,	mulheres	mais maduras, mulheres envelhecidas. </s><s> A luta
s> São mulheres jovens, mulheres mais maduras,	mulheres	envelhecidas. </s><s> A luta é pelas mulheres! </s><s
ras, mulheres envelhecidas. </s><s> A luta é pelas	mulheres	! </s><s> É difícil entender! </s><s> Não sei como não
Cegonha, desaparece todo esse debate, porque a	mulher	terá toda assistência possível, terá todo amparo possív
e ter atendido, não sei, acho que quatro vezes mais	mulheres	grávidas. </s><s> Então, gostaria de começar falando
iãe não quis aquela gestação, e é terrível quando a	mulher	não quer uma gestação e está grávida. </s><s> Realm
com tudo aquilo que temos de preocupação com as	mulheres	, com as adolescentes que engravidam, a pergunta é:
nça, nunca é um mal menor. </s><s> Em relação à	mulher	, se esquecem de dados fundamentais. </s><s> Em pr

FONTE: Elaborada pelas autoras.

c) *Sketch Difference*: permite a comparação entre as combinatórias de uma palavra no *corpus* como um todo, ou entre o uso do mesmo item lexical em diferentes *subcorpora*. Na figura a seguir, exibimos parcialmente a Sketch Difference para a palavra *mulher*, conforme os dois *subcorpora* correspondentes ao posicionamento (pró-SUG e anti-SUG) dos participantes. Nos extremos e em cores diferentes, constam as combinatórias que só ocorrem em um *corpus*; ao meio e em cor neutra, aparecem as combinatórias comuns a ambos os *subcorpora*.

FIGURA 10 – Sketch Difference para a palavra *mulher*



Fonte: Elaborada pelas autoras.

A próxima seção aborda as etapas metodológicas que adotamos em nosso percurso analítico.

3.3 Etapas de análise

Para investigar conceptualizações pró-escolha no contexto da SUG, tendo em vista uma análise macro e microcontextual do *corpus*, organizamos nosso percurso analítico nas etapas a seguir:

a) **Identificação das temáticas, ou nós, presentes nas falas dos participantes pró-SUG:** considerando o funcionamento da ferramenta NVivo, a qual exige a classificação do *corpus* em nós para processar inter-relações com as demais categorias, valemo-nos desse recurso para mapear as grandes temáticas que constituíam o debate nas audiências, por meio dos seguintes passos:

- releitura de cada arquivo do primeiro *subcorpus*, segmentado por participantes (vide Figura 3), após inseri-lo na interface do NVivo;
- eleção dos excertos que correspondiam a um grande tema e criação do nó correspondente na ferramenta;
- processamento integrado de todos os nós e verificação de sua predominância nas audiências como um todo.

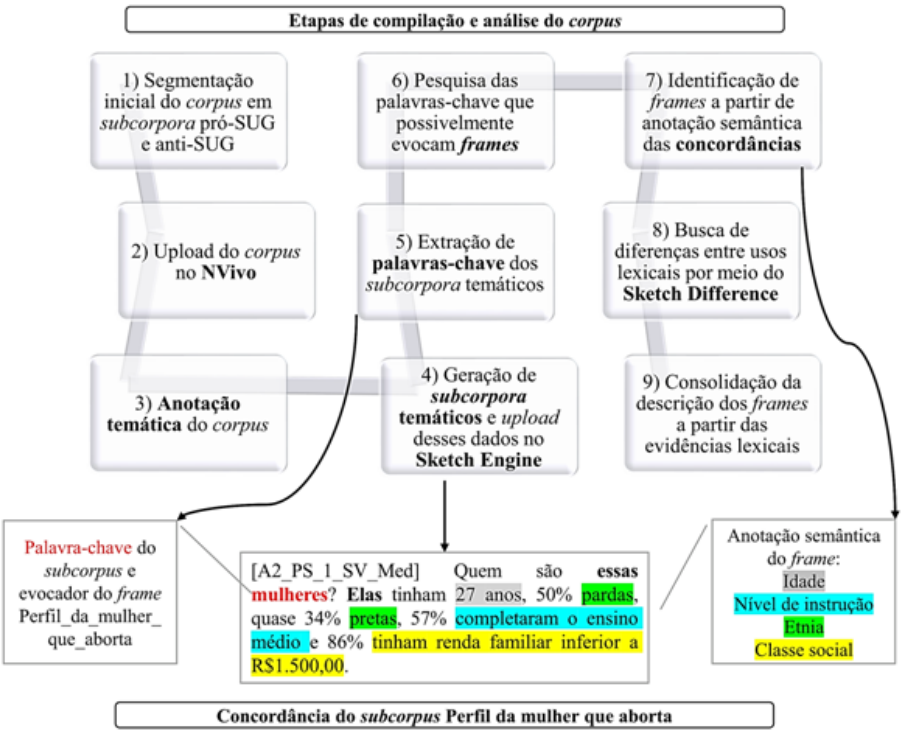
b) Identificação dos frames que emergem do discurso pró-SUG: como discutimos em nossa revisão teórica, *frames* semânticos são identificados por meio do levantamento de "*formas linguísticas*" que ativam "estruturas cognitivas – os *frames*" (FILLMORE; BAKER, 2010, p. 314). A partir de tais usos linguísticos, elencam-se os Elementos de Frame instanciados no cotexto, de modo que a descrição vai-se consolidando conforme a manipulação dos dados avança. Nesse contexto, inspirando-nos em trabalhos como o de Siman (2015, p. 5), recorremos à Plataforma FrameNet Berkeley "[...] como referência (mas não como limite) [...]" para descrever os *frames*. Isso ocorre porque, como demonstram os trabalhos de Chishman *et al.* (2018, 2019), o estudo de um domínio específico demanda a criação de *frames* de acordo com o microcontexto analisado, considerando que a FrameNet é um recurso de língua inglesa que está em permanente construção – e que, por conseguinte, não fornece todos os *frames* que emergem de um *corpus* especializado. Também destacamos que a descrição de *frames*, neste estudo, consiste na identificação da camada semântica dos Elementos de Frame – opção também explicitada nos trabalhos de Siqueira (2013), Lima e Miranda (2013), Miranda e Bernardo (2013), dentre outros –, não havendo o propósito de sistematizar os padrões sintáticos encontrados no *corpus*, nem de identificar EFs *core* e não *core*. Isso significa que não seguimos o modelo de anotação da FrameNet, que se pauta nas valências dos evocadores de *frames*. Assim, em uma anotação que se atém apenas à camada semântica, são anotados os constituintes que instanciam Elementos de Frame independentemente de sua posição na frase em relação ao evocador. Tal procedimento permite a anotação de instanciadores de EFs de forma mais ampla, com a finalidade primeira de "[...] 'remontar' as cenas

conceptuais que emergem do discurso [...]" (FONTES, 2012, p. 37). Especificamente, esse estágio de análise envolveu o uso da ferramenta Sketch Engine e compreendeu as seguintes etapas:

- identificação das unidades lexicais que potencialmente evocavam *frames*, considerando os nós elencados na etapa anterior – para os nós com mais de vinte excertos, consulta às palavras-chave do respectivo *subcorpus* no SE;
- busca dessas unidades lexicais na FrameNet e/ou de *frames* relacionados e, se necessário, descrição de novo *frame*;
- descrição dos Elementos de Frame expressos linguisticamente, ou por meio de Instanciação Nula (casos em que não é instanciado em uma concordância do *corpus*, geralmente porque foi mencionado em uma parte anterior dos dados);
- sistematização das unidades linguísticas que instanciam evocadores e Elementos de Frame, por meio do uso dos recursos da ferramenta Sketch Engine;
- se necessário, refinamento da descrição do *frame* a partir dos dados encontrados por meio da ferramenta de *corpus*.

A figura a seguir ilustra a sequência dos procedimentos analíticos e explicita o papel do NVivo e do SE nesse percurso:

FIGURA 11 – Ilustração do percurso analítico por meio das ferramentas NVivo e Sketch Engine



Fonte: Elaborado pelas autoras.

A próxima seção traz a análise e a discussão dos dados conforme as etapas elencadas.

4 Conceptualizações pró-escolha no debate da Sugestão Legislativa n.º 15/2014: análise dos dados

Seguindo as etapas metodológicas elencadas na seção anterior, o percurso de análise aqui discutido foi dividido em dois estágios. Na primeira etapa (seção 4.1), analisamos as temáticas ou *nós* do discurso dos participantes pró-SUG, obtidos por meio da exploração do *corpus* no NVivo. Por sua vez, a seção 4.2 concerne à análise dos *frames* a partir dos nós elencados, que possibilitou explorarmos o léxico que os evoca por meio do Sketch Engine.

4.1 Os temas que permeiam o discurso pró-SUG: explorando os nós do NVivo

Conforme já referido, a exploração inicial do *corpus*, após a leitura preliminar, foi realizada com auxílio da ferramenta QSR NVivo, que permite a classificação dos dados em *nós*; trata-se dos temas abordados que eventualmente originaram categorias de análise – no caso deste estudo, serviram como o ponto de partida para a descrição de *frames*. Além disso, um mesmo excerto do texto pôde ser marcado com um ou mais nós, os quais foram cadastrados na ferramenta ao longo da exploração dos textos. Analogamente, assim como a descrição de *frames* baseada nos trabalhos de Fillmore (1982, 1985) envolve a anotação semântica de Elementos de Frame, o trabalho com o NVivo parte de uma "anotação temática" do *corpus* como forma de se realizar um mapeamento geral dos dados, o qual foi se refinando a partir desse primeiro exercício analítico.

Nesse sentido, tal processo envolveu o cadastro de diferentes nós no recurso, dentre os quais elencamos, no quadro a seguir, aqueles que emergiram do *corpus* pró-SUG. Salientamos que, conforme a leitura foi avançando, eventualmente eliminamos, refinamos, renomeamos e/ou reorganizamos os nós, considerando o andamento do trabalho. Nessa etapa, já havíamos realizado um mapeamento do *corpus* por palavras-chave, que também serviu como ponto de partida para a codificação realizada no NVivo. Além disso, também levamos em conta a frequência com que cada temática ocorria no *corpus* ao selecionarmos os excertos – ligados a nós – a partir dos quais identificamos *frames* semânticos.

QUADRO 1 – Lista dos nós encontrados e breve contextualização de cada um

Nome do nó	Breve contextualização	Exemplo
Assassinato	Descreve o aborto como assassinato do feto.	"O abortamento mata uma criança inocente que não pode se defender."
Autonomia	Aborda o conceito de autonomia sob a ótica da autonomia da mulher.	"A sociedade patriarcal tem nos negado a autonomia sobre os nossos corpos e a nossa sexualidade e nos tratado como meras reprodutoras do sistema."
Escolha [subnó de Autonomia]	Aborda predominantemente as possibilidades (ou as restrições) da escolha da mulher pelo abortamento ou pela maternidade.	"É o que fazem as mulheres que decidem abortar: pensam, refletem, discutem e decidem por aquilo que lhes dita a consciência, como melhor caminho naquele momento, como recorda um teólogo latino-americano."

Coação	Trata de situações em que a mulher é coagida a abortar ou a levar adiante uma gestação indesejada.	"Nesse sentido, a coação para as mulheres não pode vir do seu namorado, não pode vir da sua família e não pode vir do Estado."
Contracepção	Trata de métodos contraceptivos sob diferentes perspectivas.	"Existem tantos meios de contracepção. Claro que nenhum deles é 99%, mas as pessoas precisam aprender a assumir as suas responsabilidades."
Pílula do Dia Seguinte [subnó de Contracepção]	Defende ou critica o uso da pílula do dia seguinte como método contraceptivo.	"Discute-se a questão da pílula do dia seguinte. É abortiva? Não é abortiva? Pode ser utilizada? Não pode ser utilizada?"
Democracia	Trata do aborto ou do debate sobre a SUG como questão democrática.	"A primeira é a garantia e a qualidade da deliberação democrática. O tema do aborto, como outros temas, não é um tema trivial e, portanto, requer respeito, escuta, abertura ao diálogo [...]"
Desigualdade	Aborda a relação entre a pauta da SUG e questões de desigualdade.	"O aborto fala de nós, de vocês, mulheres comuns. Marcadores sociais de desigualdade, como juventude, classe e cor, agudizam a precariedade da vida dessas mulheres."
Desigualdade de Gênero [subnó Desigualdade]	Tem como foco as desigualdades entre homens e mulheres.	"Nós sabemos que uma mulher que foi violentada sexualmente é duplamente penalizada pela sociedade machista, que torna a mulher um objeto nessa relação desigual, usa dessa violência."
Desigualdade Financeira [subnó Desigualdade]	Tem como foco as desigualdades financeiras entre mulheres.	"Do contexto social dessas mulheres. Quem pode paga; quem não paga vai na sorte."
Desigualdade Racial [subnó de Desigualdade]	Tem como foco as desigualdades entre mulheres brancas e mulheres negras	"Por serem inseguros, os abortos arriscam a vida e a saúde das mulheres, notadamente as negras [...]"
Direito	Aborda questões jurídicas ligadas à questão do abortamento.	"a aplicação da lei penal é seletiva, afetando de maneira mais drástica as mulheres pobres, negras e socialmente excluídas."
Direitos Humanos [subnó de Direito]	Aborda especificamente o aborto como questão de direitos humanos (da mulher ou do feto).	"Como componente da pauta mais ampla de direitos sexuais reprodutivos, que também inclui o acesso à saúde reprodutiva, o aborto está inscrito no arcabouço geral dos direitos humanos [...]"

Maternidade	Foca no conceito de maternidade.	"E eu gostaria de iniciar falando da maternidade. Pode parecer estranho que, discutindo uma proposta que torna o aborto legal dentro de certos limites, eu proponha aqui trazer o tema da maternidade."
Maternidade não hegemônica [subnó de Maternidade]	Trata da maternidade como escolha.	"A maternidade deve ser uma decisão livre e desejada, não uma obrigação das mulheres."
Modos de Abortar	Descreve formas de abortamento.	"A mulher comum, a puta ou a adolescente abortam de maneira semelhante: usam comprimidos isolados ou combinados com chás, ervas ou garrafadas."
Morte de Mulheres [subnó Riscos do Aborto Inseguro]	Trata especificamente da morte de mulheres como consequência do aborto inseguro.	"Trezentas mortes maternas por ano em função de abortamento inseguro. Aproximadamente uma mulher por dia morre em função de aborto inseguro."
Origem da Vida	Trata do aborto sob a perspectiva da origem da vida.	"Se nós colocarmos filósofos, teólogos, cientistas de vários outros campos, não há consenso sobre o que é a vida."
Perfil da Mulher que Aborta	Elenca características da mulher que recorre ao abortamento no Brasil.	"Elas têm filhos; elas são jovens; elas têm entre 22 e 29 anos; elas têm religião, [...]"
Responsabilidade	Trata da responsabilidade de diferentes atores envolvidos na questão do aborto (mulher, marido, Estado etc.).	"Então, é de responsabilidade também do sexo masculino o controle da natalidade. Não só do sexo feminino."
Riscos do Aborto Inseguro [subnó de Riscos do aborto]	Aborda especificamente os riscos do aborto inseguro.	"Por serem inseguros, os abortos arriscam a vida e a saúde das mulheres, notadamente as negras e as mais pobres."
Morte de Mulheres [subnó Riscos do Aborto Inseguro]	Trata especificamente da morte de mulheres como consequência do aborto inseguro.	"Trezentas mortes maternas por ano em função de abortamento inseguro. Aproximadamente uma mulher por dia morre em função de aborto inseguro."

Fonte: Elaborado pelas autoras.

Nessa etapa, também foi relevante observar as temáticas mais predominantes no *corpus* pró-SUG. Reiteramos que o NVivo permitiu-nos fazer esse tipo de cruzamento ao categorizarmos cada participante por meio do recurso Classificações de Caso (reproduzido anteriormente

na Figura 6). No Gráfico 1, a seguir, o tamanho de cada nó corresponde à sua frequência nos dados. Assim, os nós mais à direita, e depois os mais superiores, são aqueles mais frequentes em número de itens codificados.

GRÁFICO 1 – Os nós mais frequentes no *corpus* pró-SUG e seus nós subordinados

Fonte: Elaborado pelas autoras.

De modo geral, esse levantamento de nós, previamente à descrição dos *frames* semânticos, permitiu-nos verificar que as temáticas abordadas ao longo das audiências públicas da SUG por vezes abrangeram ou extrapolaram a pauta inicial da Sugestão. Conforme o Gráfico 1, os nós mais predominantes no discurso pró-SUG foram Desigualdade, Autonomia, Direito, Riscos do Aborto (com frequência maior dos subnós Riscos do Aborto Inseguro e Morte de Mulheres), além de Democracia, Maternidade e Origem da Vida. Diante disso, destacamos que o tema da Autonomia ganhou significativo destaque nesse primeiro levantamento, embora não seja predominante nas discussões sobre abortamento em nosso contexto; como explica Elias (2018, p. 21), no Brasil, "[...] a

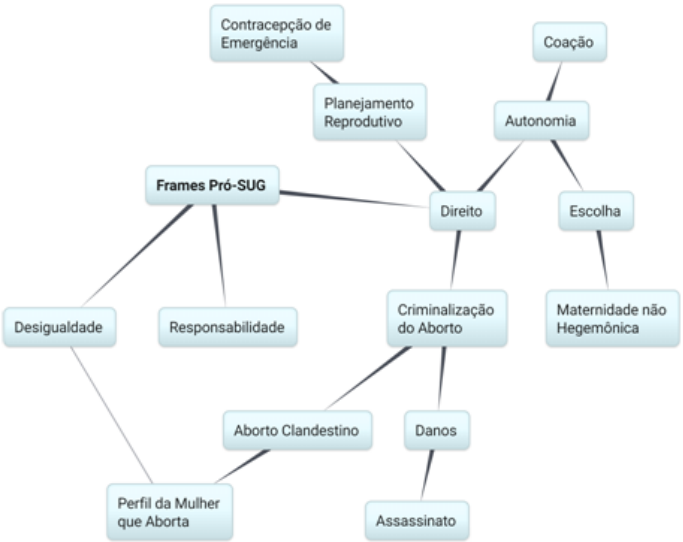
mobilização política e a reivindicação ao aborto como um direito de cidadania às mulheres têm menos destaque [...]" em comparação à abordagem do tema como questão de saúde pública.

Na próxima seção, considerando os nós sistematizados nesta etapa, identificamos os *frames* que emergiram do discurso pró-SUG nas audiências públicas investigadas.

4.2 Entre a questão de saúde pública e os direitos das mulheres: os *frames* pró-SUG

Conforme abordamos na seção 3.1, a motivação inicial da proposta da SUG nº 15 teve como foco principal o abortamento clandestino como questão de saúde pública. A partir desse dado e considerando a descrição de *frames* realizada por meio da organização do *corpus* em nós, identificamos os enquadramentos que refletem ou que ampliam esse foco. Embora alguns dos *frames* aqui discutidos sejam evocados também por participantes anti-SUG – e eventualmente reenquadrados, como buscamos mostrar nas seções posteriores –, os cenários de que tratamos neste artigo, concernem ao modo como participantes pró-SUG defenderam a proposta. A figura a seguir exibe a rede conceptual de *frames* que analisamos nesta seção:

FIGURA 12 – Rede conceptual de *frames* Pró-SUG



Fonte: Elaborado pelas autoras.

Os primeiros *frames* que destacamos estão relacionados principalmente à realidade sociocultural do abortamento como um todo e, mais especificamente, como uma problemática brasileira, inserida nos meandros de desigualdade – racial, econômica, de gênero – que atravessam o País e atravancam o avanço dos direitos das mulheres. Nesse âmbito, o *frame* Aborto_Clandestino, originado a partir do processamento do nó Modos de Abortar, trata das formas que as mulheres utilizam para interromper a gestação na clandestinidade, elencando os meios e instrumentos utilizados, bem como os locais onde ocorre o aborto. O quadro a seguir exibe parcialmente a descrição do referido *frame*, trazendo sentenças anotadas[13] – procedimento que seguimos ao reproduzir todos os demais *frames*. Nas concordâncias, entre colchetes, indicamos o participante que evocou o *frame*, seguindo a nomenclatura do *corpus* mencionada na seção 3.1.

QUADRO 2 – *Frame* Aborto_Clandestino

***Frame* Aborto_Clandestino**	
Definição:[14] Um agente usa um meio de interromper a própria gestação, ou a gestação de outrem.	**EFs e definições:** Agente — Mulher que aborta Instrumento — Instrumento realizado para abortar Local — Local onde ocorre o aborto clandestino Resultado — Resultado do ato de abortar
Evocadores: abortar, aborto, expulsar, operação, procedimento, provocar aborto	
[A3_PS_1_DD_Acad] A mulher comum , a puta ou a adolescente **abortam** de maneira semelhante: usam comprimidos isolados ou combinados com chás, ervas ou garrafadas. [A3_PS_1_DD_Acad] Quanto mais jovem for a mulher , o Cytotec é o método mais comum [...]. [A3_PS_2_JW_Pol] Ela entrou nessa clínica clandestina ; a **operação**, o **procedimento** deu errado; ela morreu; e eles deram fim no corpo dela carbonizando-o. [A5_PS_1_MA_Med] porque as mulheres favorecidas, de boa condição socioeconômica, têm acesso a medicamentos e recorrem a clínicas clandestinas [A5_PS_1_MA_Med] as mulheres pobres, as mulheres negras, as mulheres pardas , sem acesso à educação, [...] elas recorrem a soluções perigosas para **provocar o aborto**	

Fonte: Elaborado pelas autoras.

[13] Por questões de espaço, no caso de *frames* mais frequentes, elegemos os exemplos anotados mais ilustrativos para compor os quadros no corpo do texto.

[14] Neste estudo, não adotamos o termo "definição" de uma perspectiva lexicográfica; trata-se de uma breve contextualização de cada frame que complementa a respectiva lista de Elementos de Frame.

Chama a atenção o fato de que dos 78 participantes apenas três evocaram o *frame* Aborto_Clandestino, que é crucial ao entendimento da questão levantada pela SUG; especificamente, trata-se de uma acadêmica, cujas pesquisas sustentam os dados trazidos sobre os métodos de abortamento; de um parlamentar; e de uma médica – todos pró-SUG. Além disso, a acadêmica é a que mais evoca o respectivo *frame*.

O EF Agente do *frame* Aborto_Clandestino tem como instanciadores diferentes grupos de mulheres que abortam – aspecto que evidencia o entrelace com o *frame* Perfil_da_Mulher_que_Aborta, identificado a partir do nó homônimo, concernente às evidências, já sistematizadas por renomados estudos (DINIZ; MEDEIROS, 2010; DINIZ; MEDEIROS; MADEIRO, 2017), que situam o abortamento como ato comum a todas as classes sociais das mulheres, independentemente de idade, credo, nível de instrução, histórico reprodutivo, número de filhos, dentre outros aspectos. Interessante observar que a maioria dos participantes que evocam tal *frame* pertence aos domínios médico e acadêmico, valendo-se de dados estatísticos e outras categorias provenientes de pesquisas científicas para sustentar sua exposição.

QUADRO 3 – *Frame* Perfil_da_Mulher_que_Aborta

<table>
<tr><th colspan="2">Frame Perfil_da_Mulher_que_Aborta:</th></tr>
<tr><td>Definição: Este frame contém características da mulher que aborta.

Relações entre frames: subframe de Pessoa (FrameNet)</td><td>EFs e definições:
Idade — Idade da mulher que aborta
Nível de instrução — Nível de instrução da mulher que aborta
Religião — Religião da mulher que aborta
Etnia — Etnia da mulher que aborta
Classe social — Classe social da mulher que aborta
História reprodutiva — Histórico reprodutivo da mulher que aborta
Estado civil — Estado civil da mulher que aborta</td></tr>
<tr><td colspan="2">Evocadores: ela, elas, mulheres, mulheres que abortam</td></tr>
<tr><td colspan="2">[A1_PS_1_HS_Med] Das que informaram ter realizado ao menos um procedimento ao longo da vida, 15% se declararam católicas; 13% evangélicas, e 16% de outras religiões.
[A2_PS_1_SV_Med] Quem são essas mulheres? Elas tinham 27 anos, 50% pardas, quase 34% pretas, 57% completaram o ensino médio e 86% tinham renda familiar inferior a R$1.500,00. Situação típica das capitais nordestinas.
[A2_PS_1_TG_Med] São especialmente as mulheres em condições menos favorecidas aquelas que se submetem aos riscos da prática do aborto realizado em condições precárias.
[A3_PS_1_DD_Acad] Elas têm filhos, elas são jovens; elas têm entre 22 e 29 anos; elas têm religião, como aquelas que hoje aqui estão representadas para falar contra o aborto; elas têm um companheiro.</td></tr>
</table>

Fonte: Elaborado pelas autoras.

Além disso, os instanciadores dos EFs Nível de Instrução, Etnia e Classe Social mencionam um perfil menos favorecido de mulheres que abortam – trata-se das mulheres negras, pobres, com baixo nível de instrução formal e de baixa renda, que são as maiores vítimas do aborto clandestino. É por meio de tais Elementos de Frame que identificamos o entrelaçamento entre os instanciadores dos EFs do *frame* Perfil_da_Mulher_que_Aborta e o *frame* Desigualdade, que partiu do nó homônimo e de seus respectivos subnós, cuja anotação é exibida parcialmente no quadro a seguir. É possível perceber que tal *frame* é evocado mais uniformemente por todas as categorias socioprofissionais de participantes pró-SUG. Assim, médicos, políticos, ativistas, religiosos, acadêmicos e advogados desse grupo abordam as desigualdades entre mulheres negras/pobres e mulheres brancas/ricas que se concretizam em diferentes situações (EF Situação): na questão do abortamento em geral, no acesso a um aborto seguro (mesmo ilegal, clandestino), nos casos de morte materna e de penalização por crime de aborto, entre outros.

QUADRO 4 – *Frame* Desigualdade

***Frame* Desigualdade:**		
Definição: Este *frame* designa uma comparação desigual entre dois agentes, de modo que um está em desvantagem em relação ao outro no que se refere a algum atributo	**EFs e definições:**	
	Agente em desvantagem	Agente em posição de desvantagem
	Agente em vantagem	Entidade em posição de vantagem
	Situação	Contexto no qual se estabelece a desigualdade
	Meio	Meio pelo qual se estabelece a desigualdade
Evocadores: desigualdade**,** apartheid, desiguais, dominação, dominar		
Excertos do *corpus*: [A1_PS_1_AC_Med] não é uma escolha da civilização que mantém esse **apartheid** de direitos entre mulheres e homens, entre mulheres ricas e não ricas quando se trata da questão do aborto. [A1_PS_1_RT_Rel] Os dados têm mostrado que são as mulheres negras e pobres as que têm sofrido as consequências da criminalização do aborto, porque as mulheres que têm dinheiro vão para fora do País fazer a interrupção em um país onde é legalizado ou mesmo em clínicas onde elas podem pagar o preço estipulado funcionam dentro de padrões de higiene adequados, e elas abortam seguramente, enquanto [A2_PS_1_LM_Acad] Todos deveriam pensar nas mulheres pobres, negras, em Salvador, no Norte e no Nordeste, que têm que enfrentar condições de vida **desiguais**, menos acesso às políticas públicas, às condições de trabalho, à oportunidade educacional		

[A5_PS_1_MA_Med] Existe uma grande **desigualdade** regional e uma grande **desigualdade** econômica, porque praticamente não vemos mortes por aborto nos hospitais privados nem nas regiões mais ricas do mundo, e essas mortes por aborto nos países pobres também têm uma **desigualdade** dentro dos próprios países.
[A5_PS_2_GC_Ativ] porque esta Casa, infelizmente, ainda é uma casa marcada pela ordem patriarcal... tem praticamente só homens. Ficamos caladas aqui e ainda levamos lição de moral todo tempo.

Fonte: Elaborado pelas autoras.

Como mostram os exemplos, os participantes pró-SUG não se atêm a abordar as desigualdades de raça e classe entre mulheres que abortam, mas também instanciam esse *frame* para tratar das relações assimétricas existentes entre homens e mulheres quando se trata da responsabilidade por evitar uma gravidez; bem como entre países desenvolvidos e subdesenvolvidos – estes últimos, em virtude de desigualdades sociais, apresentam maior número de mortes decorrentes de abortamentos clandestinos (GUTTMACHER INSTITUTE, 2017). Além disso, com indica o último exemplo do Quadro 4, essa desigualdade de gênero também é abordada no que se refere a contextos específicos como o Senado – afinal, estava-se discutindo uma pauta que evidencia a necessidade de luta pelos direitos das mulheres em um ambiente elitista, historicamente dominado por homens brancos heterossexuais, sendo alguns deles autores de pautas retrógradas quanto aos direitos das mulheres (SANTOS, 2020). Tal aspecto aponta que as constrições desse contexto institucional podem limitar consideravelmente o alcance da voz dos defensores das mulheres e de seus direitos.

Algumas expressões que instanciam os EF Entidade em Desvantagem (*frame* Desigualdade) e evocam o *frame* Perfil_da_Mulher_que_Aborta também instanciam o EF Paciente do *frame* Danos, cuja descrição teve como ponto de partida o nó Riscos do Aborto, conforme exibimos no quadro a seguir. Nesse enquadramento, o aborto inseguro instancia o EF Causa – trata-se do causador dos danos à mulher que aborta na clandestinidade.

QUADRO 5 – *Frame* Danos

***Frame* Danos:**		
Definição: Um agente ou uma causa afetam um paciente de tal maneira que o paciente fica em um estado anômalo, geralmente não desejado.	**EFs e definições:** Causa Paciente	Causa do dano ao paciente. Parte afetada pelo agente, sofrendo danos.
Evocadores: morrer, mutilação, hemorragia, infecções, perfuração uterina, hemorragia, infecção, choque séptico, perfuração de vísceras, traumatismos genitais, dor pélvica, infertilidade		
Excertos do *corpus*: [A2_PS_1_SV_Med] As taxas de complicação por aborto, ou seja, os motivos da complicação nessas mulheres são **hemorragia** e **infecções**. IN [aborto inseguro] [A3_PS_2_JW_Pol] E o aborto é quarta causa de **mortalidade materna** hoje no Brasil e a primeira entre mulheres pobres e negras. Ou seja, esse é um problema de saúde pública colocado aqui. IN [aborto inseguro] [A5_PS_1_MA_Med] São **complicações** de abortos mal feitos, de abortos inseguros, de abortos clandestinos, com métodos obsoletos, que não se utilizam mais; métodos perigosíssimos que deveriam ser proibidos [...]. Incluem, além de **perfuração uterina**, **hemorragia**, **infecção**, **choque séptico**, **perfuração de vísceras**, **traumatismos genitais**, e as mulheres podem sobreviver com **sequelas** [...].		

Fonte: Elaborado pelas autoras.

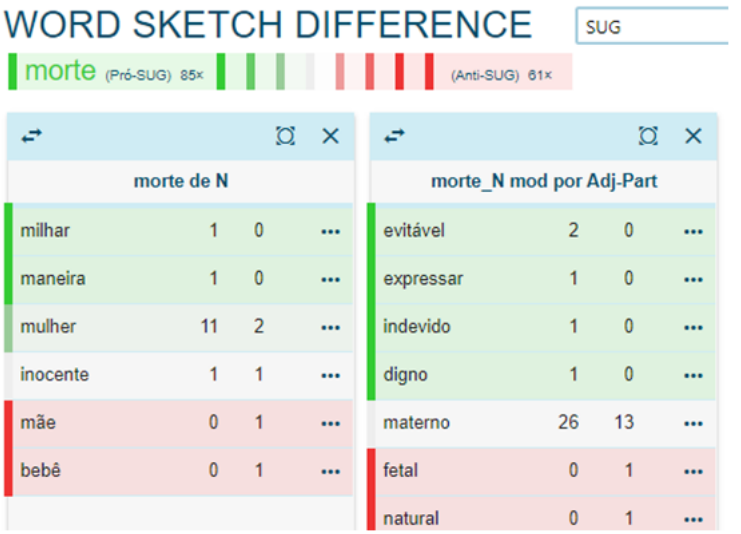
Os excertos que exibimos no Quadro 5 têm como instanciador do EF Causa, especificamente, o aborto inseguro ou clandestino. Conforme mencionamos na seção 3.3, por vezes um instanciador pode aparecer na posição de Instanciação Nula – ou seja, não é explicitado no texto, geralmente porque foi mencionado em uma parte anterior. Nesses casos, quando relevante, assinalamos com uma expressão entre colchetes o termo instanciado, conforme conseguimos recuperá-lo ao expandirmos as concordâncias do Sketch Engine.

É por meio do *frame* Danos que o abortamento como questão de saúde pública é evidenciado, considerando-se as sequelas do aborto clandestino e o índice de mortalidade materna ocasionado pela recorrência de procedimentos realizados nessas condições no País. Observamos que a maior parte dos excertos anotados é de autoria de participantes médicos(as), de modo que o léxico especializado, atinente à área da Saúde, também é preponderante nesse segmento dos dados.

Os pró-SUG são os únicos que abordam a morte de mulheres em virtude do abortamento clandestino, como é possível verificar no Sketch Difference de “mulher”, a seguir. Mais especificamente, esses

participantes se utilizam onze vezes da combinação "morte de mulheres"[15] (vide terceiro item da coluna à esquerda). São também os únicos que tratam das "mortes evitáveis"[16] de mulheres que recorrem ao abortamento clandestino (vide primeiro item da coluna à direita).

FIGURA 13 – Sketch Difference para a palavra *morte*



Fonte: Elaborada pelas autoras.

O *frame* Danos, principalmente por ter como evocadores expressões como "mortes maternas", entrelaça-se ao *frame* Assassinato (originado do nó homônimo) em alguns casos, os quais exploramos a seguir. Trata-se de excertos em que a ilegalidade do abortamento, em virtude dos danos causados, torna-se um instrumento de execução de mulheres. Os instanciadores do EF Assassino, responsável pelos casos de mortalidade materna, são lexicalizados duas vezes: na primeira, temos "vocês do pró-morte", em que a participante se refere a todo o movimento antiescolha como responsável pelas mazelas do aborto clandestino. Na segunda ocorrência, "o Estado" é o grande executor da "pena de

[15] Exemplo de concordância: "[...] o Estado brasileiro é responsável pelas mortes das mulheres em situação de risco, em abortamento inseguro".
[16] Exemplo de concordância: "[...] dentro da mortalidade materna, temos de pensar nas mortes evitáveis."

morte" (evocador do *frame*) contra "mulheres pobres" – a partir desse instanciador, temos outro entrelaçamento com o *frame* Desigualdade.

QUADRO 6 – *Frame* Assassinato

***Frame* Assassinato:**	
Definição: Um assassino ou causa ocasiona a morte da vítima.	**EFs e definições:** Assassino — Agente responsável pela morte da vítima Instrumento — Usado para matar a vítima Vítima — Morre como resultado do assassinato Meio — Método ou ação que resulta na morte da vítima
Evocadores: eliminar, matar, interromper a vida, assassinado, roubar a vida, sacrificar	
Excertos do *corpus*: [A1_PS_1_IT_Adv] os corpos que escolhemos deixar morrer, as mulheres que **escolhemos deixar morrer** em decorrência de procedimentos malsucedidos de abortamento. [A4_PS_1_MN_Rel] a ilegalidade do aborto como instrumento de morte. É essa **morte** mulheres brasileiras que eu não quero que continue a acontecer na escala em que acontece, entre outras razões, mas muito fortemente, pela ilegalidade do aborto em nosso País. [A5_PS_2_PV_Ativ] Nós queremos que vocês dos pró-morte, com seus dogmas religiosos e violadores do Estado laico, que promovem a tortura diária das mulheres, que **promovem a morte** das mulheres, que promovem mais e mais abortos clandestinos e inseguros, sejam responsabilizados por isso. [A3_PS_1_MT_Acad] Fácil criminalizá-las, fácil **matá**-las, fácil para o Estado não se responsabilizar por essa **pena de morte** contra mulheres pobres.	

Fonte: Elaborado pelas autoras.

No *corpus*, também identificamos o *frame* Responsabilidade, cuja descrição partiu do nó homônimo, no qual o EF Parte Responsável tem a incumbência de cumprir determinado dever, ou é responsável por determinado acontecimento. A partir dessas conceptualizações, os participantes pró-SUG salientam a responsabilidade não só do Estado por "mais e mais abortos clandestinos e inseguros" e dos movimentos antiescolha pela morte de mulheres, mas também tratam da responsabilidade dos homens pela contracepção, pelo "controle da natalidade", de modo que tal compromisso não deve ser atribuído somente às mulheres.

QUADRO 7 – *Frame* Responsabilidade

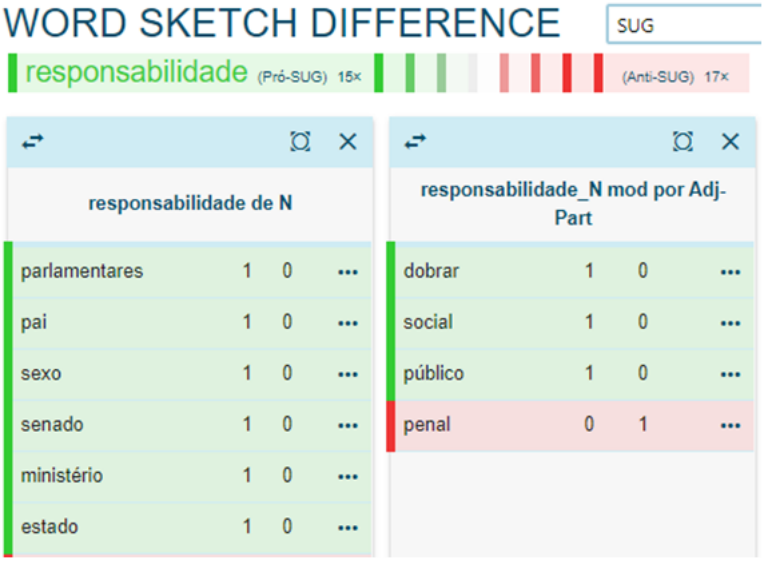
***Frame* Responsabilidade**	
Definição: uma parte responsável é requerida a cumprir um dever. Origem: *frame* Being_ Obligated (FrameNet)	**EFs e definições:** Parte responsável — Pessoa que deve cumprir um dever Responsabilidade — Dever a ser cumprido, ou evento/entidade pela qual a parte é responsável
Evocadores: assumir a responsabilidade, responsáveis, assumir as consequências, obrigação, responsabilização	
[A5_PS_1_LL_Adv] Esse é um dado importante quando nos damos conta da pouca **responsabilização** dos homens na vida reprodutiva. Cai somente nos ombros das mulheres, nos úteros das mulheres essa **responsabilidade** [A5_PS_2_PV_Ativ] Nós queremos que vocês dos pró-morte, com seus dogmas religiosos e violadores do Estado laico, que promovem a tortura diária das mulheres, que promovem a morte das mulheres, que promovem mais e mais abortos clandestinos e inseguros, sejam **responsabilizados** por isso. [A3_PS_1_MT_Acad] Fácil criminalizá-las, fácil matá-las, fácil para o Estado não se **responsabilizar** por essa pena de morte contra mulheres pobres.	

Fonte: Elaborado pelas autoras.

Buscando pelo termo "responsabilidade" no Sketch Difference, verificamos que somente os pró-SUG combinam esse termo com os itens "parlamentares",[17] "pai", "senado",[18] "Ministério"[19] e "Estado", conforme exibimos na figura a seguir (coluna à esquerda). São também os únicos que falam em responsabilidade social e pública (vide coluna à direita). Já os anti-SUG tratam apenas de responsabilidade penal (em uma ocorrência, conforme coluna à direita).

[17] Concordância: "Então, queria chamar a atenção e a responsabilidade dos nossos Parlamentares, das Casas Legislativas e das Lideranças políticas [...]".
[18] Concordância: "[...] coloca também a responsabilidade do Senado Federal em fazer este debate [...]".
[19] Concordância: "[...] atenção humanizada ao abortamento. É da responsabilidade do Ministério da Saúde fazer isso."

FIGURA 14 – Sketch Difference para a palavra responsabilidade



Fonte: Elaborada pelas autoras.

É a partir de tais enquadramentos – Aborto_Clandestino, Perfil_da_Mulher_que_Aborta, Desigualdade, Danos, Assassinato, Responsabilidade – que os participantes pró-SUG conceptualizam a criminalização do aborto como mecanismo ineficaz para reduzir o número de abortamentos clandestinos, conforme disposto no Quadro 8 – o *frame* Criminalização teve como origem parte do nó Direito. A maioria dos excertos é de autoria de advogados pró-SUG e enfatiza, por meio do Elemento de Frame Avaliação, o "descompasso" entre a Lei Penal brasileira e a "nossa realidade social"; o "anacronismo" do Código Penal, que "não resolve" o problema do número de abortos clandestinos e perigosos; a inconstitucionalidade e a crueldade da lei ao criminalizar mulheres, impedindo "o acesso ao aborto seguro"; e a seletividade da Lei Penal, que criminaliza mulheres "pobres, negras e socialmente excluídas".

Salientamos ainda que somente os pró-SUG lexicalizam a *criminalização*, (o problema da) *ilegalidade*[20] e a necessidade de *descriminalização*[21] do abortamento, conforme mostra a Figura 15 (exibida logo após o Quadro 8).

QUADRO 8 – *Frame* Criminalização_do_Aborto

***Frame* Criminalização_do_Aborto**		
Definição: Ato de criminalizar um agente ou uma ação.	**EFs e definições:** Protagonista Ação Avaliação Base	 Parte criminalizada Ato criminalizado Avaliação do processo de criminalização referido Base jurídica para a criminalização
Evocadores: criminalização, criminalizar, tratamento criminal, aplicação da lei penal		
Excertos do *corpus*: [A1_PS_1_IT_Adv] **tratamento criminal** que se dá à questão do aborto no Brasil, reflete certo anacronismo da nossa legislação, um descompasso existente entre a legislação penal que criminaliza a mulher que pratica o aborto e a nossa realidade social [A1_PS_1_IT_Adv] **tratamento criminal** que se dá à questão do aborto no Brasil, reflete certo anacronismo da nossa legislação, um descompasso existente entre a legislação penal que criminaliza a mulher que pratica o aborto e a nossa realidade social [A5_PS_1_MA_Med] A segunda razão é que o principal fator para impedir o acesso ao aborto seguro é a **criminalização**. [A1_PS_1_IT_Adv] nós estamos afirmando que a perspectiva é de o Brasil querer **criminalizar** essas mulheres [A5_PS_1_EA_Adv] Portanto, e aí o argumento tem uma reviravolta, **criminalizar** é inconstitucional. [A3_PS_1_SC_Ativ] **a aplicação da lei penal** é seletiva, afetando de maneira mais drástica as mulheres pobres, negras e socialmente excluídas.		

Fonte: Elaborado pelas autoras.

[20] Exemplo de concordância: "É essa morte das mulheres brasileiras que eu não quero que continue a acontecer na escala em que acontece, entre outras razões, mas muito fortemente, pela ilegalidade do aborto em nosso País."

[21] Exemplo de concordância: "[...] reúne todas as evidências que levaram a Federação Internacional dos Ginecologistas e Obstetras a defender a descriminalização do aborto como uma medida de saúde pública [...]."

FIGURA 15 – Sketch Difference para a combinatória "*de(o) aborto*"

Fonte: Elaborada pelas autoras.

Complementando esse aspecto, o *frame* Direito (segundo mais evocado do *corpus*), que partiu do nó homônimo, é agenciado pelos grupos pró-SUG para defender as reinvindicações das mulheres, tais como direitos "sociais e reprodutivos", incluindo os direitos fundamentais citados na Constituição de 1988 (BRASIL, 1988): direito à vida, à saúde, à autonomia, entre outros. Em suma, a mulher instancia, predominantemente em falas de advogados pró-SUG, o Elemento de Frame Protagonista, a quem devem ser garantidos direitos humanos.

QUADRO 9 – *Frame* Direito

***Frame* Direito:**	
Definição: Um protagonista é revestido do direito de exercer algum poder concedido pela Lei.	**EFs e definições:** Protagonista — Agente revestido do direito Direito — Direito de ter ou fazer algo de acordo com a Lei Base — Base jurídica para o direito concedido Dimensão — Extensão ou limite do direito concedido
Evocadores: direito, proteção jurídica, garantir direito, exercício, exercer	
Excertos do *corpus*: [A5_PS_1_EA_Adv] E o Tribunal faz uma afirmação enfática, que é muito relevante para esta Comissão: a inviolabilidade do **direito** à vida, que está escrito no art. 5º da nossa Constituição Federal, se refere exclusivamente a um ser já personalizado. [A5_PS_1_EA_Adv] E se estou falando que o aborto é um **direito**, um direito com base na dignidade humana, com base na autonomia, com base na liberdade, significa que alguém tem a obrigação de **garantir esse direito**. [A5_PS_1_LL_Adv] mesmo que haja um conflito de **direitos** entre os **direitos** da mulher e os direitos do embrião, esse conflito tem que ser decidido levando em consideração que a mulher já é uma vida plena, que a mulher já é o sujeito de direito e que o embrião no máximo tem uma expectativa de **direitos**. [A3_PS_1_SC_Ativ] A perspectiva feminista, que é a minha, que reivindica o **direito** de decisão reprodutiva às mulheres, repudia, de maneira forte, as leis e políticas de aborto compulsório	

Fonte: Elaborado pelas autoras.

A Sketch Difference da palavra *direito* mostra-nos que direito de *escolha*, de *opção*; bem como os direitos *iguais*, *sociais* e *sexuais* são unicamente lexicalizados por grupos pró-SUG, que situam, assim, o debate da Sugestão para além da questão de saúde pública – corroborando o que já havíamos verificado ao elencarmos os nós do *corpus* na etapa correspondente à seção 4.2.

FIGURA 16 – Sketch Difference para a palavra *direito*

Fonte: Elaborada pelas autoras.

Nesse sentido, o instanciador "autonomia", do EF Direito, é também evocador do *frame* homônimo – decorrente do nó também denominado Autonomia –, que exibimos a seguir, evocado predominantemente por médicos e advogados pró-SUG.

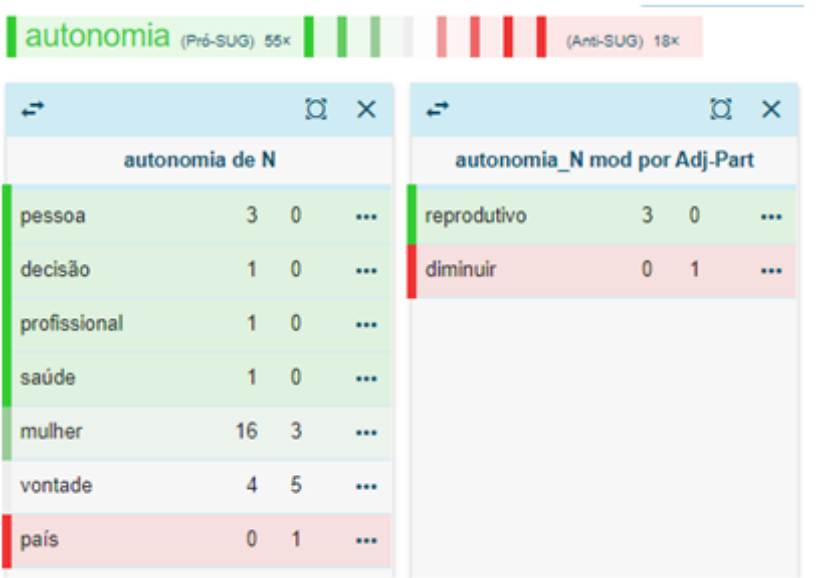
QUADRO 10 – *Frame* Autonomia

***Frame* Autonomia**		
Definição: estado ou condição de um ser autônomo para se autogovernar.	**EFs e definições:** Protagonista Avaliação Extensão	Pessoa que tem direito a autonomia Avaliação da autonomia Extensão da autonomia
Evocadores: autonomia, autônomo, autodeterminação		
[A1_PS_1_HS_Med] deixamos muito bem claro e frisamos que não se decidiu serem os Conselhos de Medicina favoráveis ao aborto, mas, sim, discutimos a **autonomia** da mulher e do médico, o que é nossa obrigação. [A2_PS_1_JB_Ativ] o interesse das mulheres que tomam decisões **autônomas**, concentradas no seu cotidiano, na sua vida e na sua livre consciência. [A5_PS_1_EA_Adv] que permitam que essa escolha seja feita com segurança e com preservação da **autonomia** da mulher [A5_PS_1_LL_Adv] nossa proposta é uma política de respeitar a **autonomia** reprodutiva das mulheres, a **autodeterminação** das mulheres		

Fonte: Elaborado pelas autoras.

Destacamos que os pró-SUG são os que mais se utilizam da colocação "autonomia da mulher" (16 vezes); e os únicos a mencionar o conceito de autonomia reprodutiva, conforme é possível verificar na Figura 17.

FIGURA 17 – Sketch Difference para a palavra *autonomia*



Fonte: Elaborada pelas autoras.

Mais especificamente, identificamos também o *frame* Escolha, originado do subnó homônimo e subordinado ao *frame* Autonomia, que tem como foco as alternativas disponíveis ao sujeito que escolhe, bem como a escolha realizada e suas circunstâncias. Tal enquadramento é exibido no quadro 11.

QUADRO 11 – *Frame* Escolha

***Frame* Escolha**		
Definição: um cognoscente faz uma escolha dentre uma série de possibilidades.	**EFs e definições:** Agente Escolha Alternativa Circunstância	 Pessoa que faz a escolha Escolha realizada Alternativas disponíveis para escolha Circunstância em que a escolha é realizada
Evocadores: escolha, escolher, decidir, optar, desistir, pseudoescolha, não querer		
[A5_PS_2_GC_Ativ] legitimamente, exercemos a nossa autonomia de **decisão** sobre quando e se queremos parir [A5_PS_2_GC_Ativ] Nós somos solidárias com as mulheres na hora em que elas **decidem** pela maternidade e não têm o apoio do Estado. [A1_PS_1_RT_Rel] Então, se a mulher **decidir** por uma interrupção da gravidez, é ela e sua consciência. [A4_PS_2_PV_Ativ] Nós queremos que as mulheres possam **escolher**. Apoiamos a Regiane ou quem for que queira **escolher** manter sua maternidade, seguir com sua maternidade, com qualidade de vida. [A5_PS_1_EA_Adv] "A gravidez não deve ser forçada, deve ser **escolha**"		

Fonte: Elaborado pelas autoras.

Observamos que somente os pró-SUG se valem das colocações "liberdade de escolha",[22] "exercício de escolha",[23] "questão de escolha"[24] e "direito de escolha",[25] conforme exibe a Sketch Difference na figura 18.

[22] Concordância: "[...] o exercício da sua liberdade de escolha."
[23] Concordância: "[...] que você tenha essa possibilidade do exercício da escolha de ter ou não ter filhos."
[24] Concordância: "Então, a questão da escolha difícil, pois ninguém é a favor do aborto."
[25] "[...] a necessidade de se garantir autonomia, direito de escolha às mulheres [...]."

FIGURA 18 – Sketch Difference para a palavra *escolha*

Fonte: Elaborada pelas autoras.

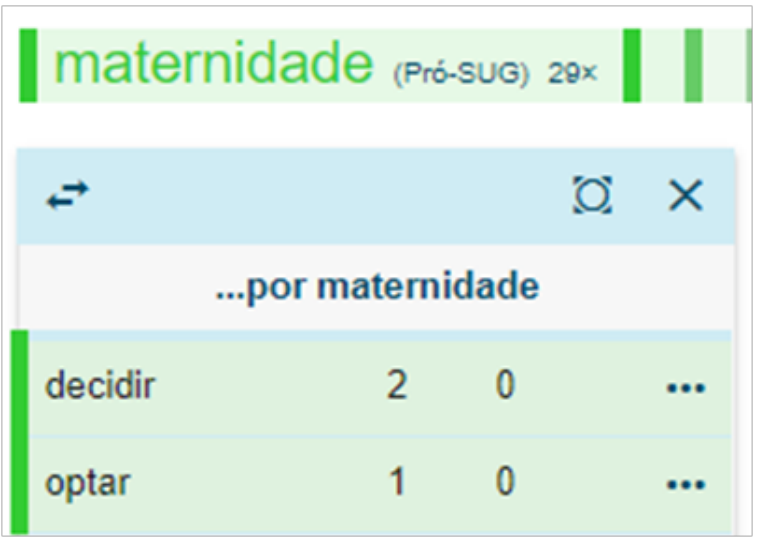
A maternidade como escolha, que instancia o EF Alternativa no Quadro 11, é uma característica do modelo de maternidade não hegemônica, que pressupõe igualdade entre homens e mulheres nas relações de trabalho, bem como partilha da responsabilidade parental (BELTRAME, 2016). O termo também consiste em um nó e em um dos *frames* identificados no *corpus*, conforme descrevemos no Quadro 12. Ao encontro disso, observamos que as combinatórias entre os verbos *decidir*/*optar* e *maternidade* só ocorrem nesse *subcorpus* em específico, como mostra a Figura 19.

QUADRO 12 – *Frame* Maternidade_não_Hegemônica

***Frame* Maternidade_não_Hegemônica**		
Definição: condição vista como uma opção à mulher, que pressupõe igualdade entre homens e mulheres nas relações de trabalho, bem como partilha da responsabilidade parental.	**EFs e definições:** Mulher Características	Mulher que tem a opção de ser mãe Características da maternidade não hegemônica
Evocadores: maternidade, gravidez		
[A2_PS_2_CB_Ativ] A **maternidade** deve ser uma decisão livre e desejada, não uma obrigação das mulheres. [A4_PS_1_MN_Rel] No entanto, para que a **maternidade** seja considerada em sua grandeza, é absolutamente necessário que compreendamos como resultado de uma decisão, de uma escolha, como uma opção entre tantas outras de realização das mulheres. [A4_PS_1_MN_Rel] Só compreendendo, portanto, a **maternidade** como resultado de opção e de escolha é possível entender o alcance ético de uma proposta que permite às mulheres acederem a um aborto quando assim considerarem necessário.		

Fonte: Elaborado pelas autoras.

FIGURA 19 – Sketch Difference parcial para a palavra *maternidade*



Fonte: Elaborada pelas autoras.

Tanto os *frames* Autonomia e Escolha quanto o *frame* Maternidade_não_Hegemônica opõem-se ao *frame* Coação, originado principalmente do nó Manipulação, que é evocado pelos pró-SUG justamente para condenar quaisquer ações que coajam a mulher – levando-a a seguir com uma gestação forçada; ou mesmo obrigando-a a abortar. Nesse caso, os principais instanciadores do EF Coagente (responsável pela coação) são leis e medidas estatais, bem como familiares.

QUADRO 13 – *Frame* Coação

***Frame* Coação:**		
Definição: Ato de coagir um agente, impondo que ela aja contra sua vontade	**EFs e definições:** Coagente Coagido Resultado	 Responsável pela coação Ser coagido Resultado da coação
Evocadores: coagir, coação, forçado, obrigar, pressionar		
[A3_PS_1_SC_Ativ] A perspectiva feminista, que é a minha, que reivindica o direito de decisão reprodutiva às mulheres, repudia, de maneira forte, as leis e políticas de aborto compulsório, assim como também medidas estatais que **coagem** as mulheres à procriação compulsória [A5_PS_1_LL_Adv] Nenhuma mulher deve ser **obrigada** a fazer um aborto, nenhuma mulher pode ser **coagida** a fazer um aborto, como nenhuma mulher deve ser **obrigada** e **coagida** a não interromper a gestação [A2_PS_1_LM_Acad] Nesse sentido, a **coação** para as mulheres não pode vir do seu namorado, não pode vir da sua família e não pode vir do Estado.		

Fonte: Elaborado pelas autoras.

Como abordam os participantes defensores da SUG, para que a mulher exerça plenamente seus direitos de cidadã, é preciso que haja políticas de Planejamento_Reprodutivo (Quadro 14), que incluem acesso à Contracepção_de_Emergência (Quadro 15), enquadramento originado do nó Pílula do Dia Seguinte. Esse método contraceptivo é considerado fundamental para reduzir a gravidez indesejada e evitar o abortamento inseguro – incluindo casos de estupro, nos quais as mulheres brasileiras têm direito à atenção humanizada, que abrange a prescrição de contracepção de emergência (BRASIL, 2011).

QUADRO 14 – *Frame* Planejamento_Reprodutivo

***Frame* Planejamento_Reprodutivo**		
Definição: política pública de saúde que desenvolve ações que possam propiciar o planejamento reprodutivo da população	**EFs e definições:** Cidadão Ação Circunstância	 Pessoa que deve ter acesso ao planejamento reprodutivo Ações realizadas por políticas de planejamento reprodutivo Circunstâncias em que ocorrem as ações de planejamento reprodutivo
Evocadores: planejamento reprodutivo, planejamento familiar		
[A1_PS_1_AC_Med] todos os métodos contraceptivos devem estar disponíveis para todas as mulheres, em todas as idades [A1_PS_1_RT_Rel] Então, o **planejamento reprodutivo** deve estar à disposição da população. Todos os métodos que a ciência conseguiu até hoje elaborar devem estar disponíveis à população.[A5_PS_1_LL_Adv] nós temos direitos ao **planejamento familiar** sem coação e com o dever do Estado de fornecer os métodos e os meios necessários para o exercício desse direito.		

Fonte: Elaborado pelas autoras.

QUADRO 15 – *Frame* Contracepção_de_Emergência

***Frame* Contracepção_de_Emergência:**		
Definição: trata da pílula do dia seguinte e de perspectivas sobre seus efeitos	**EFs e definições:** Contraceptivo Usuária Características	 Pílula do dia seguinte Usuária da pílula do dia seguinte Atributos da pílula do dia seguinte
Evocadores: pílula do dia seguinte, anticoncepção de emergência		
Excertos do *corpus*: [A1_PS_1_MS_Acad] e, muitas vezes, nem **a pílula do dia seguinte**, que poderia evitar um conjunto de danos e decisões conflituosas às mulheres, como a situação de aborto, mesmo essa medida mínima a gente nem sempre consegue. [A1_PS_1_MV_Med] Nós temos que ampliar a atenção integral às mulheres em situação de violência sexual, [...] através da **pílula de emergência**, que é um grande dispositivo para reduzir a gravidez indesejada e o aborto inseguro.		

Fonte: Elaborado pelas autoras.

Esta seção objetivou analisar os *frames* instanciados no *corpus* da SUG que vão ao encontro da intenção legislativa da Sugestão, cuja proposta parte principalmente da conceptualização do abortamento como questão de saúde pública. De modo geral, os resultados evidenciam

que, embora a iniciativa da Sugestão tenha partido da conceptualização do abortamento como questão de saúde pública, os participantes que a defendem não se atêm a esse aspecto ao agenciarem *frames* segundo seus propósitos comunicativos (TOMASELLO, 2003). Além disso, há *frames* que são evocados exclusivamente pelos pró-SUG, quais sejam: Aborto_Clandestino, Perfil_da_Mulher_que_Aborta, Criminalização_do_Aborto, Planejamento_Reprodutivo e Maternidade_não_Hegemônica. Vale ainda pontuar que, ao evocarem o *frame* Desigualdade, os participantes pró-SUG salientam não apenas as disparidades de raça, cor e escolaridade que determinam o tipo de serviço de aborto acessado pela mulher, mas também enfatizam questões como a desigualdade de gênero, a qual se reflete inclusive na configuração de espaços supostamente democráticos como o Senado.

Nesse sentido, ao tratarem dos diferentes tipos de Desigualdade que permeiam a questão do Aborto_Clandestino; das inúmeras características que incluem, no Perfil_da_Mulher_que_Aborta, a mulher casada, religiosa e com filhos (DINIZ; MEDEIROS; MADEIRO, 2017); dos Danos causados por esse procedimento inseguro; do Assassinato de mulheres que abortam (geralmente negras e pobres); da Responsabilidade do Estado para com as mulheres e dos homens para com a corresponsabilidade diante de uma gravidez indesejada; dos efeitos perversos da Criminalização_do_Aborto, que ocasionam o Assassinato de mulheres abortando na clandestinidade; da gama de Direitos nem sempre garantidos às mulheres, dada essa criminalização – incluindo o direito ao Planejamento_Reprodutivo, à Autonomia e o consequente direito de Escolha por abortar ou levar a cabo uma gestação –; e da luta por uma noção de Maternidade_não_Hegemônica, que respeite a mulher como sujeito pleno de direitos e que iniba qualquer tipo de Coação, os participantes pró-SUG fazem mais que um clamor à visibilização da cruel realidade do abortamento clandestino: realizam uma exposição fundamentada e uma reivindicação concreta para que mulheres não sejam mais punidas pela criminalização seletiva do abortamento.

A seção a seguir traz algumas considerações acerca do percurso de análise aqui realizado.

5 Considerações finais

Este artigo teve como objetivo principal discutir alguns desdobramentos analíticos de um estudo que investigou *frames* semânticos em audiências públicas da Sugestão Legislativa n.º 15/2014. Por meio de um recorte voltado à identificação de *frames* no discurso dos defensores da proposta da SUG nº 15, buscamos elucidar as possibilidades analíticas viabilizadas pela integração de uma ferramenta de análise qualitativa de dados – o NVivo – ao recurso Sketch Engine. A opção por tal articulação foi realizada tendo em vista a necessidade de segmentação do *corpus* em unidades temáticas para posterior processamento dos dados no concordanciador e no *Sketch Difference*.

Ponderamos que esta proposta possa ser considerada uma abordagem *middle-out* de exploração dos dados (CHISHMAN *et al.*, 2018), ou seja, busca-se o "caminho do meio" entre análises *bottom-up* (que têm como único ponto de partida léxico) e as *top-down* (que se valem de aspectos macrocontextuais como base inicial para a análise). Assim, a partir de uma segmentação em *subcorpora* de temas possibilitada pelo recurso NVivo (direcionamento metodológico *top-down*), realizamos uma descrição preliminar dos *frames* semânticos, a qual foi revista e consolidada por meio do processamento de listas de palavras-chave na ferramenta Sketch Engine (direcionamento *bottom-up*), da análise de concordâncias e da anotação semântica de excertos que evocavam os respectivos *frames*. Nesse percurso, utilizamo-nos também do recurso Sketch Difference, que permite a comparação entre usos linguísticos nos *subcorpora* selecionados, para observar a ocorrência de combinatórias lexicais peculiares ao *corpus* pró-SUG.

Na seção dedicada à análise dos dados, descrevemos *frames* instanciados no *corpus* que vão ao encontro da intenção legislativa da Sugestão, verificando que os participantes pró-SUG não se ativeram ao tema do abortamento como questão de saúde pública – via *frames* como Aborto_Clandestino, Desigualdade, Perfil_da_Mulher_que_Aborta, Assassinato e Criminalização_do_Aborto. O abortamento também foi conceptualizado como questão de autonomia da mulher, por meio da evocação do *frame* homônimo, do *subframe* Escolha e do enquadramento Maternidade_não_Hegemônica. Tais evidências refletem a conceptualização do abortamento não apenas como questão de saúde pública – aspecto que motivou a pauta da SUG –, mas também de justiça

social, especificamente de justiça reprodutiva (SILLIMAN *et al.*, 2016). Nesse sentido, ao se abordarem questões permeadas por desigualdade social e racial, como é o caso da pauta da SUG n.º 14, vislumbra-se a defesa de uma noção de escolha centrada em diferenças concretas entre mulheres, considerando-as como sujeitos-cidadãos atrelados a coletividades e historicidades distintas.

Conforme discutimos ao longo do artigo, essa abordagem permitiu um tratamento dos dados previamente ao *upload* do *corpus* no Sketch Engine, o que resultou, por sua vez, em *subcorpora* temáticos processáveis por essa ferramenta. Diante disso, pontuamos o ineditismo da proposta, que não apenas utilizou o NVivo como recurso de identificação de unidades temáticas – o uso mais prototípico da ferramenta em estudos que partem de fontes textuais –, mas o integrou ao processo de compilação de *corpus*, com vistas a uma análise cognitivo-discursiva dos dados.[26] Dessa forma, a proposta de uso do NVivo constituiu, ao mesmo tempo, uma primeira etapa de análise de *frames* e um processo de (re)compilação do *corpus* para fins de exploração dos evocadores no Sketch Engine. Além disso, observamos que, embora a relevância do SE em pesquisas semântico-cognitivas tenha sido extensamente corroborada (CHISHMAN *et al*., 2014, 2015, 2018), é a primeira vez que utilizamos um recurso novo da ferramenta, o Sketch Difference, para comparar o uso de algumas combinatórias lexicais entre os *subcorpora*, atrelando tais evidências à descrição de *frames*. Diante disso, salientamos que a pertinência dos procedimentos aqui discutidos para outras análises cognitivo-discursivas, ou para outros tipos de *corpora*, ainda necessita ser verificada.

Agradecimentos

Agradecemos à Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), que concedeu à primeira autora uma bolsa de doutorado CAPES/PROSUC (Código de Financiamento 001); e à Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), que financiou a aquisição de uma licença de uso do *software* NVivo.

[26] Para conferir as aproximações e os distanciamentos entre nossa proposta metodológica e estudos anteriores, sugerimos a leitura de Santos (2020).

Declaração de contribuição de cada autora

Este artigo é um desdobramento da tese de doutorado desenvolvida pela primeira autora e orientada pela segunda autora. O referencial teórico foi escrito por Aline Nardes dos Santos e modificado após sugestões de Rove Chishman. No desenho metodológico, o ponto de partida foi a proposta *middle-out* que Chishman tem adotado em suas pesquisas. Conjuntamente, as autoras aplicaram esse aporte à metodologia do estudo. No percurso analítico, o ponto de partida foram os dados da tese de Santos, os quais foram revisitados na proposta deste artigo. Por fim, as demais seções foram planejadas e revisadas colaborativamente, assim como a redação final do texto.

Referências

BELTRAME, P. B. *Aborto*: a controvérsia das feminilidades. 2016. 106f. Dissertação (Mestrado em Antropologia) – Programa de Pós-Graduação em Antropologia, Universidade Federal de Pernambuco, Natal, 2016.

BERBER SARDINHA, T. Linguística de Corpus: Histórico e Problemática. *D.E.L.T.A.,* São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://doi.org/10.1590/S0102-44502000000200005. Disponível em: http://www.scielo.br/pdf/delta/v16n2/a05v16n2.pdf. Acesso em: 25 set. 2020.

BOOTH, K. J. The Meaning of the Social Body: Bringing George Herbert Mead to Mark Johnson's Theory of Embodied Mind. *William James Studies*, [*S.l.*], v. 1, n. 1, p. 1-18, 2016. Disponível em: https://www.jstor.org/stable/26203794?seq=1#metadata_info_tab_contents. Acesso em: 17 fev. 2020.

BRASIL. Constituição (1988). *Constituição da República Federativa do Brasil*, 1988, Brasília, DF: Presidência da República, 1988. Disponível em: http://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm. Acesso em: 11 set. 2020.

BRASIL. Ministério da Saúde. *Atenção Humanizada ao Abortamento*: Norma Técnica. Brasília: Ministério da Saúde, 2011.

BRASIL. Senado Federal. Sugestão n° 15, de 2014. Atividade Legislativa. Brasília, 2014. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/119431. Acesso em: 1 mar. 2020.

BYBEE, J. Usage-Based Models in Linguistics: an Interview with Joan Bybee. Entrevista concedida a Tiago Timponi Torrent. *Revista Linguíſtica*, Rio de Janeiro, v. 8, n. 1, p. 1-6, 2012.

CHISHMAN, R. *et al.* Field – Dicionário de Expressões do Futebol: um recurso lexicográfico baseado no aporte teórico-metodológico da Semântica de Frames e da Linguística de Corpus. *Signo*, Santa Cruz do Sul, v. 39, n. 67, p. 25-35, 2014. DOI: https://doi.org/10.17058/signo.v39i67.5128. Disponível em: https://online.unisc.br/seer/index.php/signo/article/view/5128. Acesso em: 22 out. 2020.

CHISHMAN, R. *et al.* The Relevance of the Sketch Engine Software to Build Field – Football Expressions Dictionary. *Revista de Estudos da Linguagem*, Belo Horizonte, v. 23, n. 3, p. 769-796, 2015. DOI: https://doi.org/10.17851/2237-2083.23.3.769-796. Disponível em: http://www.periodicos.letras.ufmg.br/index.php/relin/article/view/8918. Acesso em: 21 out. 2020.

CHISHMAN, R. *et al.* Dicionário Olímpico: a Semântica de Frames encontra a lexicografia eletrônica. *In*: FINATTO, M. J. B.; REBECHI, R. R.; SARMENTO, S.; BOCORNY, A. E. P. (org.). *Linguística de Corpus*: perspectivas. Porto Alegre: Instituto de Letras - UFRGS, 2018. p. 265-298.

CHISHMAN, R. *et al.* Challenges and Difficulties in the Development of Dicionário Olímpico (2016). *In*: ELEX CONFERENCE, 2019, Sintra. *Proceedings* […]. Sintra: Lexical Computing CZ s.r.o., Brno, Czech Republic, 2019. p. 622-641.

DINIZ, D.; MEDEIROS, M.; MADEIRO, A. Pesquisa Nacional de Aborto 2016. *Ciência & Saúde Coletiva*, Rio de Janeiro, v. 22, n. 2, p. 653-660, 2017. DOI: https://doi.org/10.1590/1413-81232017222.23812016. Disponível em: http://www.scielo.br/scielo.php?pid=S141381232017000200653&script=sci_abstract&tlng=pt. Acesso em: 8 fev. 2020.

DINIZ, D.; MEDEIROS, M. Aborto no Brasil: uma pesquisa domiciliar com técnica de urna. *Ciência & Saúde Coletiva*, Rio de Janeiro, v. 15, supl. 1, p. 959-966, 2010. DOI: https://doi.org/10.1590/S1413-81232010000700002. Disponível em: http://www.scielo.br/pdf/csc/v15s1/002.pdf. Acesso em: 03 jul. 2020.

DUQUE, P. H. Discurso e Cognição: uma abordagem baseada em frames. *Revista da ANPOLL*, Florianópolis, v. 1, n. 39, p. 25-48, 2015. DOI: https://doi.org/10.18309/anp.v1i39.902 Disponível em: https://revistadaanpoll.emnuvens.com.br/revista/article/view/902/829. Acesso em: 10 set. 2020.

ELIAS, M. L. G. G. R. Conservadorismo, feminismo e o judiciário como arena em disputa: debate sobre aborto. *In*: ENCONTRO DA ABCP, 11., 2018, Curitiba. *Anais* [...]. Curitiba: ABCP, 2018. p. 1-26.

FILLMORE, C. J. An Alternative to Checklist Theories of Meaning. *In*: ANNUAL MEETING OF THE BERKELEY LINGUISTICS SOCIETY, 1., Berkeley. *Proceedings* […]. Berkeley: Berkeley Linguistics Society, 1975. p. 123-131. DOI: https://doi.org/10.3765/bls.v1i0.2315

FILLMORE, C. J. Frame Semantics and the Nature of Language. *In*: CONFERENCE ON THE ORIGIN AND DEVELOPMENT OF LANGUAGE AND SPEECH, 1976, New York. *Proceedings* […] New York: New York Academy of Sciences, 1976. p. 20-32. DOI: https://doi.org/10.1111/j.1749-6632.1976.tb25467.x

FILLMORE, C. J. Frame Semantics. *In:* THE LINGUISTICS SOCIETY OF KOREA (org.). *Linguistics in the Morning Calm*. Seoul: Hansinh Publishing, 1982. p. 111-137.

FILLMORE, C. J. Frames and the Semantics of Understanding. *Quaderni di Semantica*, [S.l.], v. 6, n. 2, p. 222-254, 1985.

FILLMORE, C. J.; BAKER, C. A Frames Approach to Semantic Analysis. *In*: HEINE, B.; NARROG, H. (ed.). *The Oxford Handbook of Linguistic Analysis*. New York: Oxford University Press, 2010. p. 313-339.

FONTES, M. R. *Frames e valores*: um estudo sobre a normatividade no espaço escolar. 2012. 157f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de Juiz de Fora, Juiz de Fora, 2012. Disponível em: http://www.ufjf.br/ppglinguistica/files/2009/12/FONTES-Mariana-Rocha-2012-Disserta%C3%A7%C3%A3o.pdf. Acesso em: 20 nov. 2020.

GEERAERTS, D.; KRISTIANSEN, G.; PEIRSMAN, Y. Introduction. Advances in Cognitive Sociolinguistics. *In*: _____. (ed.). *Advances in Cognitive Sociolinguistics*. Berlin; New York: De Gruyter Mouton, 2010. p. 1-22. DOI: https://doi.org/10.1515/9783110226461

GIL, A. C. *Como elaborar projetos de pesquisa*. 4. ed. São Paulo: Atlas, 2008.

GUIZZO, B. S.; KRZIMINSKI, C. O.; OLIVEIRA, D. L. L. C. O Software QSR NVIVO 2.0 na análise qualitativa de dados: ferramenta para a pesquisa em ciências humanas e da saúde. *Revista Gaúcha de Enfermagem*, Porto Alegre, v. 24, n. 1, p. 53-60, 2003. Disponível em: https://seer.ufrgs.br/RevistaGauchadeEnfermagem/article/view/4437. Acesso em: 10 jun. 2018.

GUTTMACHER INSTITUTE. *Abortion in Latin America and the Caribbean*. New York: Guttmacher Institute, 2017. Disponível em: https://www.guttmacher.org/sites/default/files/factsheet/ib_aww-latin-america.pdf. Acesso em: 15 mar. 2020.

HANKS, W. F. O que é contexto. *In*: BENTES, A. C.; REZENDE, R. C.; MACHADO, M. R. (org.). *Língua como prática social*: das relações entre língua, cultura e sociedade a partir de Bourdieu e Bakhtin. São Paulo: Cortez, 2008. p. 169-203.

KOCH, I. V.; MORATO, E.; BENTES, A. C. Ainda o contexto: algumas considerações sobre as relações entre contexto, cognição e práticas sociais na obra de Teun van Dijk. *Revista Latinoamericana de Estudios del Discurso*, Brasília, v. 11, p. 79-92, 2011. DOI: https://doi.org/10.35956/v.11.n1.2011.p.79-91. Disponível em: http://raled.comunidadaled.org/index.php/raled/article/view/93. Acesso em: 20 set. 2020.

KOESTER, A. Building Small Specialised Corpora. *In*: MCCARTHY, M.; O'KEEFE, A. (ed.). *The Routledge Handbook of Corpus Linguistics*. London; New York: Routledge, 2010. p. 66-79. DOI: https://doi.org/10.4324/9780203856949-6

LAGE, M. C. Utilização do *software* NVivo em pesquisa qualitativa: uma experiência em EaD. *ETD - Educação Temática Digital*, Campinas, v. 12, p. 198-226, 2011. DOI: https://doi.org/10.20396/etd.v12i0.1210. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/etd/article/view/1210/pdf_57. Acesso em: 14 mai. 2020.

LANGACKER, R. W. *Cognitive Grammar*: A Basic Introduction. New York: Oxford University Press, 2008. DOI: https://doi.org/10.1093/acprof:oso/9780195331967.001.0001

LANGLOTZ, A. *Creating Social Orientation Through Language*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2015. DOI: https://doi.org/10.1075/celcr.17.

LIMA, F. R. O.; MIRANDA, N. S. O frame semântico como uma ferramenta analítica de compreensão de experiências sociais educacionais. *Revista Gatilho*, Juiz de Fora, v. 8, p. 1-14, 2013. Disponível em: http://www.ufjf.br/revistagatilho/files/2013/05/O-frame-sem%C3%A2ntico-como-ferramenta-anal%C3%ADtica.pdf. Acesso em: 25 mar. 2020.

MIRANDA, N. S. Domínios conceptuais e projeções entre domínios: uma introdução ao Modelo dos Espaços Mentais. *Veredas: Revista de Estudos Linguísticos*, Juiz de Fora, v. 3, n. 1, p. 81-95, 1999. Disponível em: https://veredas.ufjf.emnuvens.com.br/veredas/article/view/500. Acesso em: 03 mar. 2020.

MIRANDA, N. S. O caráter partilhado da construção da significação. *Veredas*: *Revista de Estudos Linguísticos*, Juiz de Fora, v. 5, n. 1, p. 57-81, 2001. Disponível em: http://www.ufjf.br/revistaveredas/files/2009/12/artigo49.pdf. Acesso em: 08 jul. 2020.

MIRANDA, N. S.; BERNARDO, F. C. Frames, discurso e valores. *Cadernos de Estudos Linguísticos*, Campinas, v. 55, n. 1, p. 81-97, 2013. DOI: https://doi.org/10.20396/cel.v55i1.8636596. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/cel/article/view/8636596. Acesso em: 16 mar. 2020.

MORATO, E. M. A noção de frame no contexto neurolinguístico: o que ela é capaz de explicar? *Cadernos de Letras da UFF*, Niterói, n. 41, p. 93-113, 2010. Disponível em: http://www.uff.br/cadernosdeletrasuff/41/artigo4.pdf. Acesso em: 10 jan. 2020.

SALOMÃO, M. M. Gramática e interação: o enquadre programático da hipótese sociocognitiva sobre a linguagem. *Veredas: Revista de Estudos Linguísticos*, Juiz de Fora, v. 1, n.1, p. 23-39, 1997. Disponível em: https://bit.ly/2tZRwlH. Acesso em: 25 mar. 2020.

SALOMÃO, M. M. Teorias da linguagem: a perspectiva sociocognitiva. *In*: FÓRUM DE LINGUAGEM, 2., Rio de Janeiro. *Anais* [...]. Rio de Janeiro: Universidade Federal do Rio de Janeiro*,* 2006. p. 1-13.

SALOMÃO, M. M. Teorias da linguagem: a perspectiva sociocognitiva. *In*: MIRANDA, N. S.; SALOMÃO, M. M. (org.). *Construções do Português do Brasil*: da gramática ao discurso. Belo Horizonte: UFMG, 2009. p. 20-32.

SANTOS, A. N. *A Sugestão Legislativa nº 15/2014:* entrelaçamentos e reenquadramentos de frames semânticos no debate sobre os direitos reprodutivos das mulheres no Brasil. 2020. 291f. Tese (Doutorado em Linguística Aplicada) – Programa de Pós-Graduação em Linguística Aplicada, Universidade do Vale do Rio dos Sinos, São Leopoldo, RS, 2020. Disponível em: http://www.repositorio.jesuita.org.br/handle/UNISINOS/9111. Acesso em: 20 nov. 2020.

SANTOS, A. N.; CHISHMAN, R. L. O. Frames de compreensão e corpora: estudo de caso com uso do Sketch Engine. *In*: FINATTO, M. J. B.; REBECHI, R. R.; SARMENTO, S; BOCORNY, A. E. P. (org.). *Linguística de Corpus*: perspectivas. Porto Alegre: Instituto de Letras da UFRGS, 2018. p. 183-206.

SILLIMAN, J. *et al. Undivided Rights*: Women of Color Organizing for Reproductive Justice. Chicago: Haymarket Books, 2016.

SILVA, A. S. Discurso na mente e na comunidade. Para a sinergia entre a Linguística Cognitiva e a Análise (Crítica) do Discurso. *Revista Portuguesa de Humanidades*, Braga, v. 9, n. 1, p. 53-78, 2015.

SIMAN, J. H. *Frames de doença de Alzheimer*. 2015. 155f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Estadual de Campinas, 2015. Disponível em: http://taurus.unicamp.br/bitstream/REPOSIP/270612/1/Siman_JosieHelen_M.pdf. Acesso em: 16 mar. 2020.

SIQUEIRA, A. C. T. *A Semântica de Frames na análise do discurso discente*: traçando o perfil do professor de português. 2013. 152f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de Juiz de Fora, 2013. Disponível em: https://repositorio.ufjf.br/jspui/bitstream/ufjf/900/1/amandacristinatestasiqueira.pdf. Acesso em: 20 ago. 2020.

TANNEN, D. What's in a Frame? Surface Evidence for Underlying Expectations. *In*: FREEDLE, R. (ed.). *New Directions in Discourse Processing*. Norwood: Ablex, 1979. p. 137-181.

TOMASELLO, M. *The Cultural Origins of Human Cognition*. Cambridge: Harvard University Press, 1999.

TOMASELLO, M. *Constructing a Language*: A Usage-Based Theory of Language Acquisition. Cambridge; London: Harvard University Press, 2003.

TOMASELLO, M. *Origins of Human Communication*. Cambridge; London: The MIT Press, 2008. DOI: https://doi.org/10.7551/mitpress/7551.001.0001

VEREZA, S. Mal comparando…: os efeitos argumentativos da metáfora e da analogia numa perspectiva cognitivo-discursiva. *SCRIPTA*, Belo Horizonte, v. 20, n. 40, p. 18-35, 2016a. DOI: https://doi.org/10.5752/P.2358-3428.2016v20n40p18. Disponível em: http://periodicos.pucminas.br/index.php/scripta/article/view/13964. Acesso em: 9 mar. 2020.

VEREZA, S. Cognição e sociedade: um olhar sob a óptica da Linguística Cognitiva. *Linguagem em (Dis)curso*, Tubarão, v. 16, n. 3, p. 561-573, 2016b. DOI: https://doi.org/10.1590/1982-4017-160303-0416d15. Disponível em: http://www.scielo.br/pdf/ld/v16n3/1518-7632-ld-16-03-00561.pdf. Acesso em: 07 mar. 2020.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1191-1228, 2021

De *marcar la cancha* a *una canchereada* na metaforização da política pelo futebol: análise de unidades fraseológicas especializadas em *corpus* jornalístico argentino

From "marcar la cancha" to "una canchereada" in the metaphorization of politics by football: analysis of specialized phraseological units in Argentinean journalistic corpus

Ariel Novodvorski
Universidade Federal de Uberlândia (UFU), Uberlândia, Minas Gerais / Brasil
arivorski@ufu.br
http://orcid.org/0000-0003-1370-8334

Cleci Regina Bevilacqua
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul / Brasil
cleci.bevilacqua@ufrgs.br
http://orcid.org/0000-0002-1002-9080

Resumo: Este artigo apresenta uma análise de termos e unidades fraseológicas especializadas, do âmbito do futebol, em processos de metaforização com o domínio alvo da política. A pesquisa utiliza um *corpus* jornalístico monolíngue em espanhol rio-platense, da coluna *Humor Político* do jornal argentino *Clarín*. A utilização de programas computacionais para análises lexicais característicos da pesquisa com a Linguística de *Corpus* viabilizou a identificação e extração de termos e fraseologias especializadas presentes no *corpus*. A análise das unidades em contexto apontou para a construção metafórica do campo mais abstrato da política por meio de figuras mais concretas do meio futebolístico. A consulta a *corpora* disponíveis *on-line* corroborou as premissas e os achados no *corpus* de estudo.

Palavras-chave: terminologia; unidades fraseológicas especializadas; metáfora; linguística de *corpus*; *corpus* jornalístico.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1191-1228

Abstract: The following paper provides an analysis of soccer terms and specialized phraseological units in metaphorization processes with the target area of politics. The research uses a monolingual journalistic *corpus* in Spanish rio-platense, from the politics section *Humor Político* of the Argentinean newspaper *Clarín*. The use of computer programs for lexical analysis typical of *Corpus* Linguistics research has made it possible to identify and extract terms and specialized phrases found in the *corpus*. The analysis of units in context pointed to the metaphorical construction of the most abstract field of politics by means of more concrete figures of the soccer environment. The *corpora* consultation available online corroborated the premises and findings in the *corpus* of study.

Keywords: terminology; specialized phraseological units; metaphor; *corpus* linguistics; journalistic *corpus*.

Submetido em 31 de agosto de 2020

Aceito em 19 de outubro de 2020

1 Introdução

Este artigo é um recorte da pesquisa de Pós-doutorado do primeiro autor deste texto, junto ao PPG-Letras da UFRGS, com conclusão prevista para dezembro de 2020. Essa pesquisa nasce da exploração de diferentes *corpora*, monolíngues, comparáveis e paralelos, no par linguístico espanhol/português, compilados a partir da seção de opinião de jornais argentinos e brasileiros, com a trama política desses países como tema principal e a observação das características lexicais, como objeto de estudo. A motivação para a escolha da temática e consequente compilação dos *corpora* partiu de nosso interesse, por um lado, pelo acompanhamento da situação e trama política na região, enquanto leitores de diferentes jornais de ampla circulação e de livre acesso *on-line* e, por outro lado, pela an*á*lise das combinatórias lexicais identificadas a partir de unidades terminológicas, em uma perspectiva contrastiva e comparável.[1] Apresentamos parte desse material e dos resultados em

[1] Do ponto de vista da Metáfora Conceptual, das características do *corpus* e da utilização de ferramentas da Linguística de *Corpus*, os trabalhos de Berber Sardinha (2007a, 2008, 2010) e de Sperandio (2009, 2010) são pontos de contraste relevantes com esta pesquisa, como será observado na seção teórica, por abordarem o estudo empírico de metáforas no plano da política brasileira, entre outros, do ex-presidente Lula e do Movimento dos Trabalhadores Rurais Sem Terra (MST).

diferentes eventos, nos últimos quatro anos. Para o presente trabalho, utilizaremos exclusivamente o *corpus* integrado por textos coletados da seção de opinião do jornal *Clarín*, em específico a coluna dominical *Humor político*, do jornalista argentino Alejandro Borensztein,[2] que passamos a denominar *Corpus AleBores*. Essa coluna, como observa o próprio autor, é publicada todo domingo há 12 anos.

A partir do contato inicial com o *corpus*, ainda enquanto leitores, pudemos perceber a riqueza lexical e a dificuldade que poderia representar a compreensão do sentido de determinadas unidades fraseológicas, de uso especializado no contexto da trama política. Por um lado, isso ocorreria pelo fato das diferentes áreas de especialidade convergentes e implicadas nos textos; por outro lado, pelo conhecimento sócio-histórico e cultural exigido, pela alusão a fatos que deveriam fazer parte da memória dos leitores, fundamentais para o estabelecimento das relações na construção dos sentidos. Cabe considerar, ainda, questões relacionadas à função comunicativa de determinadas ocorrências nos textos, isto é, à função pragmática implicada. Some-se a isso, no ato da leitura, a necessidade de reconhecimento e compreensão dessas unidades fraseológicas constituídas nos textos, isto é, o conhecimento mais estritamente linguístico, para reconhecimento e/ou identificação dos constituintes que formam as fraseologias, função necessária para compreensão da relação existente, muitas das vezes metafórica, entre os domínios vinculados por essas unidades fraseológicas.

Para exemplificar, em "También le pasó a Scioli en 2015. Se la dejaron abajo del arco y la tiró por arriba del travesaño" (BORENSZTEIN, 2018, nosso destaque). No âmbito da política argentina, Daniel Scioli foi o candidato à presidência argentina pelo Kirchnerismo, no final do segundo mandato de Cristina Kirchner. As fraseologias em destaque ilustram uma jogada de futebol em que alguém deixou a bola para outro jogador, na porta do gol ("se la dejaron abajo del arco"), ou seja, bastaria que o jogador chutasse para fazer o gol; contudo, chutou para fora, acima do travessão ("la tiró por arriba del travesaño"). Caberia a pergunta, aqui, qual a necessidade ou razão para utilizar a imagem de uma jogada de futebol para ilustrar uma situação do mundo da política? Tais unidades fraseológicas do domínio do futebol ("dejársela [la pelota/a

[2] Disponível em: https://www.clarin.com/autor/alejandro-borensztein.html. Acesso em: 10 ago. 2020.

bola] abajo del arco" e "tirarla [la pelota/a bola] por arriba del travesaño") metaforizam, por meio de uma imagem mais concreta, uma situação mais abstrata do meio político, em que um candidato perdeu as eleições presidenciais, em circunstâncias que estariam bastante facilitadas, pois teriam deixado todo o caminho pronto para que fosse eleito. Em linguagem futebolística, *deixaram na cara do gol, deram o gol servido, só faltava chutar para o gol*.

Realizamos a exploração do *Corpus AleBores* em diversas direções, antes mesmo de encontrar o rumo que passou a conduzir a pesquisa, por meio da utilização das ferramentas *WordList, KeyWords* e *Concord*, do programa para análises lexicais *WordSmith Tools* (WST), versão 7,0 (SCOTT, 2016), em suas diferentes funcionalidades. Também recorremos ao suporte de recursos disponíveis *on-line* para consulta, a saber, o *Corpus del Español* (DAVIES, 2016, 2018), em sua versão dialetal, e o *Sketch Engine* (KILGARRIFF, 2019), utilizados fundamentalmente como *corpora* de consulta e contraste, para corroborar ou reformular hipóteses, a partir dos achados em nosso *corpus*. Como objeto de estudo tomamos, num primeiro momento, o gênero textual artigo de opinião, pelo prisma da Linguística de *Corpus* (LC). Num segundo momento, as Unidades Fraseológicas Especializadas (UFEs), caracterizadas pela presença de unidades léxicas que adquirem um valor especializado na área do futebol, enquanto candidatos a termos, mas utilizados metaforicamente em referência ao domínio da política. Ambos os trabalhos foram feitos na perspectiva das análises descritivas e empírico-dedutivas, com suporte tecnológico dos recursos anteriormente mencionados.

Por meio da extração e análise da lista de palavras-chave com a ferramenta *KeyWords* do WST, a pesquisa sobre o gênero possibilitou a identificação de características que apontam para a estabilidade e organização interna dos diferentes textos que compuseram o *corpus*. Algumas das características apontadas por Berber Sardinha (2009, p. 25-26) foram observadas, como o fato de os gêneros serem social, cultural e historicamente definidos, sequenciados internamente e compostos por uma lexicogramática distinta. No entanto, o maior destaque foi a identificação de áreas temáticas que se mostraram salientes nos resultados, em especial, o domínio do futebol, pela recorrência de termos do meio futebolístico.

A análise das palavras-chave extraídas dos *corpora* jornalísticos que compilamos especificamente para dois trabalhos, um apresentado no V Congresso Internacional de Fraseologia e Paremiologia, realizado na USP (NOVODVORSKI, 2018),[3] e outro apresentado no XIV Encontro de Linguística de *Corpus*, realizados na UFRGS/Unisinos (NOVODVORSKI, 2017),[4] a partir da filtragem dos resultados, possibilitou primeiro a identificação de diversos vocábulos da área do futebol, atestada pela frequência registrada por esses termos. Tudo se configurou como forte indício para pressupor que a representação do domínio futebolístico, nos textos que compõem o *corpus*, tenha uma relação direta na definição do domínio da política. Por meio das linhas de concordância, a Figura 1 ilustra uma aproximação para a análise contextualizada de um dos candidatos a termo *(palo/palos)*.

FIGURA 1 – Linhas de concordância com *palo/palos* na ferramenta *Concord*

Fonte: A pesquisa (WST, versão 7.0)

[3] A política pela ótica do futebol: uma análise término-fraseológica em corpus jornalístico de humor político. Trabalho apresentado no V Congresso Internacional de Fraseologia e Paremiologia, em São Paulo, 2018.

[4] Humor político: o gênero pelo prisma da Linguística de Corpus.Trabalho apresentado no XIV ELC./IX EBRALC, em Porto Alegre, 2017.

Como se observa na figura anterior, as linhas de concordância geradas a partir da busca pelo vocábulo *palo* revelam mais de um sentido: no âmbito do futebol, *pegar en el palo* equivale a *bater na trave* (linhas 01, 07-10); no campo da economia, *palo(s)* equivalem a *milhão(ões)*, e *palo(s) verde(s)* a *milhão(ões) de dólares*; já em *moler a palos* temos a manifestação de um ato violento, correspondente a *dar uma paulada*. Além dessas, outras ocorrências como *ni a palos*, que equivale a *de jeito nenhum*, podem ser observadas na figura.

Das ocorrências registradas nessas linhas de concordância, tomamos a primeira linha, para ilustrar o aspecto de nossa proposta de análise: a metaforização da política pelo futebol. Pressupomos que há uma relação metafórica entre dois domínios ou áreas de conhecimento, em que as características de uma passam a ser assimiladas para a compreensão da outra. A frase "...si no es Stiuso le pega en el palo" (*...se não for Stiuso, bate na trave*) apresenta, por um lado, uma UFE formada por um termo da área do futebol, *palo/trave*, que, como já apontado, equivale a *bater na trave*. Para uma compreensão mais completa do fragmento, é necessário saber que Stiuso foi um agente secreto, dos serviços de inteligência, durante vários governos na Argentina. A suspeita, na época da publicação do texto jornalístico, era de que poderia haver alguma relação com o assassinato do procurador federal Alberto Nisman, que apareceu morto logo após denunciar Cristina Kirchner, presidenta na época, na véspera de sua declaração no Congresso Nacional. A relação metafórica existente entre os domínios fonte (futebol) e alvo (política) apontam que essa suspeita de envolvimento do agente Stiuso seria uma quase certeza, assim como uma bola que vai em direção ao gol e que, se não entrar, bateria na trave, o que seria quase um gol.

Como se observa, pela extração dos termos da área de futebol e pela identificação das UFEs, com auxílio das ferramentas do WST e dos recursos utilizados como *corpora* de consulta (DAVIES, 2016, 2018; KILGARRIFF, 2019), o passo seguinte correspondeu à identificação dos ambientes de ocorrência para completude da significação. Isto é, o sentido metafórico das UFEs é compreendido, quando acionada a perspectiva pragmática relacionada à referência ao meio político em questão, que demanda, também, um conhecimento contextual (social, histórico e cultural), além da percepção dos aspectos funcionais do texto em si, das funções comunicativas implicadas (CABRÉ, 1993). Desse

modo, as três perspectivas se encontram interligadas, a da linguagem, do conhecimento e da comunicação.

A problematização que deriva de toda a contextualização anterior e que delineia, de algum modo, os rumos de nossa pesquisa e do texto aqui apresentado é: De que maneira acionamos, enquanto leitores e membros de uma comunidade sociocultural, as informações que são veiculadas nos textos, a tal ponto de conseguirmos compreender os sentidos construídos para o tratamento de questões políticas, por meio da linguagem futebolística? Em outras palavras, como ocorre a metaforização entre duas áreas de conhecimento, a política e o futebol, num *corpus* de textos de opinião?

Com base nas questões e tomando como ponto de partida Cabré (2002), formulamos nossa hipótese deste modo: existem aspectos cognitivos, linguísticos e pragmáticos entrelaçados e englobados por uma dimensão cultural mais ampla, acionada a partir de traços e vestígios identificados na leitura do texto. Isso equivale a dizer que o reconhecimento de determinadas marcas linguísticas nos textos, especificamente as UFEs, aciona nas lembranças do leitor o domínio do futebol, transferindo características dessa área, assimiladas na compreensão do domínio metaforizado, o campo da política. Considerando a perspectiva da comunicação, o autor dos textos promove a representação de um domínio por meio das características do outro, acarretando uma percepção mais acessível das questões políticas. Como o próprio autor aponta em diversas oportunidades "el fútbol lo explica mejor" (*com o futebol dá para explicar melhor*).

Dado o alcance da cultura do futebol, talvez pela plasticidade imagética que evocam determinadas jogadas e regras do jogo desse domínio e por estarem de algum modo no repertório popular, em especial em países como Argentina e Brasil, por considerar o âmbito concernente a este trabalho, expressões típicas desse domínio funcionam para metaforizar tanto situações mais corriqueiras do cotidiano quanto diversos âmbitos mais abstratos, não apenas o político. Assim, quando dizemos que *alguém pisou na bola*, ou que *precisamos vestir a camisa da empresa ou instituição*, ou que *alguém precisa baixar a bola* ou que *levou cartão vermelho*, sabemos que não estamos falando de futebol, mas por analogia nos referimos a um domínio mais abstrato, metaforizado. Com isso, muitas fraseologias características do futebol circulam na vida das pessoas, que recorrem à linguagem do futebol para falar de outras

coisas. Para ilustrar, algumas frases típicas em espanhol foram tratadas e podem ser encontradas na matéria do jornal uruguaio *El País*, "Voces: El lenguaje de la cancha está presente en el de todos los días".[5]

Desse modo, para além da descrição do *corpus* e dos procedimentos metodológicos envolvidos na pesquisa, nossos objetivos neste trabalho são (a) identificar e descrever termos e fraseologias especializadas, em cujos usos subjazem metáforas conceptuais, nos textos que perfazem o *corpus* de estudo, pelo cotejo dos sentidos promovidos nas áreas de especialidade relacionadas, e (b) analisar os aspectos referentes à metaforização entre os domínios da política e do futebol. Para tanto, apresentaremos no artigo um recorte dos resultados alcançados, em torno de duas Unidades Terminológicas (UT) referentes ao espaço destinado para a prática do futebol, a UFE "marcar la cancha" (*marcar o campo de jogo*) e a UT "canchereada", que pode ser traduzida como *malandragem* ou *catimba*, no meio futebolístico.

Por meio do quadro teórico-metodológico que será descrito a seguir e com o suporte das ferramentas, recursos e abordagem próprios da LC, buscaremos alcançar esses objetivos.

2 Fundamentação teórica

A presente pesquisa engloba as seguintes fontes bibliográficas, conforme cada uma das respectivas áreas teóricas: (1) Terminologia (CABRÉ, 1993, 2002, 2005; KRIEGER; SANTIAGO; CABRÉ, 2013); (2) Fraseologia Especializada (BEVILACQUA, 1998, 1999, 2004; CABRÉ; ESTOPÀ; LORENTE, 1996; ORENHA, 2009; ORENHA; CAMARGO, 2009); (3) Fraseologia (CORPAS PASTOR, 2010); (4) Metáfora (BERBER SARDINHA, 2007a, 2008, 2010; SPERANDIO, 2009, 2010, DEIGNAN, 2005, 2012; LAKOFF; JOHNSON, 1980); (5) Gramáticas Descritivas e Dicionários de Usos da língua espanhola e portuguesa (ANANÍA, 2005; BORBA, 2002; BOSQUE; DEMONTE, 1999; FONTANARROSA; SANZ, 1994; GOVERNATORI; LAROCCA, 2014; HOUAISS, 2009; MOLINER, 2008) e (6) Linguística de *Corpus* (BERBER SARDINHA, 2004, 2009; PARODI, 2008, 2010).

[5] Disponível em: https://www.ovaciondigital.com.uy/futbol/voces-lenguaje-cancha-presente-todos-dias.html. Acesso em: 14 out. 2020.

Para além dessas áreas, duas referências complementam este quadro, por cobrirem aspectos culturais, históricos e sociais da Argentina, em especial em torno da política, do futebol e da linguagem, escritos a partir da visão de dois jornalistas brasileiros com ampla experiência no país vizinho: Ariel Palacios e Guga Chacra. Ambos os livros foram publicados pela editora *Contexto*: *Os hermanos e nós* (PALACIOS; CHACRA, 2014) e *Os argentinos* (PALACIOS, 2015). É importante destacar, também, o lugar relevante que ocupam os diferentes programas computacionais, utilizados nas pesquisas desenvolvidas no âmbito da LC. Nesse sentido, também se incluem o *software* WST,[6] versão 7,0 (SCOTT, 2016), as plataformas *on-line Corpus del Español*, nas versões dialetal[7] e *NOW*[8] *– News on the Web* (DAVIES, 2016, 2018), e *Sketch Engine*[9] (KILGARRIFF, 2019).

Para situar a presente pesquisa no plano da Terminologia e, em particular, relacioná-la à especificidade do *corpus* de estudo, destacamos a delimitação feita por Cabré (1993) no tangente às peculiaridades das linguagens de especialidade, em contraponto à linguagem geral. A autora afirma que as linguagens especializadas se caracterizam em função da temática, dos falantes e das situações comunicativas, e que áreas como o comércio e o esporte também fazem parte da especialização, embora pudesse parecer que apenas temas científicos ou técnicos tivessem caráter de especialidade. Segundo a autora, dentre os aspectos linguísticos, funcionais e pragmáticos, são estes últimos os que possibilitam uma distinção mais clara entre as linguagens de especialidade e a língua comum, uma vez que permitem diferenciar termos de palavras. Assim, termos e palavras se diferenciam, considerando aspectos exclusivamente pragmáticos, a saber: pelos usuários, pelas situações de uso, pela temática e pelo tipo de discurso em que costumam aparecer.

Segundo Cabré (2005), o tema de uma comunicação é o ponto que determina o caráter de texto especializado. Para além das matérias científicas ou técnicas, âmbitos especializados de atividade como o esporte, dentre outros, produzem tipos de textos que se diferenciam dos textos considerados gerais, próprios de situações não profissionais. Em

[6] Disponível em: https://lexically.net/LexicalAnalysisSoftware/. Acesso em: 3 ago. 2020.
[7] Disponível em: https://www.corpusdelespanol.org/web-dial/. Acesso em: 10 ago. 2020.
[8] Disponível em: https://www.corpusdelespanol.org/now/. Acesso em: 10 ago. 2020.
[9] Disponível em: https://www.sketchengine.eu/. Acesso em: 30 jul. 2020.

termos lexicais, continua a autora, os textos especializados são específicos pela terminologia e pela fraseologia utilizadas e, também, em função da semântica das unidades terminológicas que integram os textos. Assim, "cada unidade terminológica corresponde a um nó cognitivo dentro de um campo de especialidade, e o conjunto de nós cognitivos, conectados por relações específicas, constitui a representação conceptual de determinada especialidade" (CABRÉ, 2005, p. 25).

A esse respeito, Bevilacqua (2004, p. 43) complementa que o Núcleo Terminológico (NT), no reconhecimento das Unidades Fraseológicas Especializadas (UFEs), "deveria ser uma unidade nominal que corresponda a um nó cognitivo do âmbito tratado" no alcance do *corpus* textual da pesquisa. Esta consideração é relevante, por possibilitar o tratamento das UFEs como unidades de representação e transmissão de conhecimento especializado. Desse modo, continua Bevilacqua (2004, p. 11), "é possível investigar como transmitimos e adquirimos conhecimento especializado, por meio de outras unidades linguísticas, que não são exclusivamente os termos". Estas últimas observações respaldam plenamente os objetivos que nos propomos alcançar na presente pesquisa, tal como observado na Introdução. Como já apontado, pressupomos que o reconhecimento de determinados agrupamentos lexicais nos textos, especificamente as UFEs, ativam no leitor o domínio do futebol. Deste domínio são transferidas características, que passam a ser assimiladas para a compreensão do outro domínio, que é metaforizado, o campo da política.

Cabré (2002) reforça que o caráter específico das unidades terminológicas reside em aspectos pragmáticos e que as significações são resultado de uma negociação entre especialistas, produzida no âmbito de um discurso especializado, por meio da realização de predicações que definem os significados das unidades. Com isso, a autora propõe uma teoria que possibilita um tratamento multidimensional dos termos, que contempla de modo integrado aspectos cognitivos, linguísticos, semióticos e comunicativos. Desse modo, o termo, enquanto unidade, é formado por uma dimensão semiótica e linguística, por outra cognitiva e por uma terceira vertente comunicativa.

Por outro lado, é importante a característica recursiva e dinâmica dos termos, que podem se deslocar de uma área de especialidade para outra, assim como as unidades do léxico comum passam para o léxico especializado, respectivamente *banalização* e *teminologização* (CABRÉ,

2005). Em entrevista (KRIEGER; SANTIAGO; CABRÉ, 2013), a pesquisadora observa que tem tratado a "terminologicidade" como um valor associado às unidades do léxico, a partir de uma concepção de termo enquanto unidade do léxico que ativa um sentido preciso, num contexto sociocomunicativo específico. Destacamos, portanto, que "os termos não são unidades diferentes das unidades do léxico, e sim unidades do léxico que adquirem características específicas em seu uso discursivo" (KRIEGER; SANTIAGO; CABRÉ, 2013, p. 331). A autora ainda observa que o objetivo da terminologia aplicada consiste tanto na compilação das unidades de valor terminológico, sobre uma temática e situação determinadas, quanto no estabelecimento das características, conforme essa situação.

Em se tratando da compilação e extração (semi)automática de unidades de valor terminológico, é importante recuperar e destacar as dificuldades encontradas e alguns dos critérios apontados, no trabalho empírico e seminal de Cabré, Estopà; Lorente (1996), no que diz respeito às UFEs. Em particular, chamamos a atenção para especificidade temática, quanto à determinação de uma unidade ser ou não um termo, e à delimitação do segmento formal que compõe uma unidade terminológica, no caso de unidades sintagmáticas. Por meio de uma ampla experimentação e centradas no estudo das unidades terminológicas polilexemáticas (UTP), em contraposição às UFEs e, no intuito de poder distinguir entre termos sintagmáticos e construções fraseológicas especializadas, as autoras assumem ser resultante essa distinção da aplicação de critérios como a categoria gramatical, a estrutura interna, a frequência, o grau de fixação e a variação dos componentes. Nesse ponto, destacam a regularidade observada na condição de termo dos núcleos dos sintagmas nominais, no caso das UTP, e da condição de termo dos complementos, nos sintagmas verbais, no caso das UFEs. Ainda reforçam não acreditarem na existência de termos nem fraseologia especializada de modo abstrato, uma vez que sempre adquirirão o valor de UT, UTP ou UFE, no âmbito de uma área de especialidade. Estes últimos apontamentos guardam relação direta com a proposta de nossa pesquisa.

Ressaltando o caráter cognitivo e comunicativo da proposta teórica de Cabré, a Teoria Comunicativa da Terminologia (TCT), Bevilacqua (2004, p. 10) destaca a importância do tratamento da fraseologia especializada no campo da Terminologia e afirma que, assim como as Unidades Terminológicas (UTs), também as UFEs "podem

ser consideradas um objeto de estudo poliédrico, multifuncional e multidimensional". Dessa maneira, a autora aponta a possibilidade de análise das UFEs, a saber: (1) pela perspectiva cognitiva, enquanto "unidades transmissoras de conhecimento especializado", que no contexto dessa abordagem teórica são denominadas Unidades de Conhecimento Especializado (UCE); (2) enquanto unidades que ocorrem em situações especializadas, denominadas Unidades de Comunicação Especializada (UNICOME); e (3) pela perspectiva da linguagem (abordagem também adotada nesta pesquisa), em que o tratamento das UFEs implica a percepção, descrição e análise, conforme a gramática da língua em questão e a partir da materialidade linguística, isto é, a utilização de critérios morfológicos, sintáticos, semânticos e pragmáticos, na análise de um *corpus* de textos produzidos por especialistas, em torno de temáticas específicas e em contextos de uso real da língua. Considerando o uso especializado, são denominadas Unidades de Significação Especializada (USE).

Bevilacqua (1998, 1999, 2004) define as UFEs como unidades sintagmáticas, integradas por um termo, com determinado grau de fixação e de frequência num *corpus* ou domínio especializado. A pesquisadora (2004) propôs, especificamente em sua tese, as unidades fraseológicas formadas por um núcleo eventivo, que denominou unidades fraseológicas especializadas eventivas (UFE eventivas). Essas unidades sintagmáticas são formadas, conforme suas propriedades, por um ou mais termos, que constituem o Núcleo Terminológico (NT), e um Núcleo Eventivo (NE), que é realizado textualmente por um verbo, um nome deverbal ou um particípio. Orenha (2009) e Orenha e Camargo (2009) analisam termos, colocações e colocações especializadas, por meio da extração de UFEs com subsídios da LC, em *corpora* paralelos bidirecionais e comparáveis, no par linguístico inglês/português, envolvendo traduções juramentadas e não juramentadas, de documentos de contratos e estatutos sociais. Considerando aspectos referentes aos profissionais da Tradução, as autoras apontam o valor de inclusão em obras terminográficas dos padrões de UFEs encontradas.

Corpas Pastor (2010) também destaca que os fatores de frequência de ocorrência e de coocorrência são elementos que compõem e possibilitam a identificação das UFs, além da institucionalização, a estabilidade, a idiomaticidade e a variação que tais unidades apresentam em diferente grau. Essa autora define a UF como uma combinação estável

de, pelo menos, duas palavras, cujo limite superior será o sintagma ou a oração composta e apresentará como traços inerentes a fixação ou a idiomaticidade por si mesmas, ou uma combinação de ambos os critérios (CORPAS PASTOR, 2010, p. 126).

Acerca da consideração do fator frequência e, em particular, pelo fato de definir as UFEs como "unidades que adquirem valor especializado *pelo* e *no texto* em que são utilizadas" (BEVILACQUA, 2004, p. 44), a autora destaca que, na seleção das unidades no *corpus*, há outros fatores que já seriam suficientes para "mostrar o valor especializado das unidades extraídas" (*idem*). Ou seja, o critério frequência não deveria ser conclusivo para a seleção das UFEs. Essa observação é de vital importância para nossa pesquisa, uma vez que, a depender da extensão do *corpus* de estudo, itens de baixa frequência ou que reportam uma única ocorrência (*hapax legomena*) poderiam ser desconsiderados, mesmo apresentando caráter especializado. Desse modo, assumimos que a frequência será observada, mas não como fator determinante para a constatação de que uma unidade seja considerada UFE, como será descrito na seção de *Corpus* e Metodologia.

A taxonomia proposta (CORPAS PASTOR, 2010, p. 127-136) para classificação das UFs define um primeiro nível de estruturação em três esferas: (1) as *colocações*, fixadas pelo uso, com algum grau de restrição combinatória; (2) as *locuções*, fixadas no sistema; e (3) os *enunciados fraseológicos* (parêmias e fórmulas), fixados na fala, formam parte do acervo sociocultural da comunidade do falante. Estes últimos se diferenciam das colocações e locuções pelo fato de chegarem a formar enunciados completos em si mesmos e a realizarem atos de fala, independente da combinação com outros elementos no discurso.

Já encerrando esta breve introdução e complementando o enquadramento teórico da pesquisa proposta, enfocamos o estudo da Metáfora pela visão cognitiva, como um recurso natural e intrínseco ao ser humano, por meio do qual se busca entender o mundo, processando mentalmente conceitos abstratos, partindo de conceitos concretos. Nesse sentido, mais do que se caracterizar como um traço da linguagem, a metáfora estabelece relações entre dois conceitos diferentes, que se unem por associação para compreendermos um deles a partir das características do outro. Portanto, "a essência da metáfora é entender e experimentar um tipo de coisa em termos de outra" (LAKOFF; JOHNSON, 1980, p. 41). Esses autores chegaram à conclusão de que nossa vida cotidiana está

impregnada de metáforas, e que nosso sistema conceptual ordinário, em torno do qual pensamos e atuamos, é de natureza metafórica. Pelo fato de lidarem com conceitos e de estabelecerem uma conexão entre duas áreas ou domínios semânticos, um concreto e outro abstrato, recebeu o nome de Metáfora Conceptual.

Deignan (2005, p. 14-29) destaca pontos que são essenciais para nossa pesquisa, quanto à estruturação do pensamento por meio das metáforas e com relação ao estudo de padrões pelo viés empírico que proporciona a LC. A autora observa que, geralmente, pesquisadores buscam metáforas conceptuais através da linguagem, na tentativa de encontrar padrões em frases e palavras, como evidência de metáforas conceptuais subjacentes. Desse modo, estabelece uma distinção fundamental entre *metáforas linguísticas* e *metáforas conceptuais*, afirmando que as metáforas linguísticas *realizam* as metáforas conceptuais, mas que isso não deve conduzir à interpretação de que a Teoria da Metáfora Conceptual tenha sido desenvolvida para explicar padrões linguísticos, pois o percurso é exatamente o inverso (DEIGNAN, 2012). O sentido de uma metáfora linguística é descrito em termos de *veículo*, que é o significado literal de uma palavra ou frase no domínio fonte (concreto), e *tópico*, que é o significado de uma palavra ou frase no domínio alvo (abstrato). Retomando o exemplo da Introdução deste artigo, "Si no es Stiusso le pega en el palo", temos como veículo a UFE "le pega en el palo" (bater na trave), do domínio fonte e mais concreto do futebol, e como tópico "Si no es Stiusso" (Se não for Stiusso, um agente de inteligência argentino), do domínio alvo e mais abstrato da política. Essa metáfora linguística realiza a metáfora conceptual mais genérica, em que POLÍTICA É FUTEBOL, grafada em caixa alta como convenção geral.

Aqui cabe uma breve observação quanto ao uso dos vocábulos *realizar* ou *realização*, que adquirem um sentido técnico, especializado, a partir dos estudos sistêmico-funcionais hallidayanos. Conforme a Linguística Sistêmico-Funcional, a linguagem verbal é um sistema sociossemiótico estratificado, que resulta da formulação e troca de significados. Os contextos mais amplos de cultura e de situação se realizam nos estratos da linguagem, passando pelos níveis da semântica do discurso, na construção de significados, e pela organização no estrato lexicogramatical, alcançando o nível da expressão, por meio de um texto escrito ou oral, realizado no plano gráfico ou fonético. Nesse sentido,

contexto sociocultural e linguagem são indissociáveis, uma vez que estão imbricados numa relação de mútua realização.

Assim como afirma Deignan (2005) e vários outros autores, as metáforas linguísticas são o principal tipo de evidência dada para a existência de metáforas conceptuais, atestada pela frequência de ocorrência. Talvez o principal tipo de evidência resida no potencial de as metáforas conceptuais revelarem a sistematicidade do léxico, no sentido de motivarem uma significativa quantidade de metáforas linguísticas, mapeáveis por meio de relações no interior ou entre campos lexicais. Além desse tipo de evidência, também a possibilidade tanto de criação como de compreensão de novas metáforas, por analogia a metáforas convencionais já existentes, entre determinados domínios (DEIGNAN, 2012). Por outro lado, o fato de a LC se ocupar da exploração de *corpora* de textos autênticos, em detrimento de frases inventadas ou descontextualizadas, corrobora a probabilidade de ocorrência das metáforas linguísticas em usos linguísticos reais. Num *corpus* de estudo, a autora menciona ter analisado palavras-chave referentes à temática de corrida de cavalos e jogos de apostas, em que os dados do *corpus* mostraram, por meio das colocações e linhas de concordância, frequentes usos metafóricos a respeito de campanhas políticas em inglês (DEIGNAN, 2005). É também relevante para nosso trabalho a observação feita por Deignan (2005), quando aponta ser provável que as pessoas, às vezes, explorem deliberadamente mapeamentos metafóricos com o propósito de criar efeitos de humor ou estilísticos.

Berber Sardinha (2007a, 2007b, 2008, 2009, 2010) aponta que o estudo da metáfora conceptual, tal como descrita em termos cognitivos, oferece grandes desafios e oportunidades para a LC, se considerado que, pelo fato de ser um fenômeno corriqueiro, da vida cotidiana, deverá estar presente nos *corpora* que compilamos. Nesse ponto, o autor se questiona acerca dos procedimentos necessários para exploração de *corpus* e extração dessas metáforas, por meio das ferramentas da LC, uma vez que depreendemos o sentido das palavras a partir do uso e que, nesse aspecto, a LC tem oferecido evidências abundantes e consistentes quanto a isso. Por outro lado, Berber Sardinha (2009, p. 43) destaca que, sendo a metáfora conceptual um fenômeno cognitivo, sua exploração por meio da LC seria um modo de "conseguir inferir o processamento mental a partir das instâncias de uso" e que essa já poderia se configurar

como uma resposta, ainda que parcial, à crítica de que a LC teria pouca capacidade de teorizar a respeito da linguagem.

No âmbito da pesquisa nacional, em português brasileiro, são diversos os trabalhos que guardam relação direta com nossa pesquisa e que servem de motivação como pontos de contraste e de comparação, tanto pelo viés empírico da exploração de metáforas conceptuais à luz da LC, quanto pelo uso de *corpora* jornalísticos. A abordagem das questões culturais implicadas também é relevante, por se tratar de países vizinhos, Brasil e Argentina, com histórias e relações sócio-políticas muito próximas. Como exemplo, Berber Sardinha (2007b) indica que as metáforas são culturais, portanto, relacionadas a determinada cultura, civilização ou ideologia, não havendo, nesse sentido, verdades absolutas para a Teoria da Metáfora Conceptual. Nesse texto (BERBER SARDINHA, 2007b, p. 168) e em referência às metáforas do presidente Lula à época, o autor destaca que ao dizer que "vamos vestir a camisa de um setor da sociedade, estamos conceituando e entendendo a sociedade como um esporte, as pessoas como jogadoras, outros grupos como adversários, a convivência como uma partida e a conduta desejada como as regras do jogo, observadas por um árbitro". Como bem observa o autor, todos esses mapeamentos ficam subentendidos pelo campo conceitual do jogo em si, pelo fato de haver um conhecimento compartilhado do que seja uma partida, talvez de futebol, e das partes implicadas. Tudo isso possibilita o entendimento da expressão como metáfora.

Também destacamos Berber Sardinha (2007a), que descreve diferentes métodos e analisa as metáforas da imprensa de modo geral, em *corpus* do jornal *Folha de São Paulo*, e as metáforas de um jornalista em particular, Joelmir Beting. As metáforas do ex-presidente Lula também se fazem presentes em outras publicações de Berber Sardinha (2008, 2010), em relação à conquista e ao desenvolvimento, respectivamente. Para tornar mais ilustrativas as inferências resultantes dos mapeamentos, denominadas *desdobramentos* na teoria, o pesquisador apresenta, no caso do jogo em equipe "Se a partida é o comércio exterior, a vitória seria o superávit nas contas dos países do bloco econômico. Uma derrota, por outro lado, seria um déficit nas contas" (BERBER SARDINHA, 2008, p. 99). Ainda mencionamos, aqui, as pesquisas de Sperandio (2009, 2010), em que a autora analisa, no primeiro trabalho, usos metafóricos em discursos do ex-presidente Lula, relacionados ao Programa *Fome Zero* e, no segundo, metáforas com relação ao Movimento dos Sem-Terra

(MST), em duas reportagens da revista brasileira *Veja* e da estadunidense *Newsweek*, ainda que sem lançar mão dos recursos e ferramentas características da LC.

Entendemos que, para além de todas as motivações e justificativas já elencadas, a possibilidade de buscar respostas quanto ao eventual caráter teórico da LC, como auxílio à compreensão do modo como metaforizamos e compreendemos as metáforas conceptuais que subjazem às metáforas linguísticas, mapeáveis a partir da identificação e descrição das UFEs com auxílio das ferramentas do WST, é um forte indício tanto do mérito dos trabalhos que recorrem à LC para a realização de pesquisas empírico-descritivas quanto do lugar já consagrado da LC. Em nossa opinião, é tão descabido atribuir um valor exclusivamente metodológico à LC quanto ainda continuar duvidando acerca do lugar de prestígio que já ocupa na área da Linguística, tanto teórico-descritiva quanto aplicada. Formulamos nossas hipóteses, a partir da ótica da LC, que corroboramos, refutamos ou reformulamos, com base na indagação de *corpora*; hipóteses que podem derivar, ainda, na formulação ou constatação de teorias.

Na próxima seção, descrevemos tanto o *corpus* quanto os procedimentos implicados neste trabalho.

3 Corpus e Metodologia

O *corpus* de estudo de nossa pesquisa, que denominamos *Corpus AleBores*, está formado por grande parte dos textos publicados na seção de opinião *Humor Político*, do jornal argentino *Clarín*, escritos em espanhol rio-platense pelo colunista e arquiteto Alejandro Borensztein, especificamente, por todos os textos aos quais foi possível ter acesso, no período entre 2009 e 2019. Tal como é mencionado em encontro *on-line* com o autor (29/05/2020), no âmbito do *Ciclo de Diálogos com Clarín,*[10] a coluna começou a ser escrita em 2007, completando em 2020 treze anos de publicações dominicais. Como aponta o autor, sua verdadeira formação universitária e profissional é em Arquitetura, por ter feito a carreira acadêmica, de graduação e pós-graduação, nessa área. A respeito do ofício de escrita semanal de um texto a ser publicado num dos jornais de maior circulação no país, Borensztein observa que surgiu de modo

[10] Informações disponíveis em: https://www.clarin.com/dialogos/entrevista_Alejandro_Borensztein. Acesso em: 20 jun. 2020.

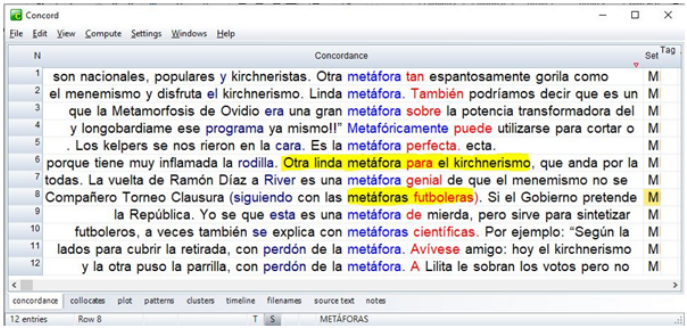
casual e que se conecta com uma experiência profissional prévia, na televisão, de aproximadamente 30 anos atrás. Com seu irmão Sebastián, reconhecido cineasta argentino, autor de filmes como *La suerte está echada* e *Un cuento chino*, entre outros, Alejandro escrevia para um programa televisivo, por indicação de seu pai, Tato Bores, humorista político renomado.

O colunista destaca, então, que o processo de escrita foi uma grande lição, porque nunca foi algo planificado em sua vida, assim como todas as demais atividades profissionais, na Arquitetura e na Televisão, para as quais sempre foi mais esquemático e ordenado. Lembra, também, que a demanda de escrita de uma matéria para ser publicada, todos os domingos, ocupando página inteira do *Clarín*, surgiu há treze anos em função de uma nota encaminhada a esse jornal, que seria uma carta de leitor e acabou sendo publicada e acarretando esse compromisso desde então, devido ao sucesso que teve.

Na reportagem, Alejandro Borensztein responde questões feitas pelos assinantes do jornal e fala a respeito dos principais personagens da coluna *Humor Político*, presidente atual ou ex-presidentes ou candidatos à presidência na Argentina, que os leitores aprenderam a acompanhar e reconhecer semanalmente, a partir destes nomes: Tío Alberto (Alberto Fernández, atual presidente); Mauri ou el Gato (Mauricio Macri), Ex-Ella, la Jefa, la Arquitecta Egipcia, Reina Hotelera, além de outros nomes (Cristina Kirchner), Compañero Lancha (Scioli, ex-candidato à presidência) e Compañero Centro Cultural, Compañero Torneo Clausura (Néstor Kirchner). Além desses nomes, sobre muitos outros é explicada a origem e evolução, como Campeonato del Pelotudo del año, el Club de los Malos, que fazem referência a manifestações ou atitudes de políticos em particular, ou de posicionamentos de partidos políticos, que muitas vezes resultam curiosas ou polêmicas e que motivam a escrita humorística da coluna dominical.

Também é feita uma alusão às metáforas presentes na escrita do colunista, ponto essencial no âmbito deste trabalho. A Figura 2 ilustra ocorrências no *corpus* que corroboram o papel das metáforas na relação entre política e futebol:

FIGURA 2 – Linhas de concordância no *corpus* de estudo com *metáfora*



Fonte: *WordSmith Tools* (SCOTT, 2016)

Dentre os principais aspectos da tipologia do *corpus* de estudo, destacamos: (a) modo escrito, em formato eletrônico; (b) contemporâneo e diacrônico, de 2009 a 2019; (c) de seleção definida pelo gênero (seção de opinião: humor político) e extensão mensurada em textos (mais de 400) e em itens (mais de 450 mil palavras); (d) de conteúdo especializado, marcado pelo campo socioprofissional (política); (e) monolíngue, na língua espanhola, em sua variedade rio-platense; (f) de autoria única, em língua nativa; e (g) para finalidade de pesquisa. A integralidade dos textos que compõem o *corpus* foi mantida. O máximo de textos publicados anualmente foi de 47. O número menor de textos coletados em alguns anos deriva de sua indisponibilidade na página do jornal. É oportuno destacar que o *corpus* foi compilado exclusivamente para fins de pesquisa, sem nenhum propósito comercial nem de disponibilização ou reprodução parcial ou integral. Os fragmentos presentes nas análises dos resultados correspondem a pequenos trechos em que são identificadas as ocorrências pertinentes para a pesquisa, que podem chegar até o nível de um parágrafo, no máximo. Para ter acesso sem registro ao jornal *Clarín*, existe uma permissão limitada a um quantitativo de textos por mês, que depende do conteúdo consumido pelo usuário, conforme consta nos termos e condições publicados no site do jornal. É necessário estar registrado, para ter acesso a um conteúdo maior de textos, e ser assinante, para ter recursos disponíveis no jornal como ouvir a leitura de notícias, copiar conteúdo, fazer comentários ou imprimir, entre outros. Embora a autoria única do *corpus* possa, certamente, trazer questões estilísticas ou a visão de um único indivíduo (objetos que não fazem parte deste trabalho),

o contraste dos resultados por meio de consulta aos *corpora* disponíveis *on-line* (DAVIES; 2016, 2018; KILGARRIFF, 2019) possibilita atestar, pela frequência, a recorrência do uso de determinadas UFEs encontradas e/ou de metáforas linguísticas que realizam metáforas conceptuais, sendo mais características de determinados países, como poderá ser observado na próxima seção.

A Tabela 1 apresenta a extensão do *corpus* de estudo e sua distribuição nos diferentes anos de publicação.

TABELA 1 – Extensão do *Corpus AleBores*

Ano	***Nº textos***	***Tokens***	***Types***	***T/T ratio***
2009-2010	08	9.698	2.824	29,12
2011	41	37.774	7.508	19,88
2012	37	35.954	7.262	20,20
2013	47	52.658	9.227	17,52
2014	47	52.022	9.215	17,71
2015	47	52.363	8.445	16,13
2016	42	46.529	8.426	18,11
2017	44	57.300	9.522	16,62
2018	47	62.471	9.785	15,66
2019	46	60.032	9.370	15,61
TOTAIS	***406***	***466.800***	***31.759***	***6,80***

Fonte: A pesquisa

A seguinte enumeração de itens procura descrever, primeiro, a sequência de etapas mais gerais desenvolvidas no âmbito desta pesquisa. Em segundo lugar, adentramos de modo mais específico nos procedimentos envolvidos na compilação, preparação e armazenamento do *corpus* de estudo, para seu tratamento com as ferramentas utilizadas no auxílio à extração de dados, à descrição e às correspondentes análises.

1) Planejamento das características do *corpus* compilado, considerando fatores como extensão e representatividade;
2) Definição de critérios e códigos para compilação e armazenamento do *corpus* de estudo;

3) Compilação, preparação e armazenamento do *corpus*;
4) Anotação de metadados e elementos histórico-contextuais referentes aos textos compilados para constituição do *corpus*;
5) Levantamento dos dados estatísticos do *corpus*;
6) Extração das palavras-chave do *corpus* de estudo, relacionadas aos diferentes campos lexicais e semânticos do futebol e da política;
7) Identificação e extração dos candidatos a termo, no domínio do futebol;
8) Identificação, extração e descrição das unidades fraseológicas especializadas do âmbito do futebol, na metaforização do domínio da política, a partir da análise das linhas de concordância e dos subsídios (colocados, clusters) que oferecem as ferramentas e utilitários dos programas computacionais;
9) Análise e descrição das relações metafóricas, à luz das UFEs extraídas, entre os domínios do futebol e da política;
10) Classificação e sistematização dos resultados, a partir da consulta tanto a dicionários gerais e especializados das línguas espanhola e portuguesa em uso, quanto a recursos *on-line* de consulta, como o *Corpus del Español* e *do Português* (DAVIES, 2016, 2018) e *Sketch Engine* (KILGARRIFF, 2019).

Os procedimentos envolvidos na compilação, preparação e armazenamento do *Corpus AleBores* partiram do acesso à página do colunista.[11] Abrimos individualmente cada uma das publicações disponíveis de *Humor Político*, selecionamos o texto, para copiar e colar seu conteúdo integral em arquivos individuais em formato TXT. Esse foi o procedimento que se mostrou mais eficaz, para evitar que fossem transportados ao *corpus* informações relacionadas a imagens ou *links*, que demandariam uma posterior limpeza dos arquivos.

Para a nomeação de cada texto do *corpus,* utilizamos uma sequência de 8 números, seguindo a sequência conforme data de publicação: ano, mês, dia (exemplo, 20141012). Os títulos dos textos foram etiquetados por meio das *tags* <T> e </T>, no intuito de separar título de conteúdo textual e para a recuperação de informações, durante as buscas com os programas do WST. Os textos foram armazenados em pastas no Explorador de arquivos do Windows, organizados por anos de

[11] Disponível em: https://www.clarin.com/autor/alejandro-borensztein.html. Acesso em: 17 jun. 2020.

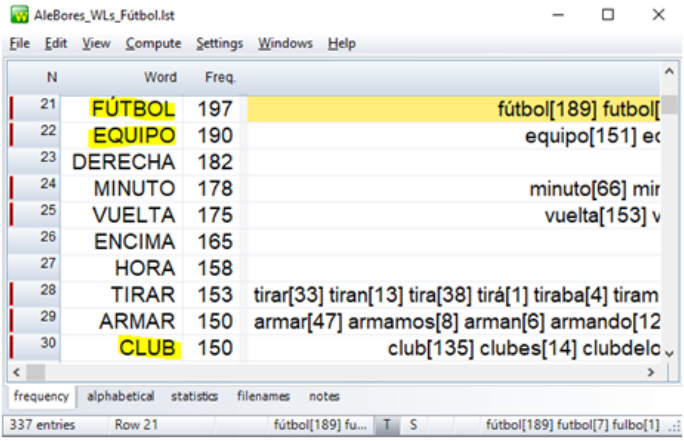
publicação e compartilhados no *OneDrive* entre os autores da pesquisa, para acompanhamento e trabalho conjunto.

Com o *corpus* compilado, preparado e armazenado, o passo seguinte foi a extração dos dados quantitativos mais gerais, primeiro, por meio da ferramenta *WordList*, como apresentado na Tabela 1. Geramos listas de palavras para cada um dos *subcorpora*, a partir dos anos de publicações coletadas e, também, uma lista de palavras com *corpus* geral. Para extração das palavras-chave, por meio da ferramenta *KeyWords*, contrastamos a lista de palavras do *Corpus AleBores*, nosso *corpus* de estudo, com a lista de palavras de um *corpus* de referência compilado no âmbito de uma pesquisa de Iniciação Científica que orientamos (ALVES, 2013). Esse *corpus* é formado pelas publicações de 6 seis *Congresos Internacionales de la Lengua Española* (1992 – 2010), portanto, correspondente ao gênero de escrita acadêmica. O *corpus* de referência possui uma extensão de 2.834.385 *tokens* e 95.649 *types*, em 813 textos, representando em torno de 6 vezes mais do que o tamanho do *corpus* de estudo, em extensão pelo número de palavras. Conforme a literatura da LC (BERBER SARDINHA, 2004, 2009), o *corpus* de referência deve ser em extensão, no mínimo, cinco vezes maior do que o *corpus* de estudo. A diferença quanto ao gênero do *corpus* de referência busca, fundamentalmente, trazer à tona o que é proeminente no *corpus* de estudo, no intuito de poder apontar para temáticas e campos lexicais em destaque. Um *corpus* de referência com características similares às do *corpus* de estudo, em termos de extensão, autoria ou gênero, tenderia a neutralizar os resultados. Considerando a representatividade do *corpus* de estudo, situado no âmbito da trama política argentina e registrada em tom humorístico numa coluna jornalística, o *corpus* de referência de escrita acadêmica utilizado, por sua extensão e diferença temática, é um ponto de comparação que se ajustou às necessidades da pesquisa.

Realizamos diferentes tipos de testes, aplicados à extração de palavras-chave, por meio de diversos ajustes na configuração da ferramenta *KeyWords*, especificamente no *p=value*, de p=0,05 a p=0,0000001, e no *max. wanted* (8.500) e *min. freq.* (3), no intuito de conferir as respectivas diferenças e de corroborar a maior ou menor presença dos itens que formariam uma lista de candidatos a termo, no domínio do futebol. Desse modo, obtivemos diferentes resultados, que variaram entre 226 e 8.161 vocábulos. Após leitura, análise e limpeza dessas primeiras listas (exclusão de todos os vocábulos não pertencentes

ao domínio do futebol) e, paralelamente, pela comparação com os resultados da lista de palavras do *corpus* de estudo geral, pouco mais de 31 mil *tokens*, foi possível constatar que as listas de palavras-chave extraídas por meio da ferramenta *KeyWords*, independentemente dos diversos tipos de ajustes realizados na configuração, não reportava a totalidade dos vocábulos candidatos a termo. Isto é, itens com baixa frequência de ocorrência no *corpus* de estudo ou cujos resultados teriam sido neutralizados pelo *corpus* de referência, simplesmente não foram identificados pela ferramenta, apesar de não deixarem de estar presentes no *corpus* e de fazerem parte do objeto de estudo desta pesquisa. Nesse sentido, decidimos realizar a análise, limpeza e extração dos candidatos a termo diretamente da lista de palavras gerada com a ferramenta *WordList*. Com a ferramenta *KeyWords*, obtivemos no resultado mais produtivo 226 palavras-chave, que resultaram em 134 candidatos a termo, após lematização. Já a análise com a ferramenta *WordList* possibilitou a extração de 337 candidatos a termo, após realização dos processos de limpeza e lematização, como se observa na Figura 3, que ilustra a alta frequência de vocábulos do domínio do futebol no *corpus*.

FIGURA 3 – Lista de candidatos a termo lematizada



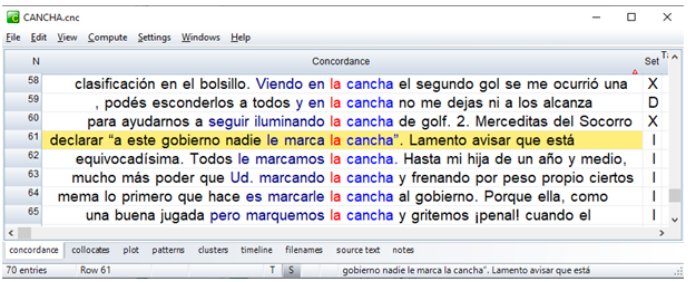
Fonte: *WordSmith Tools* (SCOTT, 2016)

Foi oportuno apreciar, durante os diversos testes, que muitos itens de ocorrência 1, *hapax legomena*, fazem parte da temática em estudo e apresentam não apenas combinatórias léxicas, como também

UFEs e processos de metaforização, relevantes para esta investigação. Assim, salvamos uma lista com os *hapax* do *Corpus AleBores* que, após o procedimento de lematização, ficaram reduzidos a 19, entre os candidatos a termo. É relevante destacar, ainda, que os nomes próprios e apelidos de clubes, jogadores, técnicos, estádios, torneios e associações de futebol, entre outros, foram salvos em lista separada, uma vez que não serão analisados como base, na formação de combinatórias léxicas ou de UFEs. A lista com nomes próprios do âmbito do futebol registrou 70 vocábulos, alguns com frequência elevada, a saber: Boca (213), River (86), Bombonera (46), Libertadores (33) e Messi (23).

A proposta de organização dos candidatos a termos do domínio do futebol em campos lexicais e semânticos, a partir dos resultados das palavras-chave e, na sequência, tomando por base a lista de palavras resultante após limpeza e lematização, foi um procedimento prévio e auxiliar à identificação das UFEs. Pensamos inicialmente em 10 campos, para organização e classificação dos resultados, após análise dos vocábulos e identificação dos candidatos a termos. Nessa etapa, a ferramenta *Concord* é de vital importância, como pode ser observado na Figura 4. A partir da busca por um item lexical específico, neste caso *cancha* (f 79 / UT 70), uma vez identificado como candidato a termo, geramos as linhas de concordância que trazem tantas linhas quantas ocorrências do termo houver no *corpus*.

FIGURA 4 – Linhas de concordância com *cancha*



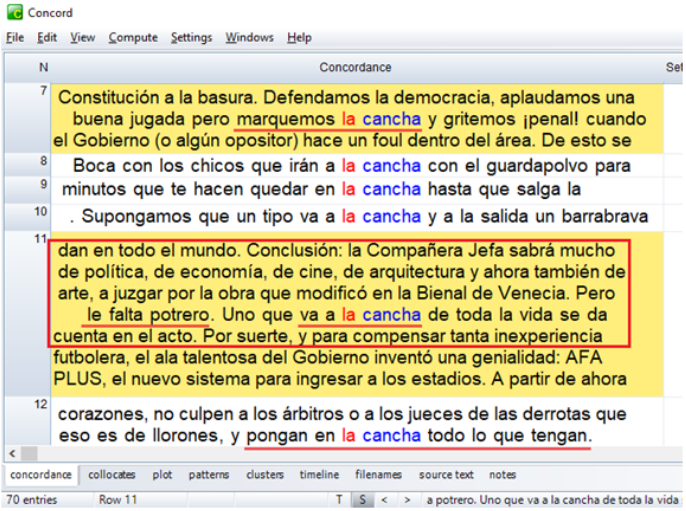
Fonte: *WordSmith Tools* (SCOTT, 2016)

É pela análise das ocorrências dos candidatos a termo em contexto, por meio das linhas de concordância, que se pode chegar à constatação de que um item lexical funciona como um termo, em determinado ambiente textual. Assim, a partir de cada um dos itens lexicais da lista, foram

geradas as linhas de concordância correspondentes, os resultados foram analisados e os arquivos foram salvos em pasta específica, identificados pelo nome do item de busca. Na ferramenta *Concord,* utilizamos a coluna *Set* como um recurso auxiliar na classificação e organização dos resultados (ver figura anterior). Por meio da inserção de um caractere (letra ou número), agrupamos os padrões identificados e separamos as ocorrências que não correspondessem a nosso objeto de estudo, quando os domínios implicados não fossem o do futebol e da política.

Procedimento análogo adotamos para a extração das combinatórias léxicas e das UFEs em que participam os termos. Nesse passo, definimos que os segmentos textuais extraídos deveriam contemplar as referências denotativas ao âmbito futebolístico, mas em relação metafórica com o domínio da política, para posterior análise dos eventuais processos de metaforização. A Figura 5 ilustra o recurso de ampliação dos contextos de ocorrência, na análise das linhas de concordância, para seleção das UFEs.

FIGURA 5 – Processo de identificação de UFEs em relação metafórica



Fonte: *WordSmith Tools* (SCOTT, 2016)

Na figura anterior, é possível apreciar a relação entre política e futebol, nos fragmentos em destaque. Assim, nas linhas de concordância 7 e 11, apenas para ilustrar o procedimento de identificação das UFEs

em relação metafórica,[12] percebemos que o assunto é a democracia ou o governo e que há alusões diretas à prática do futebol, a saber: *aplaudir una buena jugada, marcar la cancha, gritar penal, hacer un foul dentro del área, faltarle potrero, ir a la cancha, inexperiencia futbolera, ingresar a los estadios, poner todo en la cancha*.

Os campos lexicais e consequentes campos semânticos de determinados vocábulos identificados no *corpus*, com pertinência ao domínio da prática do futebol e em relação metafórica com a política, estão resumidos, a seguir: Participantes (jogadores, posições no campo de jogo, técnicos, árbitros, dirigentes, torcida etc.); Locais para a prática do futebol (campo de jogo e suas partes específicas, de treino, estádios etc.); Jogadas (ataque, defesa, estratégias de jogo...); Partidas (amistosas, oficiais, duração, campeonatos, transmissão...); Partes do corpo humano (que intervêm no jogo, pê, cabeça, peito, mãos...); Materiais esportivos utilizados na prática de futebol (camisa, chuteira, caneleira...). Estes campos lexicais buscam agrupar, num primeiro nível, as ocorrências dos candidatos a termos extraídos. Por meio da análise das linhas de concordância que geramos a partir de cada termo, pudemos apreciar em contexto a abrangência semântica de cada ocorrência, conforme o reconhecimento das diferentes UFEs. O mesmo procedimento foi aplicado para a limpeza dos resultados, isto é, para exclusão tanto dos candidatos que não se confirmassem enquanto termos, como das ocorrências de UT que não correspondessem ao domínio do futebol, mas de outros esportes. Assim, como será observado na próxima seção, embora a UT *cancha* tenha reportado determinada frequência, nem todas foram correspondentes ao âmbito futebolístico.

Para o tratamento e sistematização dos dados em planilha Excel, incluímos um número para a entrada de cada item lexical candidato a termo, conforme ordem alfabética, a frequência de ocorrência no *corpus*, a frequência atestada enquanto UT, o quantitativo de tipos de UFEs identificadas para cada UT, a descrição das combinatórias léxicas e dos fraseologismos identificados e os contextos metafóricos de ocorrência.

A próxima seção apresenta as análises das UFEs formadas com as UTs *cancha* e *canchereada*, identificadas no *corpus* de estudo na realização de processos de metaforização.

[12] As análises dos fragmentos que compõem o recorte da pesquisa apresentado neste artigo serão desenvolvidas na próxima seção.

4 Análise

Dos dados obtidos e classificados sob o campo semântico Locais para a prática do futebol, selecionamos *cancha* e um de seus derivados para esta análise. Pela análise das linhas de concordância, o item lexical *cancha* reportou 79 ocorrências no *Corpus AleBores*, das quais 70 foram corroboradas como UTs, no escopo de nossa pesquisa, pelo fato de fazerem referência específica ao futebol. As demais dizem respeito a campos de Tênis, Rugby, Polo ou Golfe, esportes também populares na Argentina.

CANCHA. ***f.*** Terreno utilizado para a prática de alguns esportes como o futebol. Por extensão, estádio. **(f 70)**

Pela análise das combinatórias léxicas com *cancha*, identificamos diversas UFEs com sentido metafórico, como *entrar a la cancha, ir a la cancha, dejar / poner (todo) en la cancha, sacar a la cancha, sacar de la cancha, salir a la cancha*, além de outras, em que os participantes envolvidos são personagens da política ou do governo, mas que serão objeto de estudo em outras publicações. Neste artigo, passamos a analisar as ocorrências identificadas com MARCAR como colocado de CANCHA. Observamos 6 ocorrências da colocação MARCAR + CANCHA, que representam 8,6% da frequência da UT (70) no *corpus*. Não encontramos, em nenhum dos dicionários consultados, a combinatória léxica dicionarizada. Num primeiro momento, definiremos a colocação em sentido denotativo, no âmbito do futebol, para depois analisá-la à luz dos segmentos de ocorrência no *corpus*.

MARCAR(LE) LA CANCHA [uma pessoa a outra] **/ ESTAR LA CANCHA MARCADA**. 1. Em campos com gramado natural, marcar com cal as linhas laterais, finais, do meio do campo, além das áreas (grande, pequena e semicírculo) dos gols, ângulos de escanteios e pontos para cobrança de pênalti em cada área e desde onde são iniciados os jogos, no meio do campo. 2. Definir uma estratégia de jogo e/ou um posicionamento de jogadores em campo, que dificultam as ações do adversário. (**f 6**)

(01) *No permita que ninguna Diana Conti le amargue la vida cuando sugiere que hay que tirar la Constitución a la basura. Defendamos la democracia, aplaudamos una buena jugada, pero* ***marquemos la cancha*** *y gritemos ¡penal! cuando el Gobierno (o algún opositor) hace un foul dentro del área.*

(02) *La Compañera Jefa acaba de declarar "a este gobierno* ***nadie le marca la cancha****". Lamento avisar que está equivocadísima.*

(03) ***Todos le marcamos la cancha****.*

(04) *Hasta mi hija de un año y medio, mi Minina!!!, en cuanto se despierta y antes de tomar su mema lo primero que hace es* ***marcarle la cancha*** *al gobierno.*

(05) *Porque ella, como todos nosotros, nació en suelo argentino, bajo la tutela de un librito que se llama Constitución Nacional donde está toda* ***la cancha marcadita****. Les guste o no les guste.*

(06) *En el momento más duro de su gobierno, después de aprobar a la fuerza el acuerdo con Irán, les salió una bolilla impensada: Francisco. Nadie se avivó que en la esquina de la Casa Rosada vivía un potencial Papa (de hecho, había salido segundo en la votación anterior), y de un día para el otro apareció un argentino con mucho más poder que Ud.* ***marcando la cancha*** *y frenando por peso propio ciertos aires de descontrol.*

A compreensão das UFEs, formadas a partir de *marcar la cancha* que identificamos no *corpus*, demandou um conhecimento prévio quanto ao sentido das marcações já existentes no campo de futebol e suas implicações. As marcas, feitas normalmente de cal, delimitam as linhas laterais e finais do campo de jogo. A linha central do meio de campo separa os setores de ataque e de defesa de cada um dos times. Dentre as demais marcações, essas já bastariam para entender que cada time tem, no jogo, diversas possibilidades de distribuir seus jogadores, tanto no ataque quanto na defesa, a partir de estratégias e esquemas ofensivos e defensivos. A UFE *marcar(le) la cancha* equivale a mostrar para o rival o modo como poderá ou não jogar, que não vai poder jogar da maneira que quiser, pois as marcações no campo estabelecem limites, assim como as estratégias de seu adversário. No fórum de discussão do dicionário *WordReference*, há uma consulta sobre o significado da frase. Entre as respostas, consta em inglês "Marcar la cancha means to put the limits to (a situation)".[13]

[13] Informação encontrada em: https://forum.wordreference.com/threads/est%C3%A1-marcando-la-cancha.2324051/. Acesso em: 16 jun. 2020.

Metaforicamente, transportado ao domínio alvo da política, esse fraseologismo futebolístico aponta para o estabelecimento de limites aos governantes, na época Cristina Kirchner. Assim, ao afirmar em (2) "a este gobierno nadie **le marca la cancha**", a presidente manifestava que ninguém iria definir limites a seu governo, que ninguém iria marcar para eles como deveriam ser as coisas. Como tópico da metáfora linguística temos "a este gobierno" e como veículo "nadie le marca la cancha". Com isso, inferimos que o governo é representado como um time de futebol pelo uso que a mandatária faz dessa colocação, e que escolhe o modo como quer jogar e não aceita que ninguém lhe imponha regras. *Cancha*, neste caso, é o terreno de ação governamental. Borensztein, colunista de *Humor Político*, reage a essa afirmação, destacando em (5) que todo aquele que habita sob solo argentino está regido pela Constituição Nacional, "donde está **toda la cancha marcadita**". Essa réplica dá a entender que a expressão destaca o modo como as coisas devem ser. Assim, as leis que estabelecem as normas de convivência no território argentino, definidas na Constituição (tópico), são um campo de futebol (veículo), com regras bem definidas que devem ser respeitadas.

Em outra referência à Constituição Nacional, o articulista aconselha o leitor a não se deixar amargurar a vida, quando uma ex-deputada kirchnerista teria sugerido jogar no lixo a Constituição, e complementa em (1) dizendo "Defendamos la democracia, aplaudamos una buena jugada, pero **marquemos la cancha** y gritemos ¡penal! cuando el Gobierno (o algún opositor) hace un foul dentro del área". Ou seja, gritar pênalti, se o governo ou a oposição, enquanto tópico, fizerem falta dentro da área, isto é, em caso de cometerem uma falta grave passível de punição, significa fazer valer as regras, isso é **marcar la cancha**, como veículo da metáfora linguística. Novamente, cometer uma infração no jogo acarreta uma punição; o uso de **marcar la cancha** corresponde a estabelecer limites com consequências previsíveis para quem infrinja as normas. Assim, o campo de ação do governo é um campo de futebol, *una cancha*, com regras marcadas.

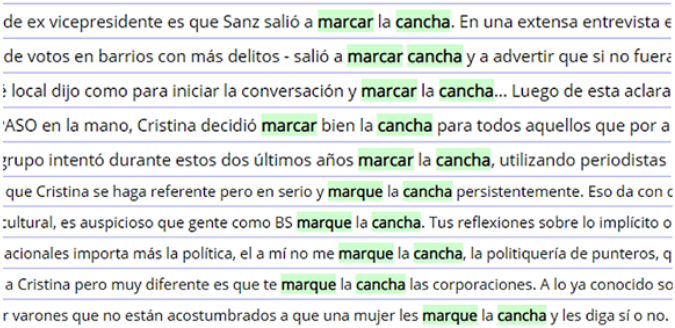
No fragmento (6), a UFE no *corpus* está relacionada ao aparecimento da figura do Papa Francisco, que surgiu **"marcando la cancha"** e com muito mais poder do que a presidente Cristina Kirchner. O colunista destaca que, embora o governo não cogitasse essa possibilidade, teria que passar a lidar com um Papa argentino, que passaria a ter ingerência quanto ao respeito às regras do jogo, no alcance das ações do

governo. Desse modo, podemos apreciar que o uso de *marcar la cancha* se aplica, por um lado, no sentido de quem não quer que ninguém marque limites a seu governo e, por outro lado, de que toda a sociedade, o Papa e a própria Constituição estabelecem marcas, definem limites, que precisam ser respeitados dentro do jogo.

Nas ocorrências (1) e (3), o articulista faz um apelo ao leitor, marcado pelo uso da primeira pessoa do plural no Imperativo afirmativo e no Presente do modo Indicativo: "aplaudamos una buena jugada pero marquemos la cancha y gritemos ¡penal!" e em "Todos le marcamos la cancha". Esses usos podem ser entendidos como referência à torcida de um time imaginário, representando o povo argentino, que tanto alentaria as boas ações dos governantes quanto questionaria a falta de limites ou suas atitudes agressivas, violentes. Desse modo, o autor da coluna poderia ser compreendido como membro ou líder de uma torcida, da qual fazem parte o próprio colunista e os leitores. Palacios e Chacra (2014, p. 83) afirmam que "O torcedor argentino usa com frequência a primeira pessoa do plural para explicar como anda seu time: *estamos indo bem* ou *hoje à tarde jogaremos contra X*". Isto é, o torcedor se enxerga como parte integrante e ativa do time, não apenas como espectador.

No intuito de verificarmos a incidência de *marcar la cancha* fora de nosso *Corpus AleBores*, por meio da busca no *Corpus del Español*, nas versões *Dialetal* e *NOW* (DAVIES, 2016, 2018), utilizando *cancha* como base do fraseologismo e *marcar* como colocado, com a função de lematização ativada num horizonte de 4 palavras à esquerda e 4 à direita, obtivemos 283 resultados na versão dialetal desses *corpora*. Cabe destacar que esses *corpora* possuem uma extensão de 2 e 5,5 bilhões de palavras, respectivamente. Como *corpus* de consulta e ponto de contraste, entre esses e o *Sketch Engine* (KILGARRIFF, 2019), o *Corpus del Español* na versão dialetal (DAVIES, 2016) foi o que se mostrou mais pertinente dos três para esta pesquisa. O país que registrou maior frequência de ocorrência foi Argentina (122), seguido pelo Uruguai (48). Os demais resultados ficaram distribuídos entre os outros países hispano-falantes, com baixa frequência, como Espanha (5) e México (4), por exemplo. Pela frequência atestada, a análise dos resultados separados por países permitiu constatar que a UFE estudada corresponde a um uso próprio rio-platense. Analisando os resultados, de modo geral, a UFE ocorreu em frases que tratam sobre política e justiça, envolvendo membros de partidos políticos, governantes, juízes, ministros da Corte Suprema, legisladores, entre outros, como pode ser apreciado na próxima figura.

FIGURA 6 – Busca por MARCAR + CANCHA



Fonte: *Corpus del Español* (DAVIES, 2016)

Dentre os derivados de *cancha* identificados no *corpus*, extraímos diversos vocábulos, alguns dos quais se configuraram UTs, como *canchero* e *cancherear*, que denotam uma ostentação de determinada habilidade (*tener cancha*), confiança, no desempenho de uma atividade, por exemplo no controle da bola, no jogo do futebol. Além dessas, também registramos o substantivo *canchereada*, que passamos a definir e para o qual tampouco encontramos registro nos dicionários consultados.

CANCHEREADA. ***f.*** Jogada feita com malandragem, com a esperteza de quem demonstra possuir habilidade, domínio, técnica ou experiência, para conseguir que uma manobra ilegal ou contra as regras (no jogo, um gol de mão, por exemplo) passe como totalmente válida. (**f 4**)

(7) *Obviamente, Macri lo vetó. ¿Está loco? No, al igual que CFK en su momento, ahora es él quien no tiene la guita para* ***financiar esta canchereada****.*

(8) *Otro ejemplo, es el bono a 100 años,* ***una canchereada del gobierno*** *para decir "mirá que confianza que nos tienen".*

(9) *Con el dólar a 15,* ***la canchereada de la arquitecta egipcia*** *nos costó 60.000 millones de pesos más.*

(10) *El tipo le mandó a la diputada Cerruti un papelito de morondanga pidiéndole simpáticamente que no meta a sus hijas menores de edad en el tema de las famosas offshores.* ***Una canchereada de pescador****. O sea, sacó la caña y empezó suavecito con el lengue lengue.*

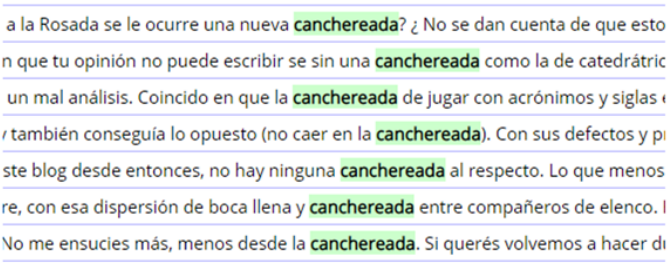
Nos quatro fragmentos anteriores, observa-se que se trata de manobras realizadas por governantes, jogadas políticas, levadas à prática com a habilidade e malandragem ou catimba de um jogador de futebol em campo, que utilizaria artimanhas para conseguir enganar o adversário, seja para fazer um gol, para fazer um drible ou tomar a bola, muitas das vezes de maneira desleal. Assim, em (7) observamos uma situação em que os ex-presidentes argentinos, CFK e Macri, por falta de capital (*guita*), precisaram recorrer a manobras que poderiam ser questionáveis, devido à ginga utilizada para esquivar um mau momento. Em (8) e (9), percebe-se que as malandragens são atribuídas ao governo diretamente ou por meio de um dos apelidos utilizados pelo colunista em referência a Cristina Kirchner, *arquitecta egipcia*. Em ambos os fragmentos há novamente uma relação com questões econômicas que, lançando mão de outras metáforas futebolísticas para explicá-las, tentam ser dribladas pelos governantes, por meio de jogadas que poderiam ser consideradas desleais. Também em (10), quando a referência à participação de familiares de políticos em empresas "*offshores*" é feita como "una canchereada de pescador", isto é, assim como *uma história de pescador*, em que também haveria uma jogada, por meio da qual se buscaria alcançar algum benefício por caminhos duvidosos. Nos quatro fragmentos em análise, os personagens do domínio político são o tópico da metáfora linguística, enquanto *canchereada* funciona como veículo, que transfere traços concretos do plano do futebol para o domínio mais abstrato da política.

Na versão dialetal do *Corpus del Español* (DAVIES, 2016), a busca lematizada de *canchereada* reportou 33 ocorrências do termo, sendo 20 no singular e 13 no plural, e 30 dessas ocorrências da Argentina, o que confirma tratar-se de um argentinismo. Num dos fragmentos encontrados nesse *corpus*, identificamos a explicação do que seria uma **canchereada**, dada por um jogador: "Bajarla [la pelota] con una mano se podía confundir con una **canchereada** también. Sí, se puede confundir, yo hacía cosas que parecían **canchereadas** pero eran recursos de una técnica futbolística que yo noté que ya la iba teniendo de chico".[14] Desse modo, é possível ver que estratégias ou manobras políticas são jogadas de futebol, feitas também com malandragem, como verdadeiras **canchereadas**. A

[14] Disponível em: http://www.elgrafico.com.ar/2012/04/25/C-4186-amadeo-carrizo-aun-no-concibo-a-river-en-la-b-llore-mucho-con-el-descenso.php. Acesso em: 17 mai. 2020.

próxima figura ilustra um recorte dos resultados encontrados no *Corpus del Español*, na versão dialetal.

FIGURA 7 – Busca por CANCHEREADA



Fonte: *Corpus del Español* (DAVIES, 2016)

5 Considerações

Encerrando este trabalho, depois da fundamentação teórica, descrição metodológica e das análises feitas a partir de um recorte de nossa pesquisa, fundamentalmente a partir do termo *cancha*, enquanto local para a prática de futebol, podemos traçar algumas considerações.

A pertinência a esta pesquisa dos pressupostos teóricos sobre os quais discorremos, na abordagem da Terminologia, em especial das UFEs, da Teoria da Metáfora Conceptual e da LC, ficou demonstrada pelo estabelecimento de diversos pontos de convergência. A recorrência de termos do domínio fonte mais concreto do futebol, em referência ao domínio alvo mais abstrato da política, a presença de um núcleo enquanto UT nas UFEs, a formação das metáforas linguísticas com tópicos do âmbito político e veículos do meio do futebol, tudo mediado pela exploração empírica de um *corpus* com os recursos, ferramentas e princípios de pesquisa da LC, dão conta desse ponto de confluência dessas diversas vertentes teóricas. Por outro lado, a presença da cultura do futebol na vida diária da sociedade, observada nos trabalhos brasileiros revistos assim como nesta pesquisa, provavelmente fazendo parte em alguma medida do vocabulário popular, acaba sendo invocada por meio de expressões típicas desse domínio, tanto pela imprensa quanto no meio político, com diversos propósitos, mas funcionando para metaforizar situações abstratas em diversos âmbitos.

Após a compilação do *Corpus AleBores*, preparação, armazenamento e tratamento com as diferentes ferramentas do programa WST, por meio do levantamento dos dados, analisamos diferentes listas de palavras e de palavras-chave, tomando por base diferentes critérios e valores adotados, no intuito de fazer a extração terminológica do domínio do futebol. Para isso, foi necessário realizar diferentes testes com o *corpus* de estudo, em contraste com um *corpus* de referência. O procedimento que se mostrou mais eficaz, isto é, que reportou o maior número de candidatos a termo, foi o escrutínio cuidadoso da lista de palavras extraída pela *WordList*.

Como descrito nas seções anteriores, a partir dos candidatos a termo geramos as linhas de concordância e, pela análise das ocorrências em contexto, identificamos as formações fraseológicas com participação dos termos. Desse modo, foi possível extrair UFEs, os fragmentos de ocorrência e descrever a formação das diferentes combinatórias léxicas, assim como a relação estabelecida nos processos de metaforização identificados e analisados. Todos os procedimentos foram detalhados do modo pormenorizado, uma vez que a replicação metodológica é um fator presente e de relevância nas pesquisas que envolvem a LC e a Terminologia. Cabe também destacar que a LC viabilizou neste trabalho mais do que uma sequência de procedimentos metodológicos ou de um modo de olhar para os dados, uma vez que aliada ao auxílio na identificação dos fatos linguísticos, observados na estrutura aparente do *corpus*, nossa introspecção foi conduzida à percepção das ocorrências metafóricas. O mapeamento da subjacência das metáforas conceptuais, enquanto fenômeno cognitivo materializado em metáforas linguísticas, a partir da exploração e indagação de *corpora*, são indícios significativos sobre o poder de teorização da LC a respeito da linguagem. Desta maneira, os trabalhos empíricos da LC contribuem para o avanço das pesquisas em metáfora conceptual e, num plano mais amplo, na Linguística Cognitiva.

Tal como discutido na seção de Análise, identificamos relações metafóricas entre os domínios do futebol e da política, que corroboram nossa hipótese inicial: existem aspectos cognitivos, linguísticos e pragmáticos entrelaçados e englobados por uma dimensão cultural mais ampla, que perpassa os domínios do futebol e da política e que somente é acionada a partir de traços e vestígios identificados no processo de leitura do texto. O reconhecimento de determinadas marcas linguísticas

nos textos, especificamente as UFEs que acionam nas lembranças do leitor o domínio do futebol, transferem características dessa área, que passam a ser assimiladas para a compreensão do outro domínio que é metaforizado, o campo da política. Nesse sentido, por meio destas análises, conseguimos identificar as seguintes metáforas conceptuais no *corpus*, que se mostraram abundantes e recorrentes: o terreno das ações políticas é um campo de futebol; marcar o campo de jogo é estabelecer limites no âmbito da política; política é futebol; políticos são jogadores; manobras ou estratégias políticas são jogadas de futebol, inclusive malandragens (**canchereadas)**.

Declaração de contribuição de cada autor

Enquanto autores, declaramos que desenvolvemos todos os trabalhos pertinentes a este artigo conjuntamente, desde a concepção e recorte, a partir de uma pesquisa maior em desenvolvimento, passando por todas as etapas de redação da fundamentação teórica, descrição metodológica, análises do *corpus*, até as considerações e revisões que se fizeram necessárias. Para facilitar os trabalhos, os autores compartilhamos uma pasta no *OneDrive*, contendo o texto do artigo, parte da bibliografia utilizada e arquivos com resultados do WST.

Referências

ALVES, M. *A representação do Brasil no ensino de espanhol*: um estudo diacrônico baseado em *corpus* de textos acadêmicos. 2013. Relatório (Iniciação Científica) – Instituto de Letras e Linguística da Universidade Federal de Uberlândia, Uberlândia, 2013.

ANANÍA, P. *Diccionario inmoral de los argentinos*. Buenos Aires: Vergara, 2005.

BERBER SARDINHA, T. As metáforas do presidente Lula na perspectiva da Linguística de Corpus: O caso do Desenvolvimento. *D.E.L.T.A.,* São Paulo, v. 26, n. 1, p. 163-190, 2010. DOI: https://doi.org/10.1590/S0102-44502010000100007

BERBER SARDINHA, T. *Pesquisa em Lingüística de Corpus com WordSmith Tools*. Campinas: Mercado das Letras, 2009.

BERBER SARDINHA, T. Lula e a metáfora da conquista. *Linguagem em (dis)curso,* Tubarão, n. 8, v. 1, p. 93-120, 2008. DOI: https://doi.org/10.1590/S1518-76322008000100005

BERBER SARDINHA, T. *Metáfora.* São Paulo: Parábola Editorial, 2007a.

BERBER SARDINHA, T. Análise de metáfora em *corpora*. *Ilha do Desterro*, Florianópolis, n. 52, p. 167-199, 2007b.

BERBER SARDINHA, T. *Lingüística de corpus.* Barueri: Manole, 2004.

BEVILACQUA, C. R. *Unidades Fraseológicas Especializadas Eventivas:* descripción y reglas de formación en el ámbito de la energía solar. 2004. 243f. Tese (Doctorado en Lingüística Aplicada) - Institut Universitari de Lingüística Aplicada, Universidad Pompeu Fabra, 2004.

BEVILACQUA, C. R. *Unidades Fraseológicas Especializadas: estado de la cuestión.* Trabajo de investigación. Barcelona: Institut Universitari de Lingüística Aplicada, 1999.

BEVILACQUA, C. R. Unidades Fraseológicas Especializadas: novas perspectivas para sua identificação e tratamento. *Organon,* Porto Alegre, n. 26, p. 1-8, 1998. DOI: https://doi.org/10.22456/2238-8915.29562

BORBA, F. S. *Dicionário de usos do português do Brasil.* São Paulo: Ed. Ática, 2002.

BORENSZTEIN, A. Barcarce, We Have a Problem. *Clarín*, Buenos Aires, 22 abr. 2018.

BOSQUE, I.; DEMONTE, V. *Gramática Descriptiva de la Lengua Española.* 2. ed. Madrid: Espasa Calpe, 1999. Tomos 1, 2 e 3.

CABRÉ, M. T. *La Terminología*: Representación y Comunicación. Barcelona: IULA / Universitat Pompeu Fabra, 2005.

CABRÉ, M. T. Terminología y Lingüística: la teoría de las puertas abiertas. *Estudios de Lingüística del Español (ELiEs),* Barcelona*,* v. 16, [s.p.], 2002.

CABRÉ, M. T. *La terminología:* teoría, metodología, aplicaciones. Traducción castellana: Carles Tebé. Barcelona: Editorial Empúries, 1993.

CABRÉ, M. T.; ESTOPÀ, R.; LORENTE, M. Terminología y Fraseología. *In*: SIMPOSIO DE TERMINOLOGÍA IBEROAMERICANA, V., 1996, Ciudad de México. *Anais* […]. Ciudad de México: Red Iberoamericana de Terminología, 1996. p. 1-23.

CORPAS PASTOR, G. *Diez años de investigación en fraseología:* análisis sintáctico-semánticos, contrastivos y traductológicos. Madrid: Iberoamericana, 2010.

DAVIES, M. *Corpus del Español:* Web/Dialectos, 2016. Disponível em: https://www.corpusdelespanol.org/web-dial/. Acesso em: 25 set. 2019.

DAVIES, M. *Corpus del Español:* NOW, 2018. Disponível em: https://www.corpusdelespanol.org/now/. Acesso em: 15 ago. 2019.

DEIGNAN, A. A gramática das metáforas linguísticas. *In*: SHEPHERD, T. M. G.; BERBER SARDINHA, T; VEIRANO PINTO, M. (org.). *Caminhos da Linguística de Corpus.* Campinas: Mercado de Letras, 2012. p. 65-86.

DEIGNAN, A. *Metaphor and Corpus Linguistics*. Amsterdam-/ Philadelphia: John Benjamins Publishing, 2005. DOI: https://doi.org/10.1075/celcr.6

FONTANARROSA, R; SANZ, T. *El fútbol argentino*: pequeño diccionario ilustrado. Buenos Aires: Clarín/Aguilar, 1994.

GOVERNATORI, G.; LAROCCA, R. ¡Qué lo parió, che!: Diccionario coloquial de los argentinos. Buenos Aires: Continente, 2014.

HOUAISS, A. *Dicionário eletrônico Houaiss da língua portuguesa.* Versão 3.0, 2009.

KILGARRIFF, A. *Sketch Engine.* Disponível em: http://sketchengine.co.uk/. Acesso em: 3 nov. 2019.

KRIEGER, M. G.; SANTIAGO, M. S.; CABRÉ, M. T. Terminologia em foco: uma entrevista comentada com Maria Teresa Cabré. *Calidoscópio,* São Leopoldo, RS, v. 11, n. 3, p. 328-332, 2013. DOI: https://doi.org/10.4013/cld.2013.113.11

LAKOFF, G; JOHNSON, M. *Metaphors we live by.* Chicago: The University of Chicago Press, 1980.

MOLINER, M. *Diccionario de uso del español.* Edición electrónica, versión, 3.0. Madrid: Editorial Gredos, S.A.U., 2008.

ORENHA, A. *Unidades Fraseológicas Especializadas*: colocações e colocações estendidas em contratos sociais e estatutos sociais traduzidos no modo juramentado e não juramentado. 2009. 290f. Tese (Doutorado em Linguística Aplicada) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, 2009.

ORENHA, A.; CAMARGO, D. C. de. A extração de Unidades Fraseológicas Especializadas a partir de *corpora* paralelos e comparáveis. *The ESPecialist*, São Paulo, v. 30, n. 1, p. 57-81, 2009.

PALACIOS, A. *Os argentinos*. São Paulo: Contexto, 2015.

PALACIOS, A.; CHACRA, G. *Os hermanos e nós.* São Paulo: Contexto, 2014.

PARODI, G. (org.). *Géneros académicos y géneros profesionales*: accesos discursivos para saber y hacer. Valparaíso: Ediciones universitarias de Valparaíso, 2008.

PARODI, G. *Lingüística de Corpus*: de la teoría a la empiria. Madrid / Frankfurt: Iberoamericana / Vervuert, 2010. DOI: https://doi.org/10.31819/9783865278715

SCOTT, M. *WordSmith Tools (7.0)* [Programa computacional]. Liverpool: Lexical Analysis Software, 2016.

SPERANDIO, N. E. As metáforas de Lula: uma forma de legitimação. In: SIMPÓSIO INTERNACIONAL DE LETRAS E LINGUÍSTICA, 2009, Uberlândia. *Anais* [...]. Uberlândia: EDUFU, 2009. p. 1-11.

SPERANDIO, N. *O Modelo Cognitivo Idealizado no processamento metafórico.* 2010. 99f. Dissertação (Mestrado em Letras) – Universidade Federal de São João del Rei, São João del Rei, 2010.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1229-1252, 2021

Analysing the behaviour of academic collocations in a corpus of research-papers: a data-driven study

Analisando o comportamento de colocações acadêmicas em um corpus de artigos científicos: um estudo dirigido por dados

Paula Tavares Pinto
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
paula.pinto@unesp.br
http://orcid.org/0000-0001-9783-2724

Diva Cardoso de Camargo
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
divaccamargo@gmail.com
http://orcid.org/0000-0001-6924-4757

Talita Serpa
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
talita.serpa@unesp.br
https://orcid.org/0000-0003-3324-9593

Luciano Franco da Silva
Universidade Estadual Paulista "Júlio de Mesquita Filho" (UNESP), São José do Rio Preto, São Paulo / Brasil
luciano.francco@gmail.com
https://orcid.org/0000-0001-7485-8657

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1229-1252

Abstract: Authors from different countries have published their papers in English, aiming to promote their research results widely and to become internationally known by their peers. It is also true that, although they are aware of the English terminology used in their respective field, some authors still struggle with some features of academic writing such as collocations. Thus, this paper presents a discussion on the underuse and overuse traces of academic collocations by Brazilian authors who had their articles published in English on an open electronic library of scientific journals. In order to analyse the collocations used by these researchers, we compiled a 906,035-word corpus from eight different academic areas. The collocations observed were statistically compared to those from an academic corpus of English writings which contains texts produced by English-speaking authors. Results showed that there are more collocations underused than overused by the authors. The analysis proved that the collocation repertoire of researchers could be broadened by being pointed out during academic writing workshops.

Keywords: academic collocations; research paper writing; corpus linguistics.

Resumo: Autores de vários países têm publicado seus artigos científicos em inglês com o intuito de promover amplamente os resultados de suas pesquisas dentre a comunidade científica internacional. É verdade que, embora estejam cientes da terminologia utilizada no respectivo campo de pesquisa, alguns autores ainda apresentam dificuldade em lidar com certas características da escrita acadêmica, como o uso das colocações. Este artigo apresenta uma discussão sobre traços de sobreuso e subuso de colocações acadêmicas utilizadas por autores brasileiros que têm seus artigos publicados em inglês numa plataforma eletrônica aberta de artigos científicos. Para analisar as colocações utilizadas por estes pesquisadores, compilamos um corpus de 906.000 palavras a partir de oito áreas científicas. As colocações analisadas foram comparadas estatisticamente com as colocações de um corpus acadêmico de inglês que contém textos escritos por autores anglófonos. Os resultados mostraram que há mais traços de subuso que sobreuso de colocações acadêmicas utilizadas pelos pesquisadores e este repertório poderia ser ampliado se fossem destacadas durante cursos de escrita acadêmica em língua inglesa.

Palavras-chave: colocações acadêmicas; escrita de artigos científicos; linguística de corpus.

Submitted on October 9th, 2020

Accepted on December 16th, 2020

1 Introduction

Authors worldwide recognise the importance of publishing academic articles in English. Although there may be some debate over

the relevance of publishing in one's native language, researchers must publish in English if they want their study results to be read by members of international scientific communities. In that sense, Brazilian authors, who wish to have their studies internationally acknowledged, need to have their articles publicised on online databases, such as *The Scientific Electronic Library Online* (*SciELO*). This platform is an electronic library for Brazilian scientific journals written in Portuguese, Spanish and English.

Taking that into account, several studies (HYLAND, 2008; NESSELHAUF, 2003; PAQUOT, 2010) have already highlighted the fact that non-native speakers may lack the necessary linguistic knowledge to use adequate academic collocations when writing in English. Haswell (1991) has claimed that the underuse of collocations in scientific papers will reveal one's "apprentice writing" which can compromise the acceptance of papers by scientific journals. On the other hand, the proper use of academic collocations would demonstrate how linguistically competent the authors are.

The definition of collocation by the *Oxford Collocations Dictionary for students of English* (LEA; CROWTHER; DIGNEN, 2002, p. vii) is the following: "collocation is the way words combine in a language to produce natural-sounding speech and writing". As examples, the authors state that, in English, it is common to say *strong wind* and *heavy rain*, but not **heavy wind* or **strong rain*.

According to Frankenberg-Garcia *et al*. (2019a), some writers are not aware of collocations or do not use them, which may lead to readers' estrangement caused by combinations such as **depend of* something, instead of **depend on* something. For this reason, the researchers developed the Collocaid Project. The main objective of this tool is to create "a lexicographic resource that is accessed from within digital writing environments to help learners write more idiomatically" (FRANKENBERG-GARCIA *et al*., 2019a, p. 24).

Another topic to be addressed is whether academic collocations stand out to non-native authors as terms and idioms do. According to Nesselhauf (2003), English collocations can be fuzzy for students, academic authors and even native speakers who are not familiar with some commonly patterned combinations. The *Oxford Collocations Dictionary for students of English* states that "collocation runs through the whole of the English language. No piece of natural spoken or written

English is free of collocation" (LEA; CROWTHER; DIGNEN, 2002, p. vii). If collocations in general English are already challenging to be noticed by non-native speakers, we wonder how it would be with academic collocations such as 'rates fell', 'the percentage dropped', 'gather information', 'funding research', among others. Consequently, we question if international researchers, who are non-native speakers of English, can proficiently combine words to produce natural collocations and, more specifically, we want to know how it happens among Brazilian researchers.

Despite the relevance of academic vocabulary and collocations in scientific texts, there are still few studies (DAYRELL 2007; PAIVA, 2009; SILVA *et al*., 2017; SILVA *et al*., 2018) that report their use in the writing of Brazilian authors. Dayrell compared collocational patterns in translated and non-translated texts. The author shows that translations from Portuguese into English draw on a small number of collocates (DAYRELL, 2007, p. 377). Paiva (2009) found evidence of overuse of specific verbs in research papers translated by Brazilian professional translators which are not frequent in articles published in high-impact journals. Babini and Silva (2012) showed that Brazilian researchers produce texts with overuse or underuse of specific lexical items which are generally expected in research papers in English. Silva *et al*. (2018) investigated the use of academic vocabulary by Brazilian (under)graduate students. They concluded that although students use a similar number of academic words compared to the Academic Word List (AWL) and the General Service List (GSL), the word forms chosen by students differ as they underuse affixation processes.

Although the four previous studies refer to the academic vocabulary produced by Brazilians, there are still several issues to be dealt with, such as the use of academic collocations by senior Brazilian researchers who have longer published papers in English. Do they tend to overuse or underuse collocations in their research papers? Are those collocations repeated over the article? These are some of the issues to be discussed in this article.

Therefore, this study seeks to shed some light on the way senior Brazilian researchers use academic collocations in their publications by presenting an investigation of data extracted from a corpus of papers in the eight major areas of research at *The Scientific Electronic Library Online* (SciELO).

The guiding research questions of this study are the following:

1. To what extent do the collocations used by Brazilian authors differ from the ones in international journals?
2. Do Brazilian authors use collocations influenced by their native language (Portuguese)?
3. Are there traces of overuse or underuse of specific collocations?

To answer those questions, we present a brief review of studies that discuss the importance of academic vocabulary and collocations.

2 Academic collocations

Previous studies have revealed that clusters, lexical bundles and collocations have been investigated in different genres of academic writing such as Master's thesis, Doctorate dissertations and research articles (ACKERMANN, CHEN, 2013; CORTES, 2004; FRANKENBERG-GARCIA *et al*., 2019a, 2019b; HYLAND, 2008; SILVA *et al*., 2017). Hyland (2008, p. 42) states that clusters are "words which follow each other more frequently than expected by chance, helping to shape text meanings and contribute to our sense of distinctiveness in a register" such as *a result of* or *it should be noted that* in academic writing. According to the author, mastering the use of these group of words, or "clusters" (SCOTT, 1996) will help non-native writers to overcome linguistic barriers which prevent their papers from reaching other members of the international community. At the same time, Cortes (2004, p. 400) states that "lexical bundles are extended collocations, sequences of three or more words that statistically co-occur in a register. Some examples of these word combinations in academic prose are: *on the other hand*, *in the case of*, *the context of the*, and *it is likely to*."

Firth (1951), in turn, was responsible for making collocations well-known and for the famous quote "you shall judge a word by the company it keeps" (*apud* PARTINGTON, 1998, p. 15). Besides, according to Nation (2001), "the term 'collocation' is used to refer to a group of words that belong together, either because they commonly occur together like *take a chance*, or because the meaning of the group is not apparent from the meaning of the parts, as with *by the way* or *to take someone in*. A significant problem in the study of collocation is determining, in a consistent way, what should be classified as a collocation" (NATION, 2001, p. 317).

Ackermann and Chen (2013) state that another difficulty in dealing with collocations is that they "often contain inflective or positional variations (e.g., *results obtained*, *broader contexts*, *achieving objectives*) which poses the great challenge of how to collate these relevant forms and present them in a uniform and consistent way" (ACKERMANN; CHEN, 2013, p. 236). The authors believe that this challenge can only be overcome by human intervention since there is still no automation method to simplify this process. The researchers mentioned above define collocation as "word combinations which co-occur more frequently than by chance across academic disciplines (hence corpus-driven) and are pedagogically relevant in an EAP[1] context (hence expert-judged)" (ACKERMANN; CHEN, 2013, p. 246). They highlight the importance of compiling a list of academic collocations based on the idea proposed by Nation (2001, p. 189-191). The author stated that academic collocations might "neither be sufficiently frequent in the language as a whole to be learnt implicitly nor part of the technical lexicon which is likely to be explicitly taught as part of subject courses".

Contrary to Nesselhauf's hypothesis (2003), which defines collocation only in its phraseological sense, in this paper we chose to adopt a frequency-based approach, which takes into account co-occurrences of words within a specific span, as Sinclair (1991) did in his work.

As far as teaching collocations is concerned, Nesselhauf (2003) suggests it is a task for teachers to make learners aware of these word combinations. The author adds that teachers should explicitly teach collocations since they do not always stand out to the learners' eyes. The criteria to be followed would be teaching the most frequent and acceptable collocations in the register on focus, in this case, academic collocations (*conduct/do/carry out a study* or *make an analysis*). Comparison to native languages (L1) is also desirable, even by highlighting functional elements such as articles and prepositions. The scholar also suggests that they should give the focus to the verb, which seems to be the cause of most mistakes. Finally, in the Brazilian context, Tagnin (2013) discusses the convention of language and dedicates part of her study to the adjective, noun, verb and adverbial collocations in Portuguese compared to English, Italian, French and Spanish. She also shows their importance in teaching and translation practice.

[1] English for Academic Purpose (EAP).

3 Methodology

The methodology followed in this study was composed of two steps: 1) compilation of the *Brazilian Academic Corpus of English* (BrACE); 2) selection of the most frequent academic collocations used by Brazilian researchers in comparison to frequent academic collocations in native English speakers' writings.

We present these steps in the following sections:

3.1 The Brazilian Academic Corpus of English (BrACE)

In order to identify the most frequent academic collocations used by Brazilian researchers in their writings, we selected papers from SciELO (an open cooperative database of journals originated in Brazil that currently features papers from several countries such as Argentina, Bolivia, Brazil, Chile, among others). This selection aimed to gather information from journals which could represent Brazilian authors' writing in all areas of research designated in Brazil. According to SciELO website:

> The Scientific Electronic Library Online - SciELO is an electronic library covering a selected collection of Brazilian scientific journals. The library is an integral part of a project being developed by FAPESP – *Fundação de Amparo à Pesquisa do Estado de São Paulo*, in partnership with BIREME – the Latin American and Caribbean Center on Health Sciences Information. Since 2002, the project is also funded by CNPq – *Conselho Nacional de Desenvolvimento Científico e Tecnológico*.[2]

The choice of SciELO as the source for our corpus is supported by the works of Neves *et al*. (2016), and Kuhn (2017). These authors also based their studies on the reliability of this procedure, since the selection of papers for SciELO lies on strict criteria and policy, based on "peer-review process, journal usage and impact factor" (KUHN, 2017, p. 194).

The website displays the main areas of domain with respective sub-areas: 1. Agricultural Sciences (*Ciências Agrárias*), with six sub-areas; 2. Biological Sciences (*Ciências Biológicas*), with 14 sub-areas; 3. Health Sciences (*Ciências da Saúde*), with nine sub-areas; 4. Physical

[2] SciELO.org – Scientific Electronic Library Online. Available from: www.scielo.org. Retrieved: May 23, 2018.

and Earth Sciences (*Ciências Exatas e da Terra*), with seven sub-areas; 5. Humanities (*Ciências Humanas*), with ten sub-areas; 6. Applied Social Sciences (*Ciências Sociais Aplicadas*), with 12 sub-areas; 7. Engineering (*Engenharias*), with 12 sub-areas; 8. Languages, Linguistics and Arts (*Linguística*, *Letras* e *Artes*), with three sub-areas.

Since some journals belonged to two or more different areas of SciELO, some of the articles were stored under the concept of interdisciplinary studies. There were some overlapping between some areas, for example, Agricultural Sciences overlapped with Chemical Engineering because some of the articles discussed soil use as well as chemical components used in agriculture. The same happened to Physical and Earth Sciences when some articles discussed topics related to Agriculture and Archaeology, which were also present in other interrelated areas. Our criterion was to follow the distinction made by the SciELO platform since they certainly had a reason for separating the publications under specific areas, as well as choose the ones with higher impact within each particular area.

The impact is based on the Qualis of journals, which is a Brazilian ranking used by the Coordination for the Improvement of Higher Education Personnel (CAPES- *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior*) to evaluate the quality of scientific journals in Brazil. In order not to have a random sample, we selected papers from, at least, two different journals from the same scientific area, starting in 2018 so as to have the most recent issues. We also observed some articles from previous years, which had been ranked as B2, B1, A2 or A1, corresponding to the highest journal impact for Qualis. This procedure would guarantee the excellent quality of these papers in each scientific community. One example was *Acta Botanica Brasilica* which had been ranked as B2 for Biodiversity and as B5 for Biology. Then, in this case, we selected five articles whose theme had to do with Biodiversity (B2) and looked for other journals that would discuss other areas of Biology related to animals, whose score for Qualis would be, at least, B2.

Following the areas of SciELO, we selected twenty (20) articles from each journal whose writings had been published in English by Brazilian authors or teams. The journals published most of the chosen papers between 2017 and 2018. However, in some areas, such as Physics and Humanities, the most recent papers were published in 2010. We decided to keep these papers to maintain the broadest range of subareas

within each domain. We accessed the electronic versions of the journals, and the texts were downloaded and saved according to criteria based on their specific areas. In a different document, we held the references for all articles used with the same tag they would have in the corpus.

Since the primary goal of compiling this corpus was to have texts that would display academic collocations and clusters, we selected the complete articles with tables, abstracts and references. The tables and figures were not a problem since the program used for analysis, *Sketch Engine*® (KILGARRIFF *et al*., 2014), does not read them.

After following the criteria previously described, we compiled a 906,035-word corpus using *Sketch Engine*. At the end of this process, the BrACE corpus data was organised as follows:

TABLE 1 – Brazilian Academic Corpus of English (BrACE)

Areas	Journals and Years	Papers	Words
1. Agricultural Sciences	1. Acta Scientiarum. Agronomy, 2018; 2. Arquivo Brasileiro de Medicina Veterinária e Zootecnia, 2018 e 2017.	20	88,740
2. Biological Sciences	1. Acta Botanica Brasilica, 2018, 2017; 2. Memórias do Instituto Oswaldo Cruz, 2018.	20	92,220
3. Health Sciences	1. Jornal Brasileiro de Pneumologia, 2018; 2.Arquivos de Neuro-Psiquiatria, 2018; 3. Brazilian Dental Journal, 2018; 4. Brazilian Journal of Pharmaceutical Sciences, 2018, 2017.	20	74,254
4. Physical and Earth Sciences	1. Brazilian Journal of Physics, 2010; 2.Revista Brasileira de Meteorologia, 2017; 3. Brazilian Journal of Oceanography, 2017; 4. Boletim de Ciências Geodésicas, 2017	20	82,440
5. Humanities	1. Ambiente & Sociedade, 2017; 2. Brazilian Journal of Political Economy, 2017; 3.Cadernos Pagu, 2010	20	151,952
6. Applied Social Sciences	1. Ambiente & Sociedade, 2017.	20	142,930
7. Engineering	1. Journal of Aerospace Technology and Management, 2018; 2. Journal of Microwaves, Optoelectronics and Electromagnetic Applications, 2017; 3. Latin American Journal of Solids and Structures, 2017; 4. Revista IBRACON de Estruturas e Materiais, 2017.	20	109,236
8. Languages, Linguistics and Arts	1. Alfa: Revista de Linguística, 2017; 2. Revista Brasileira de Estudos da Presença, 2018; 3. Ilha do Desterro, 2018, 2017.	20	164,263
TOTAL			**906,035**

Source: BrACE corpus

In the following section, we explain how we analysed the collocations in BrACE.

3.2 Selection of the most frequent academic collocations used by Brazilian researchers in comparison to frequent academic collocations in English

In this study, we used semi-automatic retrieval of collocations, that is to say, statistical information and human judgement. We used a whitelist to generate a list of words that coincided with a combination of three well-known EAP vocabulary lists:

(i) the Academic Vocabulary List (AVL-BAWE), based on the Corpus of Contemporary American English (COCA) by Gardner and Davies (2014);

(ii) the Academic Keyword List (AKL), based on the list of keywords extracted by Paquot (2010), and

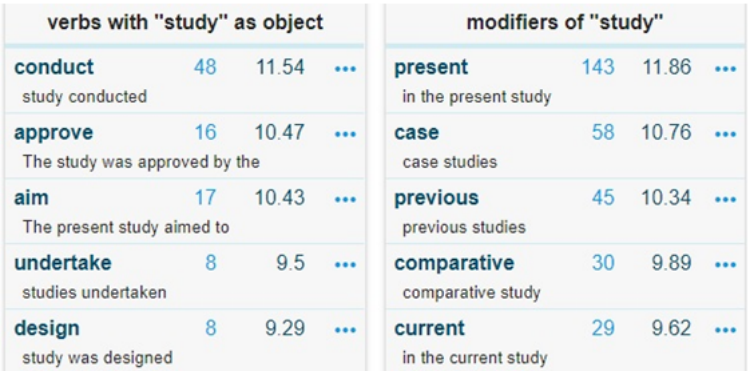
(iii) the Academic Collocations List (ACL) by Ackermann and Chen (2013).

We did this process during the time we had access to the database of the Collocaid project (FRANKENBERG-GARCIA *et al.*, 2019a) in which these lists had been used. The ColloCaid project is dedicated to developing a text-editing tool to help writers with collocations during the writing process. The research involves "investigating user needs, the visualisation of lexicographic data and human-computer interaction, and compiling an extensive database of collocation suggestions using state-of-the-art e-lexicography tools and resources".[3]

We started the selection of lexical words with nouns as base forms to observe how they would collocate most frequently in the BrACE corpus. The most frequent nouns in the list were studied. To illustrate the steps taken, we made a query with *study* as search word using a tool called WordSketch, which is a "one-page summary of a word's grammatical and collocational behaviour" (KILGARRIFF *et al.*, 2014, p. 9):

[3] Available from: https://www.collocaid.uk/.

FIGURE 1 – Screenshot of the query for "study" as a noun in the BrACE corpus



Source: Sketch Engine®

In Figure 1, we see two different lists of words that are commonly combined with the search word "study". On the left, we have verbs that co-occur with the study as an object, such as "conduct + study", "approve + study", "aim + study". On the right we see modifiers of "study" as in "present + study", case +study" and "previous + study".

We selected single words that tended to co-occur in the span of three words from the reference word, coinciding at least five times in the corpus and having a LogDice score of, at least, 7. This kind of statistical data will "indicate how strong the collocation is. The higher the score, the stronger the combination of words is. A low score means that the words in the collocation also frequently combine with many other words". This decision was taken considering previous papers that reported the statistics used in the extraction of collocations from small and large corpora (CORTES, 2004, DAYRELL, 2007; ACKERMANN, CHEN, 2013; FRANKENBERG-GARCIA *et al*., 2019a).

The next step was analysing the list of (i) "modifiers" that collocated with the search word; (ii) verbs with the search word as "object" and (iii) verbs with the search word as "subject".

The search words and their collocates were saved in a list showing the frequency of each word combination to compare them to the reference list of common collocations in English.

We excluded terms (*translation/epidemiological/environmental* study; *discourse/ scientometric* analysis) and combinations with copular or auxiliaries (be – studies *were*…, have – *have* shown). The aim was

to analyse general academic collocations instead of terms from specific areas. This way, we could retrieve collocations mostly used by Brazilian authors such as *the present study*, *case study*, *previous study* (modifier + study); *conduct a study*, *achieve/approve/aim a study* (verbs + study as object); *studies demonstrated*, this *study showed*, this *study suggests* (verbs + study as subject).

After selecting frequent collocations from BrACE, we looked for the ones that were not so frequently used by English authors in *The Oxford Corpus of Academic English* (OCAE), which is a 71,372,972-word corpus, to check if they were not used at all or if they were rarely used. The access to this corpus was possible during a period of a sabbatical break in which we worked with a research team who had this permission.

4 Results and Analysis

In this section, we present the results of our study concerning the academic collocations used by Brazilians in their papers published on SciELO, as well as characteristics of overuse and underuse.

4.1 Academic collocations overused by Brazilian researchers in comparison to frequent academic collocations in English

As presented in the methodology, we compared the wordlist of BrACE to the three academic vocabulary lists and selected the first twenty most frequent words, which were ranked from the most to the least frequent ones. We analysed them as candidates for academic collocations.

The first word class we observed from this list were nouns. We analysed collocations which had been frequently used by Brazilian authors with these nouns but were not as frequent in the three academic vocabulary lists commonly used by researchers who publish in English. We took this step to observe too frequent (overused) or uncommon (underused) collocations that had been chosen by Brazilian researchers and were not as frequent in papers originally written in English in the OCAE. As we will see, there were less overused collocations than the underused ones.

The overused collocations in BrACE that were not as frequent in the three lists of comparison were: *corroborate + study (obj.) / study (subj.) + corroborate / study (subj.) + reinforce / analysis (adj.) + finite / analysis (adj) + correlation / make + analysis (obj.) / consider + analysis (subj.) / intensive (adj.) + use* and *present (adj.) + work*

After comparing the collocations from BrACE to the ones in the three academic lists, we analysed the specific examples in the OCAE. Although they were all part of the OCAE, we wanted to confirm whether their co-occurrence and LogDice scores were similar. The collocations are presented in the following table that shows the base word and its relation to the collocate, an example from BrACE, its co-occurrence, and LogDice in BrACE and OCAE respectively:

TABLE 2 – collocations overused in BrACE in comparison to OCAE

Base (relation) + collocate **Example from the BrACE**	**Co-oc. BrACE**	**LogDice**	**Co-oc. OCAE**	**LogDice**
1. Study (obj. of) corroborate These results *corroborate* previous biomechanical *studies* that found a lower stress concentration for wide diameter implants, especially in short implants. (|Health)	6 (4.89 per million)	9.06	4 (0.05 per million)	3.78
2. Study (subj. of) corroborate Several other *studies* have *corroborated* these findings, which indicate a change in cardiac autonomic modulation, demonstrating impairment of this activity in individuals with COPD. (Health)	5 (4.07 per million)	7.92	5 (0.06 per million)	4.29
3. Study (subj. of) reinforce This *study reinforces* the lines already traced out in recent research on the need to consider multidimensional approaches when analysing human-nature relationships. (Social Sciences)	5 (4.07 per million)	7.91	6 (0.07 per million)	4.44
4. Analysis (adj.) finite *Finite element analysis* on the influence of implant surface treatments, connection and bone types. (Health)	22 (17.91 per million)	9.26	50 (0.60 per million)	5.51
5. Analysis (adj) correlation The RDC results (Figure3) were similar to the results obtained for the *correlation analysis* for the two study years and all of the soil layers measured, with a correlation of -0.88 between the RDC and Pearson's correlation. (Agriculture)	10 (8.14 per million)	8.22	108 (1.28 per million)	6.6
6. Analysis (obj. of) make In the context described above, we *analys*ed the potential impacts from the installation and operation steps of this project, considering the understanding of the oceanographic processes and possible effects on the human well-being caused by the undermining of provided ecosystem services. (Biological Sciences)	9 (7.33 per million)	8.68	161 (1.91 per million)	6.06

7. Analysis (subj. of) consider This *analysis considers* the average of 100 independent runs. (Engineering)	6 (4.89 per million)	9.12	18 (0.21 per million)	6.55
8. Use (adj.) intensive In American agriculture, the conversion of conventional tillage systems to no-till systems and the *intensive use* of glyphosate in transgenic cropping has significantly influenced the composition and populations of weeds. (Agriculture)	5 (4.07 per million)	8.97	48 (0.57 per million)	6.15
9. Work (adj. of) present In the *present work*, it was evident that cellular debris from both the uterine epithelium and the trophoblastic cells are phagocytosed and digested by active trophoblastic cells. (Biological Sciences)	27 (21.99 per million)	10.73	103 (1.22 per million)	6.57

Source: Authors

As shown in Table 2, some collocations did not have a LogDice higher than 7.0 in the OCAE despite the fact they had higher LogDice in the BrACE, which could be an indication of overuse. These collocations were: *corroborate study*; *study confirms*; *study reinforces*; *finite element analysis*; *correlation analysis, make analysis, the analysis considers* and *intensive use*.

Although we found these collocations in the OCAE, they are not so frequently used in papers written by native English authors. Besides, the same collocations were not frequent in the combination of academic lists as well.

We looked up for collocational options with the same nouns in the OCAE that could replace the ones used by Brazilians. To do so, we used the same nouns as search words in Word Sketch to look for collocations with similar meanings. However, we looked for combinations with higher frequency in the OCAE, which might sound more natural to international researchers. For the collocations with *study* + *corroborate*, the optional choices in the OCAE would be: *study* + *support* / *confirm*. So, the sentence below, taken from BrACE, could be written in the following way:

> These results [*support*] [*confirm*] previous biomechanical *studies* that found a lower stress concentration for wide diameter implants, especially in short implants.

> Several other *studies* have [*supported*] [*confirmed*] these findings, which indicate a change in cardiac autonomic modulation, demonstrating impairment of this activity in individuals with COPD.

As for the collocation *study* + *reinforce*, a similar meaning with more substantial LogDice score would be *study* + *highlight.* In this case, the sentence used by the Brazilian author would be:

> The findings of the present *study* [*highlight*] the importance of banning tobacco displays at the point of sale.

Although the collocation *finite element* + *analysis* did not show a high LogDice score (5.51) in the OCAE, we found it in the sub-corpus of Engineering, which could mean it is a discipline-specific collocation, as in the example below:

> The *finite element analysis* of any problem involves four steps: (a) discretising the solution region into a limited number of sub-regions or elements, (b) deriving governing equations for a typical feature, (c) assembling all the parts in the solution region, and (d) solving the system of equations obtained.

A similar case is the collocation *correlation* + *analysis*, which has a low LogDice score in the OCAE (6.6) but is used in the areas of Medicine, Education and Computer Sciences, as the examples below:

> To examine the role of parents and friends as sources of influence on girls' college aspirations and motivation to achieve their goals, we conducted a series of *correlation analyses* separately for girls who were sexually active and those who were not.

Although the collocation *make* + *analysis* was high, other options found in the BrACE would be more aligned with the OCAE such as *perform/conduct/apply* + *analysis*. Therefore, the sentence below would sound more natural in the following way:

> In the context described above, we [*performed*] [*conducted*] [*applied*] an *analysis* of the potential impacts from the installation and operation steps of this project (…)

The collocation *analysis* + *consider* was not present among the most common collocations of OCAE. We believe the best option, in this case, would be the expression "the analysis takes into consideration" as in:

> This *analysis takes into consideration* the average of 100 independent runs.

The collocation *intensive use* could be substituted by *widespread/ increased/unrestricted/extensive* + *use* as it is found in the OCAE.

In American agriculture, the conversion of conventional tillage systems to no-till systems and the [*widespread*] [*increased*] [*unrestricted*] [*extensive*] use of glyphosate in transgenic cropping has significantly influenced the composition and populations of weeds.

The next section of this article presents collocations that were underused by Brazilian authors.

4.2 Academic collocations underused by Brazilian researchers in comparison to frequent academic collocations in English

This time, we checked collocations that had not been so frequently used by Brazilian authors with the list of nouns we had, but were present in the EAP lists we had used in the first step. The underused collocations from BrACE that were not as frequent in the three lists of comparison were: *qualitative (adj.) + study / detail (adj.) + analysis / restrict + analysis (obj.) / extensive (adj.) + use / widespread (adj.) + use / increase (adj.) + use / support + use (obj.) /* encourage + *use (obj.) / design + system (obj.)/ system (subj.) + work / describe + process (obj.) / begin + process (obj.) / collect + data (obj.) / data (subj.) + suggest / data (subj.) + indicate / development (subj.) + occur / facilitate + development (obj.) / further (adj.) + development.*

Once again, we compared the co-occurrence of these collocations in BrACE to the OCAE, and we present the eighteen first ones below within their context in the OCAE:

TABLE 3 – Collocations underused in BrACE in comparison to OCAE

Base (relation) + collocate Example from the OCAE	Co-oc. BrACE	LogDice	Co-oc. OCAE	LogDice
1. Study (adj.) qualitative This *qualitative study* found that while knowledge about the TRiM system was not widespread, the majority of those personnel who were aware of TRiM viewed it positively and supported it being peer-delivered. (Medicine)	0	0	378 (4.48 per million)	7.82
2. Analysis (adj.) detail Poor exposures and a lack of reliable criteria prevent *detailed analysis* of the lower slope. (Earth Sciences)	2 (1.63 per million)	5.91	550 (6.51 per million)	8.73
3. Analysis (obj. of) restrict Therefore, we *restrict* our *analysis*, somewhat arbitrarily, to deflections of the form: w (x, y) = e wEu (x) + s wS (x) cos (kS y) + a wA (x) cos (kA y + A) (8.74). (Engineering)	0	0	103 (1.22 per million)	8.22
4. Use (adj.) extensive This propaganda campaign made *extensive use* of petitions as a device for expressing extra-parliamentary pressure on a public issue. (History)	0	0	188 (2.23 per million)	7.91
5. Use (adj.) widespread Although the genetics of many lower eukaryotic organisms had been studied in some detail, Beadle and Tatum's work initiated a much more *widespread use* of microbes. (Biochemistry)	1 (0.81 per million)	6.68	306 (3.62 per million)	8.74
6. Use (adj.) increase We are seeing the *increasing use* of computational modeling within historical linguistics and interdisciplinary research is not the exotic enterprise it used to be. (Linguistics)	0	0	383 (4.54 per million)	8.14
7. Use (obj.of) support The current analyses further *support* the *use* of extreme groups with an underpinning rationale. (Education)	0	0	160 (1.90 per million)	7.89
8. Use (obj.of) encourage The ease of storing, transmitting, and processing electrical data is *encouraging* the *use* of unmanned stations. (Earth Science)	0	0	110 (1.30 per million)	7.74
9. System (obj. of) design A small group of engineers (3 full-time and 4 part-time workers) used axiomatic design *to design* a *system* that can satisfy the requirements for crew survivability in a short time (5 months). (Engineering)	2 (1.63 per million)	7.29	460 (5.45 per million)	8.75

10. System (subj. of) work As I do not ever recollect an urgent request having been refused, this *system worked* very satisfactorily from our point of view. (Engineering)	0	0	133 (1.58 per million)	8.71
11. Process (obj. of) describe The *process described* there, by which lay people decide which action to take about symptoms of illness, is probably not greatly different from the general way that doctors diagnose illness. (Medicine)	3 (2.44 per million)	7.53	430 (5.09 per million)	8.84
12. Process (obj. of) begin Shot 7 initiates a new line of dramatic action that poses the question of what Lucy will do now, and also *begins a process* not exactly of rereading, but a search for a new reading of the meaning of the setups. (Media Cultural Studies)	0	0	164 (1.94 per million)	8.57
13. Data (obj. of) collect Lyons et al. (1998a) noted that geochemistry data from Lake Fryxell in the McMurdo Dry Valleys indicated an overall change in the ionic composition of the lakes when *data collected* in the mid-1990s are compared to older, but reliable, data obtained in the early 1960s. (Biological Sciences)	1 (0.81 per million)	7.83	1,786 (21.16 per million)	11.36
14. Data (subj. of) suggest We presented the 10th-grade classroom because our *data suggest* that Ms. Young fits Irvine's description of an "experienced and masterful pedagogue" who is "seeing with the cultural eye" (Irvine, 2001). (Education)	0	0	256 (3.03 per million)	9.47
15. Data (subj. of) indicate Together these *data indicated* a decline in grasslands and an increase in shrublands in the early Holocene. (Earth Science)	1 (0.81 per million)	7.21	132 (1.56 per million)	9.12
16. Development (subj. of) occur The next *development occurred* in the plateau country of Arizona, Utah, and Colorado. (Earth Science)	0	0	56 (0.66 per million)	7.24
17. Development (obj. of) facilitate It is surely in the interest of countries near and far away to *facilitate* the *development* of knowledge, skill, and freedom in these countries so they can become contributing, responsible members of the international community rather than breeding grounds for social pathology, infectious diseases, and terrorist violence. (Education)	0	0	126 (1.49 per million)	8.24
18. Further (adj. of) development The establishment and *further development* of this cascade provides us with a fertile research agenda. (Physical and Earth Sciences)	0	0	382 (4.52 per million)	8.16

Source: Authors

The collocations presented above have not been frequently used by the Brazilian researchers in their texts represented in our corpus. The examples were all taken from *The Oxford Corpus of Academic English*, which means that to have a more natural text, it would be necessary for Brazilian researchers to be aware of this use and try to incorporate these collocations into their writings.

In the next section, we discuss the general results of this study based on the observation of overuse and underuse of academic collocations used by Brazilian researchers in their articles.

5 Discussion

The discussions presented in this section seek to answer the three research questions stated at the beginning of this paper. The first one was "To what extent do the collocations used by Brazilian authors differ from the ones in international journals?". Although Brazilian researchers have had their papers published in high-impact academic journals, we could see that there are significant differences regarding underused collocations, which outnumber the overused ones. This result shows that these writers were not aware of some of the collocations mostly used by scholars in international journals. These extracts are not so different to Brazilian Portuguese such as a *detailed (adj.) + analysis, restrict + analysis (obj.), extensive (adj.) + use, widespread (adj.) + use, describe + process (obj.)* and *begin + process (obj)*. We did not expect some of the results such as the underuse of collocations as *collect + data* and *data + suggest* which are not so different from the Brazilian Portuguese. Because of that, further studies will be carried out as soon as we have more articles added to the BrACE corpus so that we can confirm or not the lack of some collocations in those articles.

The previous result leads us to the second and third questions, which are: "Do Brazilian authors use collocations influenced by their native language (Brazilian Portuguese)?" and "Are there traces of overuse or underuse of specific collocations?".

We could find evidence that indicates the influence of Brazilian Portuguese in the choice of collocations which called our attention. This is the case of s*tudy* (obj. of) + *corroborate* and s*tudy* (subj. of) + *corroborate* which were overused by the Brazilian researchers and have the equivalent in Portuguese "estudo (obj of) + corroborar" and "estudo

(subj. of) + corroborate" which are very common in articles written in this language. This result pointed out to the trace of collocation overuse. Although this combination has been found in the OCAE, it is not as frequent in research papers initially written in English, which clearly shows the influence of Portuguese in those texts.

Another comparison we can make is that Brazilians *suggest the use of* whereas authors who commonly write in English *support the use o*r *encourage* it. At the same time, instead of *data points*, Brazilians most commonly write *data indicates that*.

Upon analysing different areas of research, the collocation *qualitative study* is present in areas such as Business, Medicine and Sociology in the OCAE. In contrast, in BrACE, we find *qualitative analysis*, but not a *qualitative study*. The same happens to *regression analysis*, which is the first most frequent collocation with research in the OCAE but is not present in the BrACE. In cases like this, it is necessary to consider that the BrACE is still a small corpus of 906,035 words and some collocations not found here may start to appear as the corpus grows. These limitations do not allow us to generalise the behaviour of academic collocations as a whole but show Brazilian researchers' preferences.

It would be desirable to compare the results shown here to international authors who frequently publish in renowned journals of different domains.

Regarding the methodology, as stressed by Dayrell (2011), it would be interesting to analyse a lemmatised corpus to see the behaviour of the same lemma in different contexts as well as different span values and strength of association between nodes and collocates. Another interesting perspective would be the investigation of an additional criterion of window-sizes of collocations that could range more four words to the right and the left. It would allow us to observe longer phraseologies in research papers written by Brazilian or international researchers.

6 Final Remarks

The primary aim of this study was to identify the most frequent collocations used by Brazilian authors who had their research papers published in the eight major areas of SciELO. After identifying these collocations, we compared them to the most frequent academic ones used

by native English writers and international research groups so we could locate academic collocations that had been overused and underused by Brazilian researchers.

These results have led us to suggest further studies and actions to encourage Brazilian researchers to write more naturally in English academic style. By doing so, they will become aware of these differences in academic language that may not have been noticed in their writings.

As suggested by Nesselhauf (2003), teachers could point out the most relevant collocations through writing exercises in academic workshops or courses of academic English. The author argues that we should explicitly teach collocations since they do not always stand out to the learners' eyes. The main suggestion is to start by introducing the most frequent and acceptable collocations and, then, comparing them to native researchers' textual productions. Having these results, consequently, we could stress functional elements such as the difference between possible combinations in English and those that are more common in the students' native language. In this way, we believe that researchers would be more familiar with the language patterns used in research papers published in high-impact academic journals.

It would also be desirable to encourage students to write abstracts and papers when they are still in college so they become more and more familiar with the academic English. Also, teachers should encourage students to read as many quality papers written in English as possible so that students became aware of their specific research communities writing style. This practice would certainly enhance the use of academic collocations. Another way of stimulating the students to use more collocations would be explicitly showing them samples of sentences containing these structures.

As for senior researchers, it would be ideal to show them the collocations commonly used in their areas through writing crash courses and by teaching them to compile their corpora to be used as examples of writing in each area. By doing so, they would be acquainted not only with the language style and structure but also with genre constraints in each area.

Actions like these have already been taken as, for example, the writing masterclasses supported by the British Council in which Brazilian researchers and EAP tutors (FRANKENBERG-GARCIA *et al*., 2019b) worked together to develop their writing autonomy through the use of specialised corpora and linguistic tools.

Acknowledgements

We would like to thank Dr. Ana Frankenberg-Garcia and Dr. Geraint Rees for all valuable suggestions during the development of this research work at the University of Surrey. The authors would also like to acknowledge funding from the São Paulo Research Foundation (FAPESP/16/25198-6).

Authorship statement

This study reports on data from Dr. Paula Tavares Pinto's Post-Doctoral research at the University of Surrey. The first author was in charge of gathering data, transferring the data to spreadsheets for data analysis, and writing the first draft of the article. The four authors collaborated on interpreting results and revising the essay and the data analysis, including the statistics.

References

ACKERMANN, K.; CHEN, Y. H. Developing the Academic Collocation List (ACL). A Corpus-Driven and Expert-Judged Approach. *Journal of English for Academic Purposes*, [*S.l.*], v. 12, n. 4, p. 235-247, 2013. DOI: https://doi.org.10.1016/j.jeap.2013.08.002

BABINI, M.; SILVA, E. B. A terminologia acadêmica nos textos científicos em língua inglesa uma abordagem baseada em corpus. *In:* ISQUERDO, A. N.; SEABRA, M.C.T.C. (org.). *As ciências do léxico:* lexicologia, lexicografia, terminologia. Campo Grande: UFMS, 2012. p. 415-427.

CORTES, V. Lexical Bundles in Published and Student Disciplinary Writing: Examples from History and Biology. *English for specific purposes*, [*S.l.*], v. 23, n. 4, p. 397-423, 2004. DOI: https://doi.org.10.1016/j.esp.2003.12.001

DAYRELL, C. A Quantitative Approach to Compare Collocational Patterns in Translated and Non-Translated Texts. *International Journal of Corpus Linguistics*, [*S.l.*], v. 12, n. 3, p. 375-414, 2007. DOI: https://doi.org.10.1075/ijcl.12.3.04day

DAYRELL, C. Corpora no ensino de inglês acadêmico: padrões léxico-gramaticais em abstracts de pós-graduandos brasileiros. *In*: VIANA, V.; TAGNIN, S. (org.). *Corpora no Ensino De Línguas Estrangeiras*. São Paulo: HUB Editorial, 2011. p. 131-172.

FIRTH, J. R. *Modes of Meaning*. v. 4: Essays and Studies (English Association). Indianapolis: Bobbs-Merril, 1951. p. 118-149.

FRANKENBERG-GARCIA, A. *et al*. Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time. *ReCALL*, Cambridge, v. 31, n. 1, p. 23-39, 2019a. DOI: https://doi.org.10.1017/S0958344018000150

FRANKENBERG-GARCIA, A. *et al*. *Supporting the Internationalisation of Brazilian Research*: Curso oferecido via financiamento *Capes: Print* para a Universidade Federal do Rio Grande do Sul e a Universidade Estadual Paulista, 4-06 de jun.de 2019. 30f. Notas de aula.

GARDNER, D.; DAVIES, M. A New Academic Vocabulary List. *Applied Linguistics*, Oxford, v. 35, n. 3, p. 305-327, 2014. DOI: https://doi.org.10.1093/applin/amt015

HASWELL, R. *Gaining Ground in College Writing:* Tales of Development and Interpretation. Dallas: Southern Methodist University Press, 1991.

HYLAND, K. Academic Clusters: Text Patterning in Published and Postgraduate Writing. *International Journal of Applied Linguistics*, [*S.l.*], v. 18, n. 1, p. 41-62, 2008. DOI: https://doi.org.10.1111/j.1473-4192.2008.00178.x

KILGARRIFF, A. *et al*. The Sketch Engine: Ten Years On. *Lexicography*, Sheffield, UK, v. 1, n. 1, p. 7-36, 2014. DOI: https://doi.org/10.1007/s40607-014-0009-9

KUHN, T. A Design Proposal of an Online Corpus-Driven Dictionary of Portuguese for University Students. 2017. 421f. Tese (Doutorado em Linguística Aplicada) – Faculdade de Letras, Universidade de Lisboa, Lisboa, 2017.

LEA, D.; CROWTHER, J.; DIGNEN, S. *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press, 2002.

NATION, I. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press, 2001. (Cambridge Applied Linguistics).

NESSELHAUF, N. The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, Oxford, v. 24, n. 2, p. 223-242, 2003. DOI: https://doi.org/10.1093/applin/24.2.223

NEVES, M. L.; JIMENO-YEPES, A.; NÉVÉOL, A. The SciELO Corpus: a Parallel Corpus of Scientific Publications for Biomedicine. *In*: INTERNATIONAL CONFERENCE ON LANGUAGE RESOURCES AND EVALUTATION – LREC, 10th., 2016, Portorož, Slovenia. *Proceedings* […]. Portorož: LREC, 2016.

PAQUOT, M. *Academic Vocabulary in Learner Writing*: From Extraction to Analysis. London: Continuum, 2010.

PARTINGTON, A. *Patterns and Meanings:* Using Corpora for English Language Research and Teaching. Amsterdam: John Benjamins Publishing, 1998. DOI: https://doi.org/10.1075/scl.2

PAIVA, P. T. Uma investigação de traduções de textos da área médica sob a luz dos estudos da tradução baseados em corpus. 2009. 288f. Tese (Doutorado em Linguística Aplicada) – Instituto de Biociências, Letras e Ciências Exatas, Universidade Estadual Paulista, São José do Rio Preto, 2009.

SCIENTIFIC ELECTRONIC LIBRARY ONLINE. Available from: https://scielo.org/. Access on: Jun. 13, 2020.

SCOTT, M. *WordSmith Tools 4*. Oxford: Oxford University Press, 1996.

SILVA, E. B.; BABINI, M.; OTTAIANO, A. O. Identification of the most common phraseological units in the English language in academic texts: contributions coming from corpora. *Acta Scientiarum*, Maringá, v. 39, p. 345-353, 2017. DOI: https://doi.org/10.4025/actascilangcult.v39i4.31811

SILVA, L. G.; MATTE, M. L.; SARMENTO, S. Brazilian Students's Use of English Academic Vocabulary: An Exploratory Study. *In:* FINATTO, M. J. *et al.* (org.). *Linguística de corpus:* perspectivas. Porto Alegre: Instituto de Letras, UFRGS, 2018. p. 509-526.

SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

TAGNIN, S. E. O. *O jeito que a gente diz* - combinações consagradas em inglês e português. Barueri: Disal Editora, 2013.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1253-1288, 2021

CC BY

"Quero que vocês me acompanhem nessa jornada": análise da emergência de metáforas em narrativas sobre o câncer de mama a partir de estratégias de Linguística de *Corpus*

"I want you to come with me in this journey": analysis of the emergence of metaphors in breast cancer narratives based on Corpus Linguistics strategies

Ana Rachel Salgado
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, Rio Grande do Sul / Brasil
anasalgado@ufcspa.edu.br
http://orcid.org/0000-0001-5612-8191

Aline Aver Vanin
Universidade Federal de Ciências da Saúde de Porto Alegre (UFCSPA), Porto Alegre, Rio Grande do Sul / Brasil
alinevanin@ufcspa.edu.br
http://orcid.org/0000-0002-9984-6043

Gabriele Honsha Gomes
gabrielehgomes@gmail.com
Hospital de Clínicas de Porto Alegre – Residência Multiprofissional, Porto Alegre, Rio Grande do Sul / Brasil
http://orcid.org/0000-0002-0076-3951

Leticia Presotto
Escola Superior de Propaganda e Marketing – Câmpus ESPM Sul, Porto Alegre, Rio Grande do Sul / Brasil
letipresotto@gmail.com
http://orcid.org/0000-0001-8130-8450

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1253-1288

Resumo: O objetivo deste trabalho é investigar como a emergência de metáforas conceptuais revela as experiências subjetivas nas narrativas de pacientes que desenvolveram e tratam o câncer de mama e suas estratégias de *coping*. Para tanto, é proposta uma adaptação de método, baseado em Linguística de Corpus, para a identificação dessas metáforas encontradas no discurso dessas mulheres com base na identificação e na extração de termos candidatos a possíveis domínios conceptuais. Foram coletados textos de 31 blogs de livre acesso sobre a temática em estudo, constituindo um *corpus* de 2036 textos. Elegeu-se um dos blogs, constituído por 23 postagens, como referência para avaliar se a ferramenta escolhida e a metodologia adaptada seriam adequadas ao estudo. A partir disso, foi criada uma *keyword list* para extrair termos candidatos a domínios conceptuais, constituindo-se uma lista de referência para a análise dos demais textos. Dentre os domínios conceptuais mais frequentes, emergiram os seguintes: ENTIDADE, FORÇA DA NATUREZA, JOGO, CONTAINER, VIAGEM, VALOR MONETÁRIO E GUERRA. Também, em menor medida, elementos ligados à religião e à espiritualidade acabaram surgindo. A partir da análise empreendida, destaca-se que a realização de metáforas nas narrativas atua como estratégias de *coping*, haja vista que elas são indícios da elaboração conceitual dessas experiências.

Palavras-chave: metáfora conceptual; linguística de *corpus*; *coping*; câncer de mama.

Abstract: The aim of this paper is to investigate how the emergence of conceptual metaphors reveals subjective experiences in the narratives of patients who developed and treat breast cancer and their coping strategies. Therefore, an adaptation of a method, based on Corpus Linguistics, is proposed to identify these metaphors found in these women's discourse based on the identification and in the extraction of candidate terms for possible conceptual domains. Texts were collected from 31 freely accessible blogs on the subject under study, constituting a corpus of 2036 texts. One of the blogs, consisting of 23 posts, was chosen as a reference to assess whether the chosen tool and the adapted methodology combined with the study. Based on this, a list of keywords was created to extract candidate terms for conceptual domains, constituting a reference list for the analysis of the other texts. Among the most frequent conceptual domains, the following emerged: ENTITY, STRENGTH OF NATURE, GAME, CONTAINER, TRAVEL, MONETARY VALUE and WAR. Also, to a lesser extent, elements linked to religion and spirituality eventually emerged. From the analysis undertaken, it is highlighted that the realization of metaphors in the narratives act as coping strategies, given that they are evidence of the conceptual elaboration of these experiences.

Keywords: conceptual metaphor; corpus linguistics; coping; breast cancer.

Recebido em 09 de outubro de 2020

Aceito em 09 de dezembro de 2020

1 Introdução

A escrita em blogs se apresenta como espaço de compartilhamento não só de ideias, mas de expressão de sentimentos. Essa ferramenta, hoje em grande parte substituída pelas redes sociais, permite a expressão de ideias e percepções para diversas pessoas, que são, por vezes, desconhecidas. Desde seu surgimento, os blogs tornaram-se sistemas de publicação na internet em que qualquer pessoa pode escrever baseada nos princípios de microconteúdo: textos curtos, com informações que considera relevantes, seguindo determinado padrão, e atualizados frequentemente (PRIMO; RECUERO, 2008). Em muitos casos, elaborar textos nesse gênero revela-se uma estratégia também para lidar com dificuldades, como é o caso das vivências com o luto (KARKAR; BURKE, 2020) ou da descoberta e do tratamento do câncer (SEMINO *et al.*, 2018). Neste trabalho, nós exploramos como mulheres que desenvolveram câncer de mama relatam suas experiências com a doença e com o tratamento por meio da análise de seus textos postados em blogs abertos ao público. Tratam-se de escritas de si, direcionadas a um público imaginado, possivelmente com o intuito de encontrar estratégias para elaborar suas trajetórias e de enfrentar a doença e o tratamento – que, aqui, chamaremos de estratégias de *coping* (ANDRADE *et al.*, 2020; GUSTAFSSON *et al.,* 2019; SEMINO *et al.,* 2015; STUMM, *et al.,* 2009).

Pacientes que recebem o diagnóstico do câncer têm a tendência de sentirem-se inseguras em relação à sua condição pela crença de que o diagnóstico está relacionado à dor, a tratamentos invasivos e à morte (ANDRADE *et al.,* 2020, p. 5882). Essa doença é percebida com medo e preocupação pelas pacientes por estar ligada a sentimentos de incapacidade, incurabilidade e fatalidade, sendo também temida e estigmatizada tanto pela paciente como por sua família (RIBEIRO *et al.,* 2019). Nesse sentido, as pacientes tendem a desenvolver estratégias cognitivas para lidar com emoções e sentimentos que desenvolvem a partir da descoberta e no tratamento da doença. *Coping* refere-se à resposta emocional, cognitiva ou comportamental ao estresse ou a uma crise (GUSTAFFSON *et al.,* 2019, p. 2), isto é, o conjunto de esforços que a pessoa utiliza para confrontar determinada situação geradora de estresse (STUMM *et al.,* 2009). As estratégias para lidar com situações difíceis podem variar entre dor, sofrimento, negação, medo, sublimação, aceitação (STUMM *et al.,* 2009). Essas estratégias são também conhecidas como

mecanismos de defesa (ANDRADE *et al.,* 2020), em que a paciente pode racionalizar as experiências, lutando para dar sentido a elas, por vezes sentindo medo do desconhecido, mas ao mesmo tempo tendendo a deslocar-se da realidade. Da mesma forma, podem buscar na religião (ANDRADE *et al.,* 2020; RIBEIRO *et al.,* 2019) um suporte, e até fazer um esforço para enfrentar a doença, o que provocaria uma sobrecarga nos seus recursos internos (ANDRADE *et al.,* 2020). Além disso, a ambivalência quanto à experiência surge como forma de mobilizar mecanismos internos para elaborar a experiência (ANDRADE *et al.,* 2020; RIBEIRO *et al.,* 2019).

Ao relatar a descoberta e o tratamento com o câncer de mama, pacientes podem elaborar essa experiência e concretizá-la por meio da linguagem. A emergência de metáforas conceptuais pode indicar algumas dessas estratégias de *coping*, e é para essa direção que nosso olhar se volta. O foco do nosso trabalho está nas narrativas da experiência com o câncer de mama, permeado pela compreensão de como metáforas conceptuais – nos termos de Lakoff e Johnson (1980, 1999) e Kövecses (2010, 2011) – emergem no discurso dessas mulheres. Entendemos que a forma como nos expressamos e trazemos à tona determinados domínios metafóricos refletem estruturas do nosso sistema conceptual (LAKOFF; JOHNSON, 1999). Ao examinar a linguagem usada na experiência com doenças consideradas tabus, como AIDS e câncer, Sontag (1979) traça considerações sobre o quanto se deveria expurgar as metáforas, já que, em suas palavras, "é quase impossível fixar residência no reino dos doentes sem ter sido previamente influenciada pelas metáforas lúgubres com que esse reino foi pintado" (*s.p.*). A autora busca elucidar as metáforas e libertar-se do seu jugo, talvez sem estar consciente de que metáforas são ubíquas em nossa estrutura conceptual e não podem ser erradicadas por vontade própria (DEMJEN; SEMINO, 2016). As metáforas são parte de nossas trajetórias individuais no mundo, e, ao emergirem pela nossa fala por meio dos signos linguísticos, dão indícios de como elaboramos nossa experiência. Nesse sentido, a ocorrência de uma metáfora pode ser uma janela para o plano conceptual: uma mulher com câncer, ao referir a si própria como uma guerreira, concretiza indícios de sua percepção sobre seu momento de vida: a de que precisa ser forte, a de que entende a sua vivência como uma luta – e isso pode ter nuances positivas ou negativas.

A fim de investigar como a emergência de metáforas conceptuais revela as experiências subjetivas nas narrativas de pacientes (que as

caracterizamos como "autoras") que desenvolveram e tratam o câncer de mama, buscamos adaptar um método, baseado em Linguística de Corpus (doravante, LC) e utilizado nas pesquisas em Terminologia para a identificação e extração de candidatos a termo, para a identificação dessas metáforas realizadas nos discursos das mulheres. Para Berber Sardinha (2000, 2004), a área busca coletar e explorar dados linguísticos textuais, coletados criteriosamente, com o propósito de servirem para pesquisa de uma língua ou variedade linguística. É por meio das ferramentas desenvolvidas para análise textual que se abrem possibilidades de explorar o *corpus* em estudo, constituído por postagens em blogs do que revelam as mulheres sobre sua experiência com o câncer quando escrevem sobre ele. Ao buscar apreender metáforas em *corpora*, partimos do princípio de que, assim como acontece com outras estruturas linguísticas, é possível identificar padrões de ocorrência também nas metáforas, tais como o uso recorrente de palavras de um determinado campo semântico para referir domínios conceptuais (por exemplo, 'luta', 'guerra', 'batalha', 'guerreira', 'heroína' [GUERRA]; ou, ainda, 'viagem', 'caminho', 'estrada', 'percurso' [VIAGEM]). A pesquisa em *corpora* permite, também, identificar os padrões de combinações (colocados) recorrentes nas metáforas (BERBER SARDINHA, 2006, 2007, 2012).

As técnicas de LC podem produzir fatos sobre a linguagem que poderiam, em situações cotidianas, permanecerem escondidas (DEIGNAN, 2008), isto é, poderiam passar despercebidas diante da leitura por olhos humanos apenas. Interessa-nos, nesta pesquisa, utilizar algumas técnicas estabelecidas e ajustá-las de modo específico, de tal forma que essas nos permitam "ver" estratégias de *coping* emergindo do discurso escrito das mulheres com câncer.

Subdividimos este texto em cinco partes. Na Seção 2, trazemos para a cena a teoria da metáfora conceptual, nosso aporte teórico e tratamos da relevância da Linguística de *Corpus* como ferramental que nos auxilia a direcionar nosso olhar sobre o *corpus*. Na Seção 3, delineamos a metodologia. Neste ponto, ressaltamos o valor da pergunta de pesquisa para que as ferramentas nos auxiliem a ver fenômenos realizados pela língua. Na Seção 4, discutimos os achados do *corpus*, que captura especificidades subjetivas que são apontadas pela presença de metáforas ao longo dos textos. Por fim, a Seção 5, das considerações finais, desenha reflexões sobre este processo de investigação e aponta para trabalhos futuros.

2 Metáforas conceptuais à luz da Linguística de Corpus

A metáfora conceptual é um fenômeno ubíquo na vida cotidiana, não apenas na linguagem, mas no pensamento e na ação (LAKOFF; JOHNSON, 1980). Nosso sistema conceptual é fundamentalmente metafórico e tem papel central em definir realidades. Ao mapear um domínio conceptual em termos de outro domínio (KÖVECSES, 2002, 2010; LAKOFF; JOHNSON, 1980), realizamos coerentemente a nossa própria experiência cotidiana, na maioria das vezes sem ter consciência da natureza metafórica do próprio pensamento.

Nesse sentido, ao pensar sobre o processo de adoecimento e tratamento de uma doença como o câncer de mama, uma paciente pode referir ao fim de uma etapa da seguinte forma: "[...] agora o que quero fazer é andar para frente!",[1] em que o elemento textual "andar para frente" sugere um mapeamento do domínio-fonte TRAJETÓRIA/VIAGEM,[2] com sentido mais concreto da experiência, para o domínio-alvo VIVER, de sentido mais abstrato. Aqui, neste caso, o sentido da metáfora VIVER É SEGUIR UMA TRAJETÓRIA,[3] ou VIVER É UMA VIAGEM pode ser explicitado como se a narradora perspectivasse, na sua trajetória de vida, um futuro em que ela poderia superar as questões daquele momento. Isto é: a partir desse mapeamento primeiro, pode-se assumir que, especificamente, ANDAR PARA FRENTE É SUPERAR poderia ser uma interpretação possível das experiências ou fenômenos particulares (LAKOFF; TURNER, 1989). Na perspectiva da Teoria da Metáfora Conceptual (LAKOFF; JOHNSON, 1980), CÂNCER É GUERRA, por exemplo, é um mapeamento do qual emergem expressões metafóricas como "(...) acompanhei sua *luta* contra um câncer, sua *batalha* pela vida". Nesse exemplo, chegamos à metáfora conceptual por causa de elementos como 'luta' e 'batalha'.

Metáforas conceptuais são elementos que emergem na e pela linguagem, e sua ubiquidade demonstra que é por meio delas que construímos e moldamos nossa realidade e elaboramos significados para eventos da vida cotidiana. Esse recurso cognitivo-conceptual comunica

[1] Os exemplos foram extraídos do *corpus* deste estudo. Foram apagados dados que pudessem identificar as autoras das postagens.

[2] Costumeiramente, utiliza-se o rótulo VIAGEM para um domínio conceptual que reflita movimentos de trajetória ou caminhar, por exemplo. Manteremos esta notação.

[3] A notação dos domínios e do mapeamento metafórico conceptual é realizada no formato VERSALETE.

sobre experiências novas, complexas, abstratas e sensíveis em termos de experiências mais familiares, mais simples e intersubjetivamente acessíveis (SEMINO *et al.,* 2018).

Hendricks *et al.* (2018) pesquisaram o papel das metáforas em moldar o modo como pacientes lidam com o câncer. Nos cinco experimentos conduzidos, encontraram que elaborar o *frame* da situação de uma pessoa com o câncer como uma "batalha" encorajaria as pessoas a acreditar que aquela pessoa tem maior tendência a sentir-se culpada se não se recuperar, do que quando a pessoa doente elabora um *frame* da mesma situação como "viagem". De outro modo, o *frame* para "viagem" tem maior tendência a encorajar a inferência que a pessoa pode fazer as pazes com sua situação do que com o *frame* de "batalha".

A pesquisa conduzida por Elena Semino no projeto "Metaphor, Cancer and the end-of-life" (SEMINO *et al.,* 2018), que analisa um corpus de 1,5 milhão de palavras, tem trazido importantes *insights* a respeito da metáfora como uma ferramenta linguística e cognitiva frequentemente usada para falar e para pensar sobre experiências sensíveis e subjetivas, tais como doença, emoções e morte. Para a pesquisadora e sua equipe, tais manifestações podem ajudar ou atrapalhar a comunicação em saúde e o bem estar da paciente, dependendo de como forem utilizadas. Note-se que metáforas diferentes podem ter vantagens e desvantagens diferentes; para Hendricks *et al.* (2018), não há metáfora perfeita para se falar sobre o câncer: certas metáforas têm probabilidade de ajudar alguns pacientes mais do que outros, dependendo de uma ampla variedade de características e suas experiências, assim como suas experiências pessoais. Os dados do *corpus* da pesquisa de Semino *et al.* (2015, 2018), sobre a escrita online de pacientes e de profissionais de saúde, demonstraram que "batalha" não era inerentemente ruim, nem "viagem" era inerentemente bom para todos. Pessoas diferentes usavam cada uma das metáforas de modo empoderador e desencorajador. Não se encontrou evidência, nesses dados, de que metáforas ligadas a viagem podem ter efeitos potencialmente danosos que estão, às vezes, associados a metáforas de guerra. Compreender os impactos de diferentes metáforas, percebendo a valência (se mais positiva, se mais negativa, por exemplo) que assumem nos seus diferentes contextos pode ser uma ferramenta para que profissionais de saúde possam auxiliar os pacientes em situações específicas, com formas de pensar também particulares.

Gustaffson *et al.* (2019) analisaram um *corpus* de blogs escritos por pacientes suecos com câncer avançado, interpretando os achados a partir da análise de metáforas linguísticas. O estudo demonstra a intersecção entre a análise de metáforas e de estratégias de *coping*. Para os autores, três domínios conceptuais mais frequentes no *corpus* apresentaram diferenças de percepção: JORNADA/VIAGEM e PRISÃO aparentam ser compreendidos como mais flexíveis do que o domínio GUERRA em termos de *coping*.

A percepção de que há, para além do nível linguístico, metáforas conceptuais que permeiam a subjetividade de quem profere tais enunciados não é possível sem um olhar para o seu contexto de ocorrência. Assim, ao se proceder à leitura do texto como um todo, compreender como determinadas percepções estão daquele modo expostas, captar expressões que podem evocar expressões metafóricas pode ser chave para a interpretação de como a experiência é elaborada – no caso deste estudo, como o desenvolvimento e o tratamento da doença estão sendo elaborados e quais são as estratégias usadas para enfrentá-la. No entanto, realizar a leitura de um grande volume de posts de blogs diversos pode tomar um tempo significativo da pesquisa. A fim de otimizar a análise textual, ferramentas ligadas à área da LC são desenhadas para capturar a especificidade desse tipo de dado.

A LC, nesse sentido, pode ser uma metodologia eficaz na pesquisa de metáfora em *corpora*, visto que o seu objetivo é buscar o uso típico e habitual de formas linguísticas. Esse processo de busca através das possíveis expressões metafóricas pode nos conduzir às metáforas conceptuais (BERBER SARDINHA, 2007). Um aspecto fundamental nessa busca é o contexto de ocorrência da expressão. As palavras 'luta' e 'batalha', por exemplo, podem ter significados literais, além dos metafóricos, e apenas o contato com o contexto definirá se essas palavras são ou não metáforas. Para Deignan (2008), achados da pesquisa na área de LC indicam que metáforas linguísticas são determinadas pelo contexto, bem como pelo significado intencional do falante ou de quem escreve. Isso indica que a metáfora é um fenômeno textual, social, e cognitivo, e por isso deve ser observado levando em conta as suas dimensões contextuais. A pesquisa com *corpus* auxilia a visualizar padrões de metáforas linguísticas, haja vista que procede da acumulação de observações detalhadas da linguagem em uso para questões teóricas (DEIGNAN, 2008).

A LC é uma área da linguística aplicada que estuda a linguagem por meio da análise informatizada de *corpora* textuais. Um *corpus* é um conjunto de textos autênticos – ou seja, textos que não foram produzidos especialmente para fins de estudo (BERBER SARDINHA, 2000). Através da busca utilizando ferramentas informatizadas, também conhecidas como concordanciadores, é possível mapear expressões recorrentes no *corpus*, que podem servir como indício de expressões metafóricas. Note-se, aqui, que, sozinha, a ferramenta não revelará aspectos dos fenômenos a serem analisados sem que tenhamos um objetivo nesse processamento: a busca utilizando apenas uma das funções (lista de palavras, lista de palavras-chave, concordanciador ou colocados) não trará resultados satisfatórios, dado que uma metáfora conceptual pode se realizar de diferentes formas no uso. Ademais, uma mesma expressão pode ser utilizada de forma literal ou metafórica, dependendo do contexto – o que torna necessária a análise e interpretação dos dados por parte das pesquisadoras. Desse modo, uma abordagem livre, sem uma pergunta objetiva, revelaria uma multiplicidade de aspectos que tornaria a análise impraticável, justamente pelo fato de haver possibilidades de interpretação as mais diversas. Portanto, a questão de pesquisa formulada a partir da temática do *corpus* – a saber: como a emergência de metáforas conceptuais revela as experiências subjetivas nas narrativas de pacientes que desenvolveram e tratam o câncer de mama? – serve como balizadora para direcionar nosso olhar.

Segundo Berber Sardinha (2007), a busca por expressões metafóricas normalmente envolve o uso da intuição e da memória, bem como o conhecimento prévio e teórico sobre o fenômeno pelo pesquisador. Há diversas maneiras de analisar metáforas em *corpora* eletrônicos, como a leitura do *corpus* de forma integral, a busca a partir da intuição e do conhecimento prévio, e a investigação através de lista de palavras e da ferramenta *concordance*, dois recursos que compõem *softwares* concordanciadores, como o AntConc (ANTHONY, 2019), por exemplo. De acordo com o autor, essas formas podem ser combinadas. "Nenhuma delas em si é suficiente para dar conta desse fenômeno" (p. 197). Já Stefanowitsch (2007) propõe uma abordagem baseada em *corpus* para a investigação de domínios-alvo metafóricos baseados na recuperação de itens lexicais representativos a partir do domínio-alvo e para a identificação das expressões metafóricas associadas a eles. Para o autor, essa abordagem de análise de metáforas é superior às que levam em conta a análise de dados a partir da coleta de citações de forma

eclética ou levando em conta a introspecção. Além disso, tal abordagem permite a quantificação da frequência de metáforas individuais. Nesta pesquisa, nós também partimos da observação de domínios-alvos e da lista de frequência para explorar o *corpus*, porém seguimos um percurso diferente, conforme explicitamos a seguir.

Neste trabalho, adaptamos um método de processamento e de análise do *corpus* textual a partir do uso da ferramenta AntConc (ANTHONY, 2019), que nos possibilita olhar para o *corpus* constituído pelas postagens em blogs em busca de indícios de construções que pudessem nos remeter a possíveis metáforas. Essas são localizadas em seus contextos de forma a nos auxiliar a compreender como elaboram as mulheres a sua vivência com o câncer. Exploraremos essa proposta na próxima seção.

3 Percurso metodológico

Nossa pesquisa está baseada na análise qualitativa de um *corpus* constituído por 2.036 textos coletados na íntegra em 31 blogs de livre acesso, contendo relatos de mulheres a respeito do desenvolvimento e do tratamento do câncer de mama. Pelo fato de utilizarmos dados abertos da internet, houve a preocupação de seguir alguns princípios éticos básicos no tratamento dos dados textuais. Nesse sentido, informações sensíveis, que pudessem identificar as autoras dos posts, foram apagadas, tais como nomes próprios e a identificação do blog.

A análise preliminar, de caráter quantitativo, demonstrou que esses textos totalizam 303.088 palavras (tokens) e 31.343 types. Nesse cálculo, não foram consideradas as palavras gramaticais constantes da *stoplist*.[4] Dado o objetivo proposto, consideramos que tal amostragem cumpre com os requisitos de representatividade, uma vez que o *corpus* é composto por textos de autoria diversa (multiautoral), em que cada blog corresponde a uma autora, e os textos coletados compreendem períodos distintos de tempo, de acordo com aqueles em que cada autora esteve em tratamento – o que pode nos mostrar se houve, ou não, variação temporal no tipo de metáfora produzida pelas autoras (ALUÍSIO; ALMEIDA, 2006; BERBER SARDINHA, 2000; BIBER, 1993).

[4] A stoplist foi baixada no endereço http://miningtext.blogspot.com/2008/11/listas-de-stopwords-stoplist-portugues.html e editada pelas pesquisadoras.

O concordanciador utilizado em todas as análises por máquina foi o AntConc (ANTHONY, 2019), e as ferramentas mais utilizadas em nossa pesquisa foram a Wordlist (gerador de lista de palavras do *corpus*), a Concordance (gerador de concordâncias), a File View, que permite ver o contexto de ocorrência ampliado no texto, e a Keyword List (gerador de palavras-chave).

Após a compilação e análise preliminar do *corpus*, optamos por realizar um estudo de caso com base em um recorte constituído por um dos blogs, denominado pelas pesquisadoras de "Caraca",[5] no qual foram publicadas 23 postagens individuais sobre as vivências relativas ao câncer de mama. Tal estudo de caso foi realizado com o objetivo de verificar se a ferramenta escolhida e a metodologia elaborada seriam adequadas à pesquisa em questão. Desse modo, a análise procurou encontrar, por meio das buscas nas listas de palavras e de concordâncias, domínios conceptuais realizados discursivamente, e serviu como guia para a análise manual, na qual cada uma das pesquisadoras leu os 23 textos, a fim de identificar possíveis metáforas. A essa rodada de análises individuais, seguiu-se a comparação e discussão dos resultados, a fim de mapear os domínios mais frequentes.

Porém, não seria viável aplicar essa metodologia à análise do *corpus* completo, uma vez que a leitura detalhada de 2.036 textos pelas quatro pesquisadoras individualmente demandaria muito tempo. Da mesma forma, a busca manual por possíveis metáforas em uma lista de mais de 31.000 palavras mostrou-se inviável. Assim, optamos por alimentar o AntConc com a lista de palavras do *corpus* piloto (23 textos, 8.675 tokens, 3.143 types), e, utilizando a ferramenta Keyword List, gerar uma lista de palavras-chave.

A Keyword List é uma ferramenta utilizada com frequência na pesquisa em Terminologia (FINATTO *et al.,* 2015; JESUS *et al.,* 2017; PAIVA *et al.,* 2008), utilizada para extrair em um *corpus* os candidatos a termo. Essa ferramenta compara o *corpus* principal com um de referência e, a partir daí, utilizando critérios estatísticos, gera uma lista de palavras que são particularmente frequentes no *corpus* principal, e que podem nos dar pistas sobre os elementos discursivos utilizados.

[5] Os títulos para cada conjunto de textos coletados fazem referência a uma palavra-chave eleita para nomear cada blog. Neste caso, o blog principal recebeu este título.

Para gerar a Keyword List no AntConc, utilizamos a seguinte configuração:

1. Geramos e exportamos, em formato .txt, uma lista de palavras do *subcorpus* "Caraca" (WordList_Caraca.txt), que seria utilizada como *corpus* de referência.
2. No Menu "Settings", selecionar "Tool Preferences".
3. Em "WordList", carregar a stoplist (Use a stoplist below -> Add words from file -> Open).
4. Em Keyword List, carregar o corpus de referência (Use raw files -> Add file -> Load -> Apply)
5. Carregar o *corpus* principal, clicando em "File" e, depois em "Open Dir".
6. Na aba Wordlist, gerar a lista de palavras do *corpus* principal.
7. Na aba Keyword List, gerar a lista de palavras-chave.

Como resultado, o AntConc apresentou uma lista de 94 *types* mais frequentes (com 62.752 *tokens*), cujos contextos de ocorrência analisamos manualmente, com o objetivo de identificar itens lexicais que poderiam indicar potenciais metáforas e seus domínios fonte e alvo. Esse conjunto de palavras foi dividido entre as quatro pesquisadoras para leitura, de acordo com os critérios de frequência. Assim, a pesquisadora que deveria analisar os *types* mais frequentes da lista acabou ficando com um menor número; e a pesquisadora que ficou com os *types* menos frequentes, teria um maior número de *types* para analisar, de forma que, ao final, cada pesquisadora deveria percorrer em torno de 15.000 ocorrências ou *tokens*. Dentre os 94 *types*, a divisão de análise ficou desta forma:

Pesquisadora 1: types 1 a 9;
Pesquisadora 2: types 10 a 27;
Pesquisadora 3: types 28 a 53;
Pesquisadora 4: types 54 a 94.

Ao longo dessa tarefa, foram elaborados quadros para registro dos domínios mais frequentes, sendo anotadas: *palavra-veículo* (ex.: 'tempestade'), isto é, o elemento-chave que é metaforizado (nos termos

de Cameron, 2003); o *exemplo-ocorrência* (ex. "Na verdade como sabe e acompanhou minha 'Tempestade' não é fácil"; "Dois anos e meio praticamente do primeiro 'turbilhão', quando começava a ensaiar meus primeiros passos novamente, veio a segunda tempestade, lá estava ele: o câncer de mama atrevido!"); e a *metáfora subjacente* (ADOECIMENTO É TEMPESTADE; DIAGNÓSTICO/RECIDIVA É TEMPESTADE), anotada conforme o contexto de ocorrência.

Na próxima seção, relataremos e discutiremos os resultados desse percurso analítico, numa tentativa de compreender as estratégias de enfrentamento à doença anunciadas e concretizadas pela língua.

4 Resultados e discussão

Nesta seção, descrevemos, a partir da análise dos veículos, isto é, das palavras que fazem emergir a metáfora subjacente, os domínios conceptuais mais frequentes, os principais exemplos ligados a esses domínios e as metáforas conceptuais subjacentes. Esses aspectos estão distribuídos em seis quadros,[6] os quais são comentados em termos de percepções e estratégias de *coping* pelas autoras dos blogs, bem como se essas refletem valores semânticos que tendem a ser mais positivos ou negativos (que chamaremos aqui de valências), e a função cognitiva da metáfora que se explicita nessa escrita: se a metáfora subjacente é *ontológica* ou *estrutural*, isto é, se as experiências são compreendidas em termos de entidades ou substâncias ou se um conceito é estruturado em termos de outro, respectivamente (LAKOFF; JOHNSON, 1980). Adiantamos que não foram salientados exemplos de metáforas *orientacionais*, as quais organizam um sistema todo em relação a um outro sistema (por exemplo, de orientação espacial).

4.1 ENTIDADE: "A doença chegou sem avisar"

Em relação ao domínio ENTIDADE, encontramos 88 usos metafóricos a partir de 30 veículos, sendo que 'doença' e 'câncer' foram as de maior frequência e as metáforas subjacentes identificadas para mapeamento entre doença/câncer e entidade/personificação poderiam ser

[6] Os quadros referem-se aos seguintes domínios: ENTIDADE, FORÇA DA NATUREZA, JOGO, CONTAINER, VIAGEM e GUERRA.

representadas por DOENÇA É PESSOA e CÂNCER É PESSOA. Exemplos disso estão em sentenças como "Não se pode ficar esperando… e a doença avançando sem dó nem piedade", ou "Eu gosto de repetir a palavra câncer muitas vezes ao dia, até ele encher o saco e ir embora". Ambos os mapeamentos indicam a percepção das autoras de que a doença e o câncer, por mais que fisicamente estejam instalados em seu corpo, sugerem ser algo externo a si, como uma entidade que surge e tem características e vontades próprias, que comandam esse corpo. Trata-se de metáforas ontológicas, cuja função cognitiva é prover um status existencial, ontológico ao domínio-alvo (KÖVECSES, 2006) – neste caso, DOENÇA e CÂNCER). Como resultado, fenômenos intangíveis se tornam elementos metafóricos: a doença se personifica, adquirindo um status existencial. Alguns dos exemplos mais representativos estão expressos no Quadro 1.

QUADRO 1 – Câncer é uma entidade com vontade própria

DOMÍNIO: ENTIDADE		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Doença	"Porque o doencinha maledita. Rouba além de nossa saúde, nosso sossego, nossa vida, nosso sono"	DOENÇA É PESSOA QUE ROUBA
	"Não se pode ficar esperando… e a doença avançando sem dó nem piedade"	DOENÇA É PESSOA QUE AVANÇA DOENÇA É PESSOA QUE NÃO TEM DÓ
	A doença chegou sem avisar	DOENÇA É PESSOA QUE CHEGA
	Agradeço a Deus pela doença e por tudo que ganhei com ela	DOENÇA É PESSOA QUE DÁ
	"e quando os cabelos caem, mais ou menos 21 dias após a primeira quimio, é como se a doença estivesse te afrontando, te dizendo: Oi eu to aqui, ela diz isso pra você a para sociedade"	DOENÇA É PESSOA QUE AFRONTA
	"A doença me ensinou a valorizar os verdadeiros amigos, poucos que já existiam e muitos que tive o privilégio de passarem a fazer parte da minha nova vida"	DOENÇA É PESSOA QUE ENSINA
	só que novamente a doença me surpreendeu em outras partes do meu corpo, como seu eu já tivesse perdido controle sobre ele, ela tentava me dominar.	DOENÇA É PESSOA QUE DOMINA
	"isso possibilitou a retirada total do tumor e o processo ao qual estou sendo submetida é para prevenir que a doença volte"	DOENÇA É PESSOA QUE PODE VOLTAR
	"Essa doença é tão prepotente rs.... ela ataca justamente nossas células"	DOENÇA É PESSOA PREPOTENTE

Câncer	"por alguma razão tememos muito ter que sair de nossa zona de conforto e o câncer sabe melhor do que qualquer outra coisa arrancar a gente de lá!"	CÂNCER É PESSOA QUE ARRANCA DA ZONA DE CONFORTO
	"Acho que durante muito tempo o câncer vai andar de mãos dadas com as minhas decisões mas uma coisa eu devo agradecer, ele me tirou da zona de conforto e quer saber? Ta bom assim!"	CÂNCER É PESSOA QUE ANDA JUNTO
	"Tive que esperar ele crescer, mandar metástase e ficar cada vez mais fortinho"	CÂNCER É PESSOA QUE CRESCE E SE FORTALECE
	"Eu gosto de repetir a palavra câncer muitas vezes ao dia, até ele encher o saco e ir embora"	CÂNCER É PESSOA QUE VAI EMBORA

Fonte: Elaboração própria.

O domínio ENTIDADE indica que, como forma de enfrentamento, o câncer é concretizado como algo externo à pessoa, com o qual é preciso conviver e contra o qual se deve lutar. Estar acometida por um tumor significa ter um corpo estranho dentro de si, sobre o qual a pessoa doente não tem controle, tanto em relação a ações quanto a formas de comportamento. Percebe-se uma dificuldade de olhar para o câncer como algo que faz parte de si, já que é caracterizado como agressivo e causador de grandes males. Tornar, portanto, o câncer uma entidade externa, personificando-o metaforicamente, faz com que a pessoa em tratamento consiga se relacionar com essa parte de si e dar conta da demanda emocional suscitada por essas vivências. Esses processos sugerem uma forma de enfrentamento centrada no problema, tendo como objetivo alterar a situação que é fonte original do estresse, em que o câncer é uma entidade com a qual se deve lutar e sobre a qual se pode agir ativamente (COSTA; LEITE, 2009). Dentro desse domínio, é possível notar que as valências referentes à entidade doença, que pode ser vista como algo agressivo, prepotente, que traz o mal, que afronta e que é traiçoeira, por vezes é percebida como um ser que ensina, que muda os rumos de uma vida para melhor e que proporciona reflexões importantes. Nesse sentido, é o contexto que modula o sentido do conceito subjacente ao domínio-alvo.

Um veículo metafórico que chamou atenção foi 'fantasma', sugerindo ligação da doença com os desígnios de uma entidade ligada ao plano do sobrenatural, ou do desconhecido. Os exemplos encontrados sugerem uma valência negativa, posto que o desconhecido pode suscitar medo: "mas TODAS...sem exceção, vivem com o fantasma da doença, com os medos as angústias,", "confesso que até hoje não me acostumei

com tantas mudanças e com o fantasma da recidiva.", "O maior desafio é conviver com o fantasma de uma possível volta da doença maldita.", "pois além do medo do fantasma do câncer, ainda temos todos os medos que todos os seres humanos". Aqui, temos a construção 'fantasma d*' e o complemento 'doença', 'recidiva', 'câncer', sugerindo uma atribuição de sentido adicional a esses domínios-alvo.

4.2 FORÇA DA NATUREZA: "A sensação é de se estar numa jangada no meio de uma tormenta"

O domínio FORÇA DA NATUREZA é representativo no que tange às formas pelas quais o processo de adoecimento é percebido pelas mulheres e como essas metáforas representam uma forma de concretizar a experiência subjetiva e abstrata. Nesse domínio, foram identificadas 24 usos metafóricos identificados, a partir de 11 veículos. As metáforas conceptuais mais frequentes foram ADOECIMENTO/TRATAMENTO É TEMPESTADE; TRATAMENTO É TSUNAMI; DIAGNÓSTICO/TRATAMENTO É FURACÃO; e CÂNCER/TRATAMENTO É TORMENTA. A exemplo dos mapeamentos envolvendo o domínio ENTIDADE, anteriormente explorados, aqui temos também metáforas ontológicas, que representam a perspectiva das autoras sobre o processo pelos quais passam, isto é, uma forma de conceber um evento da vida, ou as emoções relacionadas a ele, como uma entidade. O Quadro 2 sintetiza alguns dos exemplos representativos desse mapeamento.

QUADRO 2 – Câncer é uma força da natureza

DOMÍNIO: FORÇA DA NATUREZA		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Tempestade	"Na verdade como sabe e acompanhou minha 'Tempestade' não é fácil"	ADOECIMENTO É TEMPESTADE
	"Dois anos e meio praticamente do primeiro 'turbilhão', quando começava a ensaiar meus primeiros passos novamente, veio a segunda tempestade, lá estava ele: o câncer de mama atrevido!"	CÂNCER É TEMPESTADE DIAGNÓSTICO/ RECIDIVA É TEMPESTADE
	"Abre o guarda-chuva e volta pra tempestade, porque ela vai passar."	TRATAMENTO É TEMPESTADE
Tsunami	"Agora entro numa nova fase, a fase de voltar a retomar o que ficou deixado de lado enquanto o tsunami batia"	TRATAMENTO É TSUNAMI

Furacão	"E se posso dar uma dica para quem está no meio do furacão do diagnóstico/tratamento de um câncer, vai essa: escreva"	DIAGNÓSTICO/ TRATAMENTO É FURACÃO
	"Desejo a todos que estão ainda no meio do furacão, fazendo quimio, radio ou que descobriram agora que estão com câncer muita força, coragem e fé, que Deus esteja com todos vocês"	TRATAMENTO É FURACÃO
Tormenta	"Olhando daqui, me surpreendo de ter saído de novo da tormenta, não sem me molhar, mas saí!"	TRATAMENTO É TORMENTA
	"A sensação é de se estar numa jangada no meio de uma tormenta"	TER CÂNCER É TORMENTA

Fonte: Elaboração própria.

Esse domínio indica a intensidade do impacto do diagnóstico e do tratamento na vida dessas mulheres. Tempestades, furacões, tsunamis e tormentas são fenômenos da natureza que causam destruição, desestruturação e medo por onde passam. Sugerem a sensação de desestruturação e sobrecarga psíquica sentida pelas mulheres nesse período, no qual existe um rompimento com a sensação de controle sobre a vida e a imposição de uma necessidade de reorganizar o viver. A necessidade de fazer uso de mecanismos psíquicos de enfrentamento se dá quando a pessoa doente está sobrecarregada pelas demandas emocionais frente a uma realidade que desequilibra a homeostase do psiquismo, causando estresse e desacomodação (ANDRADE *et al*., 2020; COSTA; LEITE, 2009). Essa categoria indica o uso de enfrentamento centrado na emoção, com uma mobilização interna do indivíduo para tentar regular o estado emocional associado ao estressor (COSTA; LEITE, 2009). Além disso, percebe-se, nos exemplos elencados, uma valência predominantemente negativa, cuja intensidade relacionada às forças da natureza sugeridas pelos domínios-fonte é expressa como a forma como o processo de adoecimento é percebido.

4.3 JOGO: "O tratamento é feito de etapas, cada uma, uma vitória e um novo recomeçar"

Outro domínio recorrente no corpus foi o do JOGO, com 26 usos metafóricos identificados no *corpus* a partir de 11 veículos. Neste, os veículos mais produtivos foram os seguintes: 'brincar', 'etapa', 'fase', 'regra', 'torcer', 'torcida' e 'vitória'. As metáforas identificadas foram A VIDA É UM JOGO; A DOENÇA É UM JOGO; e O TRATAMENTO É UM JOGO, como é possível ver nos exemplos do Quadro 3. Aqui, a ocorrência desta metáfora

estrutural funciona para compreender conceitos abstratos, VIDA, por meio de outro conceito mais concreto, JOGO.

QUADRO 3 – A vida/O câncer/O tratamento contra o câncer é um jogo

DOMÍNIO: JOGO		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Brinc*	"Num quero brincar mais disso não"	A DOENÇA É UM JOGO
	"Não tem jeito, é uma dança com a morte, um bailar silencioso....a gente brinca com a morte o tempo todo."	
Etapa*	"O tratamento é feito de etapas, cada uma, uma vitória e um novo recomeçar. Isso significa que estou mais perto da minha cirurgia de reconstrução definitiva da mama, em que vou passar para a segunda etapa."	O TRATAMENTO É UM JOGO
	"2012 foi um ano de recuperação, recuperei o fôlego para começar a próxima etapa, a da reconstrução e trabalhei"	
	"Mais uma etapa vencida e vamos que vamos confiantes, amém!!!"	
	"Agora se inicia uma nova etapa, o início do fim para um novo início"	
Regr*	"permiti que o medo e o meu psicológico abalado ditassem as regras do jogo."	A VIDA É UM JOGO
	"siga algumas regrinhas de ouro da sua nova vida"	
Torce*	"Daqueles que torcem pela sua vida."	O TRATAMENTO É UM JOGO
	"Torcem aí galeraaaa!!!! Para minhas taxas aumentarem estou alimentando super bem, tendo cuidados com friagem (clima)..."	
	"R. é uma pessoa admirável e sei que torceu muito pela minha recuperação, obrigada"	
Torcida	"Obrigada, mil vezes obrigada pela torcida e já peço para todas rezarem por mim!"	O TRATAMENTO É UM JOGO
Vitória	"O simples fato de tentar de novo já será sua primeira vitória."	A VIDA É UM JOGO O TRATAMENTO É UM JOGO
	"milagres na minha recuperação teve , tumor sumiu isso é VITÓRIA."	
Fase	"Mas antes de tudo dar certo tenho que passar nesta primeira fase de descobrir a doença novamente e acreditar que o plano aprovará a quimioterapia"	O TRATAMENTO É UM JOGO
	"E acho que essa fase cirúrgica acabou."	
	"o quanto seria importante nessa fase da minha vida"	
	"senso de humor para levar esta fase da vida numa boa"	

Fonte: Elaboração própria.

Ao referir à vida e ao tratamento como um JOGO, com etapas, fases, regras, torcida, a paciente parece traçar um paralelo desse momento da vida com o câncer, com um processo com mais elementos positivos do que negativos. É como se, nestes relatos, as autoras administrassem as demandas surgidas com a doença de forma a compreender e aceitar cada etapa (cf. ANDRADE *et al.,* 2020).

4.4 CONTAINER: "[...] mas no meu coração não tinha espaço pra medo"

Para o domínio CONTAINER, identificamos 34 usos metafóricos a partir de 21 veículos, dos quais os mais produtivos foram: 'vida', 'transbordar', 'encher' ('cheio'), 'esvaziar' ('vazio'), 'jorrar', 'despejar', 'corpo', 'dentro' e 'fora'. As metáforas identificadas foram CORPO É CONTAINER, PESSOA É CONTAINER, CORAÇÃO É CONTAINER e MENTE É CONTAINER, como é possível ver nos exemplos a seguir (QUADRO 4). Essa metáfora, de caráter ontológico, é compreendida como realização de que um elemento de domínios do CORPO e da PESSOA, visualmente perceptíveis, e dos domínios da MENTE e do CORAÇÃO, relacionados a fenômenos intangíveis, se tornem objetos metafóricos.

QUADRO 4 – A vida/O corpo/A mente é um container

DOMÍNIO: CONTAINER		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Vida	"gota a gota, nos preenche a vida"	VIDA É CONTAINER
	"Quero uma vida mais plena, com mais significado"	
Transbordar	"nossas almas irão transbordar de amor!"	PESSOA É CONTAINER
	"Não fiquei impaciente, pelo contrário, fiquei urgente. Urgente em não transbordar com bobagens."	
Esvaziadas	"todas as verdades que cultivamos e que às vezes estão esvaziadas de amor e concluo precisamos respeitar a amar todos sem distinção nenhuma."	VERDADE É CONTAINER
Cheio/cheia	"sempre com o coração cheio de amor e gratidão"	CORAÇÃO É CONTAINER
	"coração está cheio de alegrias, esperanças e amor!"	
	"com o coração cheio de amor"	
	"feliz natal e um ano novo cheio de alegrias"	ANO NOVO É CONTAINER
	"quem sabe sem sustos, ano novo cheio de fé, esperanças"	
	"Logo estarei ansiosa, com medo e cheia de expectativas."	PESSOA É CONTAINER
	"Essa semana eu fiquei cheia de autopiedade, sofrendo com motivos (e muitos!), mas cheia de pena de mim mesma!"	

Para fora	"mas esse blog serve também para que eu coloque para fora o que me aflige a alma."	MENTE É CONTAINER
Jorrar	"faz nascer e jorrar de mim toda energia necessária para tocar a vida nos momentos mais difíceis do tratamento."	CORPO É CONTAINER
Dentro	"E aí, quando se faz o silêncio dentro, a gente começa a ouvir coisas que não ouvia."	MENTE É CONTAINER
Espaço	"mas no meu coração não tinha espaço pra medo."	CORAÇÃO É CONTAINER
Corpo	"e demora uns três meses para sair do corpo as drogas/químicas."	CORPO É CONTAINER
	"mostrar a cara da doença para fora do seu corpo"	
	"Lembrem-se que não guardar mágoas livra o corpo de muitas substâncias que fazem mal ao sistema imunológico e ao coração."	
	"quais as drogas que vão habitar o seu corpo durante um bom tempo e que infelizmente não dão onda nenhuma"	
Despejar	"AMIGAS DO BLOG AGRADEÇO IMENSAMENTE A VOCÊS, PORQUE QUANTAS VEZES VIM AQUI DESPEJEI LITERALMENTE MINHAS ANGUSTIAS"	BLOG É CONTAINER
Tirar	"faça o que for necessario para tirar a doença de mim"	CORPO É CONTAINER

Fonte: Elaboração própria.

Nestes exemplos, encontramos um domínio bastante comum, baseado na experiência corpórea, que leva à construção de mapeamentos metafóricos primários (GRADY, 1997a, 1997b). Se O CORPO É UM CONTAINER, emoções podem enchê-lo, esse corpo pode carregar a doença, medicamentos podem sair e entrar, a doença pode habitá-lo. Ao mesmo tempo, percebe-se que a estratégia de enfrentamento adotada em grande parte dos textos que contêm esse domínio é do enfrentamento focalizado na percepção das próprias emoções.

4.5 VIAGEM: "Morrer faz parte do percurso"

No domínio VIAGEM, foram encontrados 98 usos metafóricos a partir de 27 veículos, tais como 'vida', 'processo', 'etapa', 'caminho', 'frente', 'jornada', 'estrada', 'começo', 'recomeçar', 'passo', e de formas verbais como 'caminhar', 'andar', 'passar', 'chegar' e suas derivações. A partir disso identificamos, nos blogs, metáforas conceptuais estruturais relacionadas a esse domínio, como A VIDA É UMA VIAGEM/CAMINHADA/CAMINHO; TRATAMENTO É UMA VIAGEM/JORNADA/CAMINHO; e CÂNCER É UMA VIAGEM/ODISSEIA. Nos termos de Kövecses (2006), esses são exemplos de mapeamentos metafóricos estruturais, em que observamos que o *frame*

do domínio-fonte impõe certa estrutura ao *frame* do domínio-alvo em virtude dos mapeamentos que caracterizam a metáfora. O Quadro 5, a seguir, apresenta os veículos e os exemplos de ocorrências emergidas a partir desses mapeamentos:

QUADRO 5 – A vida/O Câncer/ O tratamento é uma viagem

DOMÍNIO: VIAGEM		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Vida	"estava nas andanças da vida"	A VIDA É UMA VIAGEM / CAMINHADA/ CAMINHO
	"Ainda nos encontraremos na estrada da vida"	
Caminh*	"Desculpem as reclamações..rs já deu para perceber que to meio caída né? Faz parte da caminhada humana."	TRATAMENTO É CAMINHADA/ CAMINHO
	"Isso tudo faz parte do caminho da cura"	
	"esta semana faço a quinta quimioterapia. Ufa... está passando do meio do caminho"	
Anda*	"Mas aconteceu e agora o que quero fazer é andar para frente!"	A VIDA É UMA VIAGEM/ CAMINHADA
	"Andar lado a lado com essa solidão requer um pouco de paciência e muito, mas muito amor."	
Frente	"Mas como estou vivendo um dia de cada vez, não quero pensar agora no que ainda virá pela frente."	
Jornada	"Quero que vocês que me acompanham nessa jornada"	O TRATAMENTO É UMA JORNADA
	"Mas no meio dessa jornada toda eu sempre acreditei que iria superar aquela loucura de sentimentos e emoções."	
Estrada	"Mas já tracei de novo e rumei para a estrada que me interessa: a vida!"	A VIDA É UMA VIAGEM/ CAMINHO
Acelerar	"E que a ansiedade é nossa inimiga número um, pois é ela que te agonia, que te faz acelerar as coisas, que faz com que você queira passar por tudo isso logo, acabar finalmente os tratamentos"	
Percurso	"Morrer faz parte do percurso"	
Rumo	"HOJE FAZ UM ANO QUE MINHA VIDA TOMOU UM RUMO QUE NUNCA IMAGINEI PRA MIM!"	
Corremos	"E assim retomaremos a nossa saúde para corrermos atrás da felicidade"	
Lado	"fico meio perdida, sem saber para que lado ir."	
Decorrer	"das que você já passou no decorrer da vida"	

Passar	"medo de voltar, de ter que passar por isso de novo, acho que todo mundo que passa deve sentir, ou não?"	O CÂNCER É UMA VIAGEM
	"Só quem passa por essa doença sabe da importância de todo esse apoio"	
	"Estou focada é na cura, passar bem pelo tratamento e crescer com tudo isso!"	TRATAMENTO É UMA VIAGEM
	"Antes de passar pela quimioterapia, achei que tudo poderia ser levado tranquilamente."	
Recomeçar	"graças a Deus é possível recomeçar de novo, fazendo diferente desta vez."	A VIDA É UMA CAMINHADA. RECOMEÇAR ALGO NA VIDA É RECOMEÇAR A CAMINHADA.
Passo	"Não tenha medo de viver, pois cada passo vacilante neste novo início, te dará a força necessária para as novas conquistas que o seu coração desejar"	A VIDA É UMA CAMINHADA
	"o importante é dar o primeiro passo, e acredito que para muitos esses encontros, a troca de experiência sejam o primeiro passo para aceitação e superação da doença".	
Chegar/ chegada	"essa era minha aparência antes do ca de mama, mas ainda chego lá de novo."	O CÂNCER É UMA VIAGEM
	"Sei que a chegada dessa notícia variou muito entre todas nós…"	
Fim do túnel	"Acho que é uma luz no fim do túnel, para quem depende de tratamentos pelo SUS."	TRATAMENTO É VIAGEM
Viagem	"Obrigada por me acompanhar nessa louca viagem."	A VIDA É UMA VIAGEM
Odisseia	"A odisséia havia começado alguns anos antes quando de um acidente banal de carro, o cinto de segurança resolveu me apontar o foco de quais seriam meus problemas futuros."	O CÂNCER É UMA ODISSEIA[7]
Trajetória	"Talvez seja uma das piores lembranças e dos piores momentos no decorrer de toda a trajetória da doença."	CÂNCER É UMA VIAGEM
Passageiro	"Sei que isso é passageiro e tal. Na maior parte do tempo eu penso assim....???"	A VIDA É UMA VIAGEM
Processo	"Que eu consiga entender que a vida é feita através da magia do "processo" e que saiba, então, respeitar o seu começo, o seu meio e o seu fim."	VIDA É TRAJETÓRIA

Fonte: Elaboração própria.

Diariamente falamos sobre a nossa vida em termos de viagem, em que há diferentes caminhos para serem tomados, diferentes destinos, paradas, obstáculos, passageiros etc. (LAKOFF; TURNER, 1989). Em

[7] Mapeamento caracterizado como uma extensão do mapeamento básico CÂNCER É UMA VIAGEM.

uma viagem, acontecimentos positivos e negativos podem acontecer, pode haver caminhos mais sinuosos e complicados ou trajetos mais simples, por exemplo. Todos esses aspectos da viagem são, normalmente, associados à nossa vida e com o que nela ocorre. Os exemplos do Quadro 5 indicam esse olhar das autoras para com sua vida e seus acontecimentos. Frases como "estava nas andanças da vida" e "Ainda nos encontraremos na estrada da vida" são atualizações que emergem da metáfora A VIDA É UMA VIAGEM/ CAMINHO.

Além da vida, o câncer e o seu tratamento também podem ser experienciados através desse mesmo grande domínio, conforme podemos ver no Quadro 5. "[...] medo de voltar, de ter que passar por isso de novo, acho que todo mundo que passa deve sentir, ou não?", "Talvez seja uma das piores lembranças e dos piores momentos no decorrer de toda a trajetória da doença.", "Isso tudo faz parte do caminho da cura", "Mas no meio dessa jornada toda eu sempre acreditei que iria superar aquela loucura de sentimentos e emoções." e "Acho que é uma luz no fim do túnel, para quem depende de tratamentos pelo SUS." são exemplos que emergem dos mapeamentos CÂNCER/TRATAMENTO É VIAGEM/CAMINHO/ TRAJETÓRIA. A partir deles, é possível notar que as autoras descrevem essa jornada pela doença e seu tratamento, às vezes, como uma experiência positiva, mas também como algo negativo. Os dois primeiros exemplos aqui apresentados carregam uma carga negativa da experiência de ter câncer e fazer o tratamento da doença. Já os seguintes demonstram uma valência positiva sobre como elas encaram ou enfrentaram o processo de tratamento.

Considerando essas relações e mapeamentos, é possível entender que, no percurso dessa doença e do seu tratamento, há diversos passageiros que viajam junto das pacientes, como as médicas e médicos, as enfermeiras e enfermeiros, os familiares, o grupo de amigos, por exemplo; há diferentes estações, que são as fases do tratamento e da doença; há muitos obstáculos no caminho, como os efeitos colaterais e os sentimentos negativos que as acompanham. Utilizar esse mapeamento em seu discurso, portanto, pode ser uma forma de externalizar esse processo e tudo que o envolve, e, consequentemente, enfrentá-lo, já que conseguem enxergar um ponto de partida e um destino a qual todas esperam chegar: a cura.

4.6 VALOR MONETÁRIO: "[...] como se fizéssemos uma poupança pela vida"

A partir dos 10 veículos 'vida', 'ganhar', 'dar', 'valor', 'investimento', 'saldo', 'doar', 'perder', 'dinheiro' e 'guardar', identificamos 28 usos metafóricos que fazem parte do domínio VALOR MONETÁRIO. Com base nisso, detectamos as metáforas conceptuais TEMPO É DINHEIRO; AMOR É DINHEIRO/VALOR MONETÁRIO; VIDA É DINHEIRO/VALOR MONETÁRIO; SAÚDE É VALOR MONETÁRIO; e MOMENTOS DA VIDA SÃO VALORES. O Quadro 6 explicita esses mapeamentos.

QUADRO 6 – Tempo/Vida/Amor é dinheiro

VALOR MONETÁRIO		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Vida	"como se fizéssemos uma poupança pela vida"	VIDA É VALOR MONETÁRIO
	"as coisas de outro modo dando valor a vida"	
Ganhar	"orgulho por não ter usado a doença para ganhar amor"	AMOR É VALOR MONETÁRIO
Preço	"muitas manifestações de carinho e amor que não tem preço"	
Dar	"Até emprestar um útero, doar um órgão, dar a própria vida se preciso for."	VIDA É DINHEIRO
	"Receber e dar amor."	AMOR É DINHEIRO
Valor	"Só assim para dar valor a coisas que não damos importância"	MOMENTOS DA VIDA SÃO VALORES
	"depois do câncer a gente aprende a dar valor as pequenas coisas, é dar valor a tudo mesmo"	
	"aprendemos a dar valor em nossa própria saúde"	SAÚDE É VALOR MONETÁRIO
Investi*	"Femana lança Outubro Rosa 2010 com foco em investimento na saúde"	
	"necessidade de se investir na saúde das mulheres brasileiras"	
saldo	"Não vamos nos despedir com lágrimas, mas com sorrisos de alegria de vitória e de missão cumprida, colocando na balança, sei que o saldo foi positivo."	VIDA É VALOR MONETÁRIO

Doar	"atendimento com nutricionistas, psicólogos, especialistas que doassem seu tempo a atender os pacientes, (...)"	TEMPO É DINHEIRO
	"Doe um tempo a outras pessoas."	
	"inclusive doações de tempo mesmo, sendo voluntário."	
Perder	"Mas, que bom que estou viva para poder perder esse tempo."	
	"o radioterapeuta não queria perder mais tempo."	
	"não perca tempo, cada segundo é valioso demais."	
Dinheiro	"Mas que o dinheiro e o tempo empregado tenham retorno para pessoas com câncer."	
Guardar	"Cuide-se, revitalize, guarde um tempo pra você, por que existem coisas que só você pode fazer por você mesma…"	

Fonte: Elaboração própria.

Neste quadro, encontramos um mapeamento metafórico bastante comum em nossa cultura, associado ao domínio do VALOR MONETÁRIO ou DINHEIRO. Normalmente falamos de elementos abstratos, difíceis de serem quantificados, em termos de dinheiro. Um deles é o tempo. Uma vez que, na cultura ocidental, o trabalho está tipicamente vinculado ao tempo que leva para ser realizado, e este é quantificado com precisão, associamo-lo ao dinheiro, que é mais facilmente mensurado. Nesse sentido, o tempo não é apenas um recurso limitado, mas também algo valioso, assim como o dinheiro. Consequentemente, é comum que usemos a experiência cotidiana com dinheiro para estruturar e compreender o conceito abstrato de TEMPO (LAKOFF; JOHNSON, 1980). Nesse sentido, mapeamentos metafóricos como esses têm caráter estrutural, em que o domínio-fonte, de sentido concreto, impõe sua estrutura para designar o domínio-alvo, que é mais abstrato. No Quadro 6, vemos que do mapeamento TEMPO É DINHEIRO emergem atualizações do tipo: "[..] o radioterapeuta não queria perder mais tempo.", "Cuide-se, revitalize, guarde um tempo pra você, por que existem coisas que só você pode fazer por você mesma…" e "Mas que o dinheiro e o tempo empregado tenham retorno para pessoas com câncer.". A partir desses exemplos, é possível ver como o tempo é algo muito valioso para as pessoas que estão em tratamento de uma doença, no sentido de que tudo o que é possível deva ser feito nesse período em busca da cura ou do melhor tratamento.

Outro emprego do domínio VALOR MONETÁRIO ou DINHEIRO é em relação aos sentimentos. Assim como o tempo, o amor é algo abstrato. Falamos em termos monetários sobre o amor. O Quadro 6 apresenta atualizações desse mapeamento, como "orgulho por não ter usado a

doença para ganhar amor" e "muitas manifestações de carinho e amor que não têm preço". Outras manifestações abstratas que se concretizam por meio desse domínio são momentos da vida, a vida em si e a saúde. Conforme o Quadro 6, atualizações comuns desses conceitos são "Só assim para dar valor a coisas que não damos importância", "[...] como se fizéssemos uma poupança pela vida" e "[...] necessidade de se investir na saúde das mulheres brasileiras".

4.7 GUERRA: "Nossa vida gira em torno da luta contra o câncer"

O domínio GUERRA foi um dos que apresentou maior número de ocorrências: foram 109 usos metafóricos identificados a partir de 37 veículos. Veículos como 'luta', 'batalha', 'risco', 'encarar', 'combater', 'atacar', 'vencer', 'trégua', 'aliada' e 'mutilada' nos levaram à identificação das metáforas CÂNCER É GUERRA/BATALHA, DOENÇA É GUERRA/BATALHA, TRATAMENTO É GUERRA/BATALHA. O câncer também é referido como o 'inimigo', o 'invasor', e os tratamentos, como as 'armas' utilizadas no combate à doença. Nesse cenário, a paciente é uma combatente ('guerreira') e a equipe de profissionais de saúde, um 'exército'.

QUADRO 7 – Câncer/tratamento é guerra

DOMÍNIO: GUERRA		
Keyword (veículo)	**Exemplo**	**Metáfora subjacente**
Vida	"Ao longo de um ano e dois meses acompanhei sua luta contra um câncer, sua batalha pela vida."	TER CÂNCER É LUTAR EM UMA BATALHA
	"Eu lutando pela vida e ela atentando contra a dela."	
	"Amigos tem muitos casos de câncer de mama consultório cheio, todas ali lutando pela vida ."	
	"Fiquei fazendo companhia para elas, pois elas estavam preocupadas e tristes dei muitos abraços e muita força e vou continuar rezando por todas que estão lutando pra sobreviver."	
Risco	"é um risco muito grande que corremos pois a quimioterapia acaba com a gente e nos deixa vulnerável"	FAZER O TRATAMENTO É ARRISCAR-SE
	"Também digo que há um risco grande de morte, uma vez que ele mata algumas coisas dentro de nós"	
	"Procurando na net, encontrei sobre o assunto, quem se interessar pode e deve ler, pra ficar bem ciente do risco que corremos."	
	"caso valesse aí sim justificaria correr tantos riscos de forma consciente...a quimio cura, mas ela mata também. Destrói as células doentes e as saudáveis também."	

Encarar	“na semana seguinte já estavam no centro cirúrgico e logo depois encara”ndo uma quimioterapia”	QUIMIOTERAPIA É BATALHA VIDA É BATALHA
	“pois encarar a quimioterapia é um desafio”	
	“Diante dessas perdas se faz necessário encarar a vida de outra forma,”	
	“Tenho o desafio de encarar uma vida mais simples.”	
	“todos encaramos cirurgias, quimio, efeitos colaterais e afins”	
Combater	“Onde já se viu usar argila como forma de combate ao câncer, em vez da quimioterapia?”	TRATAMENTO É COMBATE
	“Os remédios usados na quimioterapia para combater as células doentes”	
	“porque to sem idéias pra post e como fiz quimio na sexta, estou meio que fora de combate”	
	“se caso os nodulos forem a doença, já estará sendo combatido pela quimio”	
	“A nossa mente sempre está ocupado, o nosso corpo sempre sendo utilizado pra combater essa doença e o nosso coração sempre apreensivo com tudo!!”	
Atacar	“é um dos efeitos colaterais da quimioterapia, pois ela ataca as células que estão crescendo ativamente”	TRATAMENTO É ATAQUE
	“É evidente que quanto mais cedo essa doença for atacada, maior a chance de cura”	
Lutando	“lutando contra as mazelas da quimioterapia,”	TRATAMENTO É LUTA / DOENÇA É LUTA
	“Um beijo para todas mulheres guerreiras que estão lutando contra esta doença”	
	“peço força para continuar lutando contra esta maldita doença e por tudo que ela me trouxe de ruim.”	
Vencer	“acabar a quimioterapia é vencer a pior parte dessa batalha”	TRATAMENTO É BATALHA
Batalha	“acabar a quimioterapia é vencer a pior parte dessa batalha”	
	“Foi uma batalha e tanto, são sintomas dificílimos de explicar mas só quem já os sentiu pode entender o que se passa dentro de nós.”	
	“Fiz porque venci a guerra, e celebro a vitória da batalha, fiz porque outras guerras virão, sejam do tipo que for,”	
	“Só mulheres fortes conseguem vencer essa árdua batalha!!”	
Guerreira	“Eu sei que o post ta confuso, confuso como eu, que não sei e nem quero ir contra mim e minha natureza de guerreira”	PACIENTES SÃO GUERREIRAS
	“Guerreiras de verdade não andam sozinhas!”	
	“Guerreiras não desistem facilmente.”	
	“Bem foi uma manhã, muito interessante, e repleta de mulheres corajosas e guerreiras.”	
	“Não tem como não emocionar em ver a luta de tantas mulheres fortes, guerreiras, batalhadoras que venceram esse temido câncer de mama.”	

Trégua	"não temos tido trégua faz tempo, estamos vindo de muitas lutas faz algum tempo e quando parece que as coisas vão melhorar lá vem outra"	VIDA É UMA BATALHA
	"tem aquelas horas que a alma silencia e a dor está lá, latente, sem dar trégua."	
Guerra	"E com certeza, desta guerra ambos sairão mais fortes e prontos para enfrentar o resto das suas existências."	TRATAMENTO É GUERRA
Golpe	"Foi um novo **golpe**, porém, graças a Deus, descoberto a tempo para um nova cura!"	DIAGNÓSTICO É UM GOLPE
Invasivo	"era um câncer invasivo e vários nódulos"	CÂNCER É INVASOR/ INIMIGO
	"É muito difícil de uma hora para outra saber que uma doença como essa está invadindo o corpo, sorrateiramente, traiçoeiramente."	
	"Não vamos nos cansar de agradecer a todos que rezaram para que essa doença que invadiu nossa casa fosse vencida."	
Aliado	"Então, não precisa ter medo desse grande aliado na luta contra o câncer!"	EXAMES SÃO ALIADOS DE LUTA
Mutilada	"As mulheres do mundo atual não guerreiam com armas e, quando mutilam os seios, é por causa da luta contra o câncer, defendendo a própria vida."	CÂNCER É BATALHA QUE CAUSA MUTILAÇÕES
	"mulheres que sofreram mutilação total ou parcial da mama decorrente de tratamento do câncer."	
	"Nesse tempo, eu fiquei careca, sofri a mutilação, o abandono dos amigos, o descaso do Sistema, dores físicas e psicológicas."	
	"Receber o resultado de um exame escrito "câncer" (em geral vem o termo médico, né?) e ter que encarar uma mutilação e um tratamento que mais parece um veneno na veia não é o que chamo de lindo"	

Fonte: Elaboração própria.

Neste último quadro, encontramos um mapeamento metafórico bastante comum no cenário saúde/doença. DOENÇA É GUERRA; CÂNCER É GUERRA; TRATAMENTO É BATALHA são metáforas que constituem o mesmo *frame* semântico, o bélico. Há um entendimento de que a doença convida à luta, à reação, ao combate, ao movimento e que a ele é preciso ser, e estar, forte. O recurso de elaborar o próprio processo nesses termos é amplamente utilizado por pacientes que estão diante de uma doença que ainda carrega desconhecimento e tabus. Em consonância com os achados de Semino *et al.* (2015), metáforas que relacionam violência[8] a câncer foram encontradas com maior frequência no corpus MELC (Metaphors-in-

[8] Em Potts e Semino (2017), metáforas de violência são relatadas como metáforas ligadas ao domínio GUERRA.

the-end-of-life-care, [LANCASTER UNIVERSITY, [*s.d.*]), especialmente ao se "empoderar" ou ao "desencorajar" pacientes. Como já mencionado, ocorrências de metáforas neste campo semântico surgiram também em maior número no *corpus* deste estudo. Chama a atenção que uma estratégia de *coping* utilizada, na maior parte dos casos em que elementos bélicos emergem, é (re)afirmar positivamente a sua relação com o adoecer e com o tratamento, como em "Fiz porque venci a guerra, e celebro a vitória da batalha, fiz porque outras guerras virão, sejam do tipo que for,". Isso sugere que a paciente reage em busca de uma recuperação, mas ao mesmo tempo não se pode, com isso, interpretar que o processo foi elaborado de forma positiva, mas que ela busca demonstrá-lo dessa forma para quem a lê. Esse mesmo *frame* ganha conotações negativas em alguns casos, como em "Fiquei fazendo companhia para elas, pois elas estavam preocupadas e tristes dei muitos abraços e muita força e vou continuar rezando por todas que estão lutando pra sobreviver.", mas o que se percebe é, também, que as autoras que assim se expressam também reagem diante do adoecer e do tratamento, por pior que seja o momento.

Além desses domínios mais encontrados, chama a atenção a prevalência de menções a aspectos espirituais e/ou religiosos. O apego à espiritualidade e a práticas religiosas é uma estratégia de *coping* usual, de acordo com Andrade *et al.* (2020), Ribeiro *et al.* (2019) e Caetano *et al.* (2009). Tais modos de lidar com a doença e com o tratamento podem refletir esperança e a busca por suporte, que estão apoiados em crenças derivadas de comportamentos culturalmente estabelecidos. A religiosidade, de acordo com Ribeiro *et al.* (2019), favorece novo significado ao experienciar uma doença, mudando a forma como pessoas a percebem e promovendo alívio da dor e do estresse. As mesmas autoras também afirmam que bem-estar espiritual é considerado um fator de proteção, em que atitudes positivas em relação à doença são tomadas.

Curiosamente, palavras relacionadas à religiosidade ou à espiritualidade não se destacaram no processamento da lista de palavras no AntConc. Foi necessária discussão no grupo de pesquisadoras e breve revisão da literatura sobre estratégias de *coping* para que se tornasse à busca por novos elementos. Nesse sentido, exemplos adicionais foram localizados a partir de palavras como 'deus', cujos exemplos denotam a crença de uma entidade protetora e cuidadora: "A não ser esperar um milagre, e ela esperou e acreditou até domingo, dia 15 quando ele voltou para Deus."; "Houve um tempo em minha vida que briguei com Deus, chamei o para briga, tempo de raiva, (...)", "(...) ter um emprego que

eu fizesse meus horários e Deus colocou na minha vida uma vizinha maravilhosa" e "Terminamos nosso encontro com lágrimas nos olhos, sorriso nos lábios e a certeza de que Deus cuida de cada uma de nós, (...)".

Ligados ao campo da espiritualidade, também encontramos exemplos com a palavra 'anjo'/ 'anjo da guarda', que em geral é atribuída a alguma pessoa que dá suporte, ou esperança, à autora da postagem: "Ontem a noite meu anjo da guarda Dr Marcelo Bumlai foi me visitar no hospital e me dar alta."; "o oncologista não marcou biópsia dos nódulos e porque um anjo de Deus colocou um cirurgião plástico com ele na sala,"; "Ai eu já tava tão distraída com rotina do centro cirúrgico que fui na boa, um anjo de olhos azuis veio me buscar, Dr. Julio, me acalmou os ânimos, já que a veia puncionada"; "O anestesista também foi um anjo e me deu muita ajuda"; "Doutora Sara é um anjo, ela tem o dom de me deixar mais tranquila, e me recomendou voltar"; "Um enfermeira anja conhecida da minha manicure me salvou."

Essas evidências linguísticas da crença religiosa e/ou espiritual de cada autora mostram como elas buscam auxílio ou esperança em algo que vai além do nosso plano concreto. Através dos exemplos, percebe-se, de modo geral, que recorrer a algo divino, a uma força maior que a humana, parece trazer conforto e esperança nessa trajetória de descobertas sobre a doença e sobre o tratamento.

5 Considerações finais

O desenvolvimento de um percurso metodológico baseado em LC para identificar metáforas que pudessem evidenciar estratégias de enfrentamento ao câncer de mama permitiu um olhar apurado para os contextos de ocorrência dessas metáforas, tornando possível observar os elementos co-ocorrentes, bem como o tipo de vocabulário relacionado a cada domínio conceptual.

As análises a partir de domínios conceptuais mais frequentes demonstraram que as autoras tendem a se valer de metáforas para falar sobre suas experiências, sentimentos e momentos, expondo, assim, da forma mais concreta possível, o enfrentamento da doença e tudo o que está relacionado a isso. O uso efusivo de metáforas por essas mulheres pode ser considerado como uma forma de externalizar o que se passa nesse momento de suas vidas, tanto situações positivas quanto negativas, conseguindo, desse modo, compartilhar essa experiência com outras que estão vivendo o mesmo.

Ao nos depararmos com os domínios-fonte mais frequentes (VIAGEM, GUERRA, ENTIDADE, JOGO, VALOR MONETÁRIO, CONTAINER) e suas variações, percebemos a construção metafórica de metáforas de caráter estrutural e ontológico, o que demonstra a tendência a pensar que, se no primeiro caso um domínio conceptual pode se estruturar em termos de outro, no segundo caso vemos um grande número de experiências sendo compreendidas como pelas nossas experiências com substâncias ou objetos físicos. Além disso, os dados corroboram a afirmação de Semino *et al*. (2015, 2018), para as quais metáforas podem atuar de maneiras diversas para diferentes pessoas. Assim, avaliar o contexto de ocorrência desses domínios-fonte encontrados foi ponto crucial para compreender nuances de significação implícitas nessas narrativas.

Cabe notar que a alta frequência desses mesmos domínios conceptuais em narrativas de diferentes mulheres pode denotar que, apesar de falarem de si e de seu enfrentamento à doença de modos particulares, elas também assumem narrativas que podem ser reconhecidas discursiva e culturalmente como de quem lida com tais momentos de vida.

Frente ao contexto de adoecimento e de desorganização psíquica, as pacientes desenvolvem estratégias de enfrentamento, compreendidas como recursos psíquicos internos que são acionados para dar conta de situação que avalia como estressante e geradora de uma sobrecarga de demandas emocionais (COSTA; LEITE, 2009). A partir da investigação metodológica proposta e da análise apresentada, pode-se perceber a diversidade de significados e formas de representar o adoecer utilizados pelas pacientes. Antoniazzi *et al.* (1998) trazem que o *coping* é construído em um processo relacional entre pessoa-ambiente, influenciado por características individuais de personalidade bem como características situacionais, contextuais e sociais. Com isso, podemos refletir sobre como a narrativa envolvendo o câncer é construída e enunciada dentro de um contexto social. A narrativa construída frente o adoecer por câncer é múltipla e complexa, contudo, as narrativas analisadas neste trabalho conversam entre si e com a teorização de Susan Sontag (1979), que associa o câncer a morte, dor e sofrimento. A multiplicidade das metáforas encontradas e analisadas permite expandir essa construção associada ao câncer e compreender que é a partir dessa narrativa coletiva que as pacientes irão singularizar suas experiências e concordar, ou não, com o que é posto pela coletividade.

A identificação das estratégias de enfrentamento e implicação destas na forma como a pessoa se relaciona com a evolução clínica do

quadro é visto como recurso importante para profissionais da saúde (COSTA; LEITE, 2009). É a partir da construção social em torno do adoecer por câncer que cada paciente irá construir uma narrativa individual e subjetiva que lhe permita compreender e elaborar essas vivências. Não ter um olhar preconcebido para a forma como cada paciente irá atravessar seu processo de adoecimento e tratamento oportuniza que equipes de saúde promovam um cuidado humanizado, centrado no paciente e que busca atentar para a subjetividade emergente nas falas. A formação da equipe para uma escuta ativa e cuidadosa, que consiga ser sensível para metáforas – e para os possíveis significados ligados a elas – utilizadas pela paciente é capaz de promover um cuidado singular e que faça sentido para aquela que recebe essa assistência, que encontra a paciente na sua narrativa particular. Essa atenção ao discurso e, consequentemente, para a subjetividade dessa paciente, pode contribuir para o tratamento: o percurso, muitas vezes perpassado por dificuldades e vulnerabilidades, pode ser traçado em conjunto com uma equipe sensível a essas narrativas repletas de metáforas.

Contribuições das autoras

Todas as autoras participaram ativamente da concepção, redação e revisão deste estudo. A. R. Salgado elaborou a metodologia do estudo e participou da coleta dos dados e da interpretação dos resultados; A.A. Vanin participou da organização e da discussão dos resultados; G.H. Gomes fez a coleta e organização do corpus, participou das discussões, elaborou os resultados; L. Presotto participou da discussão dos resultados.

Nota

Este trabalho foi realizado apesar das dificuldades econômicas e do pouco investimento em ciência no Brasil. Ainda assim, resistiremos.

Referências

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. *Calidoscópio*, São Leopoldo, RS, v. 4, n. 3, p. 156-178, 2006. Disponível em: http://revistas.unisinos.br/index.php/calidoscopio/article/view/6002. Acesso em: 5 set. 2020.

ANDRADE, C. J.; GALHARDI, S. R. R. B.; AVOGLIA, H. R. C. Reações defensivas de pacientes em tratamento oncológico: análise das principais formas de enfrentamento. *Brazilian Journal of Health Review*, São José dos Pinhais, PR, v. 3, n. 3, p. 5881-5899, 2020. Disponível em: https://doi.org/10.34119/bjhrv3n3-149. Acesso em: 3 set. 2020.

ANTONIAZZI, A. S.; DELL'AGLIO, D. D.; BANDEIRA, D. R. O conceito de *coping*: uma revisão teórica. *Estudos de Psicologia*, Natal, v. 3, n. 2, p. 273-294, 1998.

ANTHONY, L. *AntConc* (Version 3.5.8) [Computer Software]. Tokyo: Waseda University, 2019. Disponível em: https://www.laurenceanthony.net/software. Acesso em: 5 set. 2020.

BERBER SARDINHA, T. Lingüística de Corpus: histórico e problemática. *DELTA*, São Paulo, v. 16, n. 2, p. 323-367, 2000. DOI: https://dx.doi.org/10.1590/S0102-44502000000200005. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502000000200005&lng=es&nrm=iso. Acesso em: 5 set. 2020.

BERBER SARDINHA, T. *Lingüística de Corpus*. Barueri: Editora Manole, 2004.

BERBER SARDINHA, T. Collocation Lists as Instruments for Metaphor Detection in *corpora*. *DELTA*, São Paulo, v. 22, n. 2, p. 249-274, 2006. DOI: https://doi.org/10.1590/S0102-44502006000200002. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502006000200002&lng=en&nrm=iso. Acesso em: 5 set. 2020.

BERBER SARDINHA, T. Análise de metáfora em corpora. *Ilha do Desterro*, Florianópolis, n. 52, p. 167-199, 2007. DOI: https://doi.org/10.5007/%x. Disponível em: https://periodicos.ufsc.br/index.php/desterro/article/view/11715/11005. Acesso em: 4 set. 2020.

BERBER SARDINHA, T. A ocorrência de metáforas é previsível? *Revista de Estudos da Linguagem,* Belo Horizonte, v. 20, n. 2, p. 211–240, 2012. DOI: https://doi.org/10.17851/2237-2083.20.2.211-240.

BIBER, D. Representativeness in *corpus* design. *Literary and Linguistic Computing*, Oxford, v. 8, p. 243-257, 1993.

CAETANO, E. A.; GRADIM, C. V. C. SANTOS, L. E. S. Câncer de mama: reações e enfrentamento ao receber o diagnóstico. *Revista Enfermagem UERJ*, Rio de Janeiro, v. 17, n. 2, p. 257-261, 2009.

CAMERON, L. *Metaphor in Educational Discourse*. London: Continuum, 2003.

COSTA, P.; LEITE, R. C. B. Estratégias de enfrentamento utilizadas pelos pacientes oncológicos submetidos a cirurgias mutiladoras. *Revista Brasileira de Cancerologia*, Rio de Janeiro, v. 55, n. 4, p. 355-364, 2009.

DEIGNAN, A. Corpus Linguistics and Metaphor. *In:* GIBBS Jr., R. W. *The Cambridge Handbook of Metaphor and Thought.* New York: Cambridge University Press, 2008. p. 280-294. DOI: https://doi.org/10.1017/CBO9780511816802.018

DEMJÉN, Z.; SEMINO, E. Using Metaphor in Healthcare: Physical Health. *In:* _____. (ed.) *The Routledge Handbook of Metaphor and Language.* London: Routledge, 2016. p. 285-399.

FINATTO, M. J. B.; LOPES, L.; CIULLA, A. Extração automática de candidatos a termo do "Curso de Linguística Geral" com apoio de recursos da Linguística de Corpus e do Processamento de Linguagem Natural. *Domínios de Linguagem*, Uberlândia, v. 9, n. 2, p. 40-55, 2015. DOI: https://doi.org/10.14393/DL18-v9n2a2015-4. Disponível em http://www.seer.ufu.br/index.php/dominiosdelinguagem/article/view/31077. Acesso em: 30 set. 2020.

GRADY, B. A. *Foundations of Meaning:* Primary Metaphors and Primary Scenes. 1997. 306f. Tese (Doctorof Philosophy) – University of California, Berkeley, 1997a.

GRADY, B. A. Theories Are Buildings Revisited. *Cognitive Linguistics*, [*S.l.*], v. 8, n. 4, p.267-290, 1997b.

GUSTAFSSON, A. W.; HOMMERBERG, C.; SANDGREN, A. Coping by Metaphors: The Versatile Function of Metaphors in Blogs about Living with Advanced Cancer. *Medical Humanities*, Cleveland, OH, v. 46, n. 3, p. 267-277, 2019. DOI: https://doi.org/10.1136/medhum-2019-011656.

HENDRICKS, R. K.; DEMJÉN, Z.; SEMINO, E.; BORODITSKY, L. Emotional Implications of Metaphor: Consequences of Metaphor Framing for Mindset about Cancer. *Metaphor & Symbol,* [*S.l.*], v. 33, n. 4, p. 267-279, 2018.

JESUS, S. M.; FERREIRA, M.; ESQUEDA, M. Pesquisa e prática terminológica bilíngue na formação do tradutor. *Cultura e Tradução*, João Pessoa, v. 5, n. 1, p. 53-62, 2017. Disponível em: https://periodicos.ufpb.br/ojs2/index.php/ct/article/view/38493. Acesso: 30 set. 2020.

KARKAR, A. J.; BURKE, L. M. "It's Your Loss": Making Loss One's Own Through Blog Narrative Practices. *Death Studies,* [S.l], v. 44, n. 4, p. 210-222, 2020.

KÖVECSES, Z. *Metaphor:* A Practical Introduction. New York: Oxford University Press, 2006.

KÖVECSES, Z. *Language, Mind and Culture:* A Practical Introduction. New York: Oxford University Press, 2002.

KÖVECSES, Z. *Metaphor.* A Practical Introduction. Second edition. Oxford: Oxford University Press, 2010.

KÖVECSES, Z. Recent Developments in Metaphor Theory: Are the New Views Rival Ones? *Review of Cognitive Linguistics*, [*S.l.*], v. 9, n. 1, p. 11-25, 2011.

LAKOFF, G.; JOHNSON, M. *Metaphors We Live By.* Chicago: The University of Chicago Press, 1980.

LAKOFF, G.; JOHNSON, M. *Philosophy in the Flesh:* The Embodied Mind and Its Challenge to Western Thought. New York: Basic Books, 1999.

LAKOFF, G.; M. TURNER. *More than Cool Reason:* A Field Guide to Poetic Metaphor. Chicago: Chicago University Press, 1989.

LANCASTER UNIVERSITY. *Metaphor, Cancer and the End of Life* – Research Portal. Lancaster: Lancaster University [s. d.]. Disponível em: http://www.research.lancs.ac.uk/portal/en/publications/metaphor-cancer-and-the-end-of-life(e0824059-b68f-442c-8ee7-49fb1b0526b0)/export.html. Acesso em: 3 set. 2020.

PAIVA, P. T. P.; CAMARGO, D. C.; XATARA, C. M. Uma reflexão sobre a elaboração de um léxico bilíngüe preliminar na subárea de cardiologia a partir de termos encontrados em um corpus paralelo e em dois corpora comparáveis. *DELTA*, São Paulo, v. 24, n. 1, p. 1-22, 2008. DOI: https://doi.org/10.1590/S0102-44502008000100001. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0102-44502008000100001&lng=pt&nrm=iso. Acesso em: 30 set. 2020.

POTTS, A.; SEMINO, E. Healthcare Professionals' Online Use of Violence Metaphors for Care at the End of Life in the US: A Corpus-Based Comparison with the UK. *Corpora*, Edinburg, v. 12, n. 1, p. 55-84, 2017. DOI: 10.3366/cor.2017.0109

PRIMO, A. F. T.; RECUERO, R. C. Hipertexto cooperativo: uma análise da escrita coletiva a partir dos Blogs e da Wikipédia. *Revista FAMECOS*, Porto Alegre, v. 10, n. 22, p. 54-65, 2008. DOI: https://doi.org/10.15448/1980-3729.2003.22.3235.

RIBEIRO, G. S.; CAMPOS, C. S.; ANJOS, C. C. Y. Espiritualidade e religião como recursos para o enfrentamento do câncer de mama. *Revista de Pesquisa: Cuidado é Fundamental,* Rio de Janeiro, v. 11, n. 4, p. 849-856, 2019.

SEMINO, E.; DEMJÉN, Z.; DEMMEN, J.; KOLLER, V.; PAYNE, S.; HARDIE, A.; RAYSON, P. The Online Use of Violence and Journey Metaphors by Patients with Cancer, as Compared with Health Professionals: A Mixed Methods Study. *BMJ Supportive & Palliative Care,* London, v. 7, n. 1, p. 60-66, 2015.

SEMINO, E.; DEMJÉN, S.; HARDIE, A; PAYNE, S; RAYSON, P. *Metaphor, Cancer and the End of Life:* A Corpus-Based Study. Routledge Advances in Corpus Linguistics. New York: Taylor & Francis, 2018.

SONTAG, S. *Doença como metáfora /* AIDS e suas metáforas. São Paulo: Companhia de Bolso, 1979.

STEFANOWITSCH, A. Words and Their Metaphors: A Corpus-Based Approach. *In*: STEFANOWITSCH, A.; GRIES, S. T. (ed.). *Corpus-Based Approaches to Metaphor and Metonymy.* Berlin: Mouton de Gruyter, 2007. p. 63-105.

STUMM, E. M. F.; MAÇALAI, C.; LEITE, M. T.; LORO, M. M. Mecanismos de *coping* utilizados por mulheres mastectomizadas para lidar com o estresse. *Scientia Medica*, Porto Alegre, v. 19, n. 3, p. 108-114, 2009.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1289-1324, 2021

Frequência e distribuição de plurais irregulares no Corpus Brasileiro

Frequency and distribution of irregular plurals in the Corpus Brasileiro

Luiz Carlos Schwindt
Universidade Federal do Rio Grande do Sul (UFRGS, CNPq), Porto Alegre, Rio Grande do Sul / Brasil
schwindt@ufrgs.br
https://orcid.org/0000-0003-0533-589X

Pedro Eugênio Gaggiola
Universidade Federal do Rio Grande do Sul (UFRGS, PROBIC-FAPERGS), Porto Alegre, Rio Grande do Sul / Brasil
pedroe.gaggiola@gmail.com
https://orcid.org/0000-0002-0123-6205

Isabela Prisco Petry
Universidade Federal do Rio Grande do Sul (UFRGS, PIBIC-CNPq), Porto Alegre, Rio Grande do Sul / Brasil
isabelappetry@gmail.com
https://orcid.org/0000-0003-1082-6589

Resumo: Neste texto aborda-se a frequência e a distribuição de formas de plural irregular do português brasileiro, no âmbito da palavra, numa perspectiva descritiva. Os dados provêm do Corpus Brasileiro e estão divididos em duas amostras: Amostra L, nomes pluralizados terminados ortograficamente em vogal+is (ex. papéis), vogal+us (ex. chapéus) e is (ex. funis), e Amostra N, os terminados por ões (ex. vilões), ãos (ex. irmãos) e ães (ex. pães). O exame das variáveis fonético-fonológicas e léxico-morfológicas –

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1289-1324

número de sílabas, acento, contexto fonológico, afiliação morfológica e frequência lexical – permitiu contextualizar o comportamento das alternantes minoritárias de cada amostra, em oposição às alternantes prevalentes, vogal+is e ões, respectivamente.

Palavras-chave: plural; alomorfia; morfologia; morfofonologia; Corpus Brasileiro.

Abstract: This paper addresses the frequency and distribution of Brazilian Portuguese irregular plurals, within the scope of the word, in a descriptive approach. The data come from the Corpus Brasileiro and are divided into two samples: (i) pluralized nouns ending, in spelling, with vowel+is (eg *papéis* 'papers'), vowel+us (eg *chapéus* 'hats'), and is (eg *funis* 'funnels'), and (ii) those ending with ões (eg *vilões* 'villains'), ãos (eg *irmãos* 'brothers'), and ães (eg *cães* 'dogs'). The phonological and lexical-morphological variables analyzed – number of syllables, stress, phonological context, morphological affiliation and lexical frequency – allowed to define the main contexts for the minority alternants of each sample, in opposition to the prevalent ones, vowel+is and ões, respectively.

Keywords: plural; allomorphy; morphology; morphophonology; Corpus Brasileiro.

Submetido em 09 de outubro de 2020

Aceito em 09 de novembro de 2020

1 Introdução

Neste artigo, tratamos da frequência e da distribuição das realizações de plural de nomes do português brasileiro (PB) terminados, em sua forma singular, nas sequências ortográficas vogal+u/l e ão, a partir de dados do Corpus Brasileiro.[1]

Por conveniência metodológica, as alternantes de plural são abordadas neste estudo a partir de uma tipologia que considera as rimas das sílabas que comportam a informação morfológica de plural (isto é, a porção fonológica que inclui o núcleo silábico e todos os segmentos que o sucedem). São **Vis, Vus** e **is**, para vogal+u/l, e **ões**, **ãos** e **ães**, para ão, como se exemplifica em (1) e (2), respectivamente.

[1] http://corpusbrasileiro.pucsp.br/cb/Inicial.html

(1)		**alternante**	**plural**		**singular**	
	a.	Vis	ped**ais**	[peˀdajs]	pedal	[peˈdaw] ~ [peˈdaɬ]
	b.	Vus	chap**éus**	[ʃaˈpɛws]	chapéu	[ʃaˈpɛw]
	c.	is	fun**is**	[fuˀnis]	funil	[fuˈniw] ~ [fuˀniɬ]

(2)		**alternante**	**plural**		**singular**	
	a.	ões	bal**ões**	[baˈlõjs]	balão	[baˈlãw̃][2]
	b.	ãos	irm**ãos**	[iɾˈmãw̃s]	irmão	[iɾˈmãw̃]
	c.	ães	capit**ães**	[kapiˈtãjs]	capitão	[kapiˈtãw̃]

Trata-se de uma abordagem essencialmente descritiva com potencial para contribuir, a partir da observação do léxico em uso, para o debate sobre o papel da produtividade na definição das restrições fonológicas e morfológicas que concorrem para seleção dessas formas não canônicas de plural na língua, as quais tratamos aqui como irregulares.

O comportamento de plurais irregulares em PB foi objeto de muitos estudos, em diferentes perspectivas, em especial no que diz respeito à origem do glide, nos dois casos mencionados, e à representação da nasalidade, no caso dos ditongos em *ão* (CAMARA JR., 1969, 1970; ABAURRE GNERRE, 1983; BISOL, 1998, 2016, 2020; WETZELS,

[2] Optamos neste texto, como estratégia de simplificação, por nos referirmos aos ditongos em análise por sua representação ortográfica. Nos exemplos de (2), no que diz respeito especificamente aos casos de ão e seus plurais, adotamos uma representação fonética que considera uma sequência de vogal e glide nasalizados. Sabemos, contudo, que a questão não é tácita entre fonólogos e foneticistas. Alguns estudiosos defendem a realização da consoante nasal em coda, plena ou secundária, seguindo esse glide. O português parece situar-se, em relação à pronúncia dessa nasal, numa posição intermediária, se considerarmos, num extremo, línguas como o inglês ou o espanhol, que no mais das vezes realizam a nasal plenamente em coda silábica (ex. bu[m] 'bum'; co[n], 'com'), e, noutro, línguas como o francês, que parece exibir uma assimilação completa da nasalidade (ex. av[ã] 'avant'). Esse fato contribuiria para a defesa da hipótese de articulação secundária na língua (ex. irmãw̃ŋ). A consoante nasal é, como discutimos neste texto, subjacente para muitos autores e funcionaria como gatilho para a nasalização do ditongo, em princípio oral na origem. Não sendo apagada, isto é, preservando-se plena ou secundariamente em coda, após o espraiamento, torna-se alvo da assimilação do ponto de articulação do segmento que a sucede ou mesmo do glide que a precede, a depender do contexto.

1997, 2000; GUIMARÃES; NEVINS, 2013; entre outros), mas também na perspectiva à qual se soma esta análise, a da produtividade (HUBACK, 2010a, 2010b; CRISTÓFARO-SILVA, 2012; BECKER *et al*., 2018; GOMES; PRADO; AMARAL, 2021, no prelo; RIZZATO, 2018; entre outros).

A hipótese geral assumida aqui é a de que restrições implicadas na seleção de expoentes fonológicos concorrentes para entidades morfológicas podem ser alcançadas pela medição da produtividade de seus contextos fonológicos e morfológicos, sobretudo em distribuições imperfeitas, aquelas que não se enquadram plenamente no conceito tradicional de alomorfia. O exercício empírico que descrevemos neste artigo se dá na esfera de hipóteses subordinadas a essa, tratando de cada contexto potencialmente revelador dessas restrições, considerando-se um corpus de grande abrangência.

O texto está organizado como segue. Na seção 2, problematizamos a noção de alomorfia no que diz respeito à realização das marcas de plural exploradas neste trabalho e a relacionamos ao conceito de produtividade que adotamos. Na seção 3, resumimos brevemente alguns estudos que se ligam mais diretamente aos objetivos de nossa pesquisa. Na seção 4, detalhamos os procedimentos metodológicos. Por fim, na seção 5, apresentamos e discutimos os resultados da investigação. Seguem-se nossas considerações finais acompanhadas da agenda da pesquisa.

2 Alomorfia, produtividade e plurais irregulares do PB

Alomorfe, ou variante de morfema, define-se, na tradição, como uma das realizações de um morfema. Nem toda realização de morfema, contudo, enquadra-se nesse conceito. Segundo Bonet, Lloret e Mascaró (2015, p. 1), alomorfia existe sob duas condições: que haja mais de um morfe para cada morfema e que a divergência entre os morfes não possa ser predita pela fonologia regular da língua. Assim, cada alomorfe, por ser impredizível fonologicamente, constitui uma forma subjacente distinta. Por outro lado, alomorfes não ocorrem irrestritamente: além de, para muitos, deverem se assemelhar fonologicamente, devem também ter especificados seus contextos de ocorrência, e a expectativa é que tais contextos estejam numa relação de excludência mútua, ou distribuição complementar (BAUER, 2004, p. 15).

Nesses termos, podemos nos perguntar se as formas de plural aqui abordadas, exemplificadas em (1) e (2), podem ser consideradas alomorfes.

Para responder a essa pergunta, em primeiro lugar, precisamos examinar se é possível relacionar as formas em análise a um morfema comum, de maior abrangência na língua, no caso -s, além de nos assegurar de que o que as diferencia foneticamente não se caracteriza como regra fonológica ordinária na língua. Começando pelo segundo critério, um simples teste posicional pode responder à questão. Sabemos que as sequências Vl ou Vu podem aparecer no meio da palavra (ex. alto, incauto). Não há, contudo, evidência de processo que converta Vl em **Vis** no interior do vocábulo, e a sequência ls é estranha ao português (a palavra *solstício* seria uma exceção). **Vus**, por sua vez, parece fazer parte da representação lexical de palavras também em posição interna (ex. claustro, Fausto). A sequência ão, por outro lado, não ocorre, em princípio, em sílaba medial em português, mas apenas em final de palavra. Esse teste indica que os processos que analisamos são licenciados apenas na formação do plural, o que sugere que não estamos diante de fenômenos fonológicos regulares da língua. A complexidade do problema, porém, reside na tarefa de se explicar, por via fonológica, alternantes que modificam a forma básica de plural -s ou que lhe acrescentam substância fônica. Nessa perspectiva, várias hipóteses são exploradas na literatura, a maioria defendendo que está em jogo, mais do que alomorfia de plural, alomorfia de raiz (ex. papele+s/ limon+s) e emergência de um glide ativado pela associação de [s] à sílaba da forma singular (BISOL, 1998; WETZELS, 1997; entre outros). Em contraste está a alternativa de se lexicalizar o morfema, mais do que sua base, assumindo-se, se não alomorfia propriamente dita, a hipótese de subléxicos, ativados por restrições fonológicas ou de outra natureza, inclusive extralinguística. É, por exemplo, a análise assumida por Becker *et al.* (2018) e Rizzatto (2018), para o tratamento de plurais de palavras fechadas foneticamente por Vw, ou por Abaurre Gnerre (1983), para o tratamento de plurais de palavras fechadas por ditongo nasal.

Em segundo lugar, precisamos considerar que os contextos de ocorrência dessas formas não são cem por cento excludentes, ou seja, em princípio, as três variantes ocorrem em ambientes fonéticos e gramaticais semelhantes. Há, porém, predominâncias em relação a esses contextos que podem sinalizar para quase-regularidades.

No recorte da análise desenvolvido neste texto, vamos nos fixar nessas sub-regularidades, sem compromisso com o status mais ou menos alomórfico das alternantes que analisamos, uma vez que as propriedades fonológicas das marcas de plural, ou mesmo sua afiliação morfológica – à base ou ao sufixo –, são por ora secundárias.

Assumimos o entendimento de que produtividade de morfemas é propriedade gramatical das línguas naturais. Isso quer dizer que tanto pode ser depreendida do léxico em uso quanto pode ser predita a partir de restrições que concorrem para formação de novos itens, o léxico potencial. O léxico potencial, portanto, deve dar conta da porção transparente do léxico disponível, mas também pode incluir formações menos regulares, fruto de subconjuntos de restrições ou de restrições mais baixas na hierarquia de uma língua. Este estudo dedica-se a dados do léxico em uso.

3 Estudos sobre a produtividade de plurais irregulares em PB

A alternância dos expoentes fonológicos do plural de ditongos orais e nasais foi avaliada na perspectiva de sua produtividade em contextos fonológicos e morfológicos específicos por Huback (2010a, 2010b), Cristófaro-Silva (2012), Becker *et al*. (2018), Rizzato (2018), Gomes, Prado e Amaral (2021, no prelo), entre outros autores. Nesta seção apresentamos algumas das principais ideias desses textos em versão resumida. Por conveniência expositiva, tratamos primeiramente dos plurais em u/l, para, em seguida, tratarmos dos plurais em ão.

Plurais em u/l

Em relação a palavras terminadas em u/l, Huback (2010a) realiza um estudo envolvendo a aplicação de um teste de reação a 36 falantes nativos do PB, contando com 53 palavras-alvo distribuídas em 3 categorias relativas a frequência lexical. Com base na amostra do Corpus NILC/São Carlos, as palavras-alvo são classificadas nas frequências de ocorrência baixa, média e alta. Com foco teórico no Modelo de Redes (BYBEE, 2001), está em jogo a hesitação ou não hesitação do falante ao produzir o plural. Os resultados mostraram que 12,8% dos itens terminados em l foram pluralizados como se pertencessem aos itens terminados em u, ou seja, contando somente com a adição da expressão fonológica de plural -s. A autora observou favorecimento em particular

da hesitação para os itens *mel*, *sol* e *sal*. Apesar de haver diferenças de frequência entre esses vocábulos, o que os aproxima é o fato de serem todos monossílabos, sugerindo papel desta variável em favor da hesitação. De modo geral, a autora conclui que não há correlação entre frequência de ocorrência e hesitação, já que palavras de frequência de ocorrência baixa, média e alta inibiram hesitações. Os itens terminados em u mostraram-se mais suscetíveis a hesitações, o que pode ser explicado pelo fato de esse ser o grupo com menor frequência de tipo. Desses itens, 15,6% migraram para a pluralização em is. Itens de baixa frequência, como *jirau*, *mausoléu*, *véu*, apresentaram índices mais altos de hesitação, confirmando as expectativas do modelo adotado. O item *judeu*, contudo, de frequência alta no corpus adotado, contrariou a expectativa. As palavras menos suscetíveis a hesitação nesse grupo foram justamente as de frequência mais alta, como *meu* e *seu*, apontando para correlação entre as variáveis frequência de ocorrência e hesitação no grupo dos itens fechados por u. O estudo conclui, ainda, que a frequência de tipo, menor para os itens terminados em u, também interfere nas migrações, mais recorrentes nesse grupo.

Ainda a respeito de palavras terminadas em u/l, Cristófaro-Silva (2012), propondo uma análise à luz do que refere como modelos multirrepresentacionais, categoriza as marcas fonológicas de plural desses itens em (i) nomes em que a marca de plural ocorre a partir da adição de -s (ex. degraus) e em (ii) nomes em que o plural se realiza pela adição de um suposto morfema is (ex. sais), com alteração do radical (além de formas em que a marca de plural pode deixar de ocorrer, como em *os sal*). A autora destaca, com base nos dados de Huback (2007), que é possível se atestarem plurais pertencentes ao padrão (i) realizando-se conforme o padrão (ii) (ex. degrais).

Becker *et al*. (2018) apresentam um estudo sobre essas formas na perspectiva da aquisição de seus plurais. Os autores realizam um experimento psicolinguístico envolvendo pseudopalavras com 115 crianças de 7 a 12 anos e um grupo de controle formado por adultos. Participantes de 7 a 9 anos evitaram a alternância [w ~ j] em monossílabos, sugerindo influência de uma restrição da língua que atua na proteção da primeira sílaba, o que, na interpretação dos autores, equivale a proteção de monossílabos. O estudo mostra que essa proteção é ainda maior entre crianças de 10 a 12 anos. O grupo de controle e as crianças de 10 a 12 anos, além disso, tendem a evitar novos ditongos com baixa dispersão

de altura entre vogal e glide (ex. [ej, ew, oj, ow]), apesar de sua forte presença no léxico. Esse fato também é interpretado na perspectiva de uma restrição violável, que milita pela maior dispersão de altura em ditongos.

Gomes, Prado e Amaral (2021, no prelo) também realizam um experimento psicolinguístico valendo-se de pseudopalavras, além de um teste de produção com palavras de baixa frequência. Participaram 54 voluntários, sendo 25 com ensino superior, cursando a graduação da Faculdade de Letras na UFRJ, e 28 de um curso de Educação para Jovens e Adultos de Niterói, a fim de controlar o papel da escolaridade. Os resultados apontam para prevalência da forma de plural s em monossílabos para ambos os níveis de escolarização. Em relação a essa variável, entretanto, constataram divergência quanto à vogal [e] no núcleo da última sílaba tônica, que se mostrou significativa apenas para o grupo de participantes de nível superior, desfavorecendo o plural js.

Plurais em ão

No âmbito do ditongo nasal, Huback (2010a) verifica também a disponibilidade de ões, ãos e ães no léxico dos informantes quando perguntados a respeito do plural de vocábulos como *escrivão*, por exemplo. Os resultados obtidos a partir do experimento, descrito anteriormente, apontam para uma migração direcionada no uso das alternantes de plural do ditongo nasal: 32,5% dos vocábulos cujo étimo prevê ãos e 20,9% dos vocábulos cujo étimo prevê ães foram pluralizados pela alternante ões no experimento. Há, portanto, influência da frequência de tipo da alternante ões em sua produtividade, visto que ões participa da pluralização de um maior grupo de palavras terminadas pelo ditongo nasal em PB de acordo com dados do dicionário Houaiss observados por Huback (2010a, p. 19). A migração no sentido contrário não foi atestada de maneira expressiva: 4,1% de itens etimologicamente pluralizados pela alternante ões optaram por ãos para expressar plural, por exemplo. Houve, entretanto, hesitação nas respostas obtidas no experimento. Na análise de regressão binária, vocábulos para os quais se supõe ãos no étimo e vocábulos de baixa frequência de ocorrência se mostraram favorecedores de hesitação na obtenção de seus plurais, apontando para influência significativa de efeitos de frequência de tipo e de ocorrência no fenômeno em questão. Huback (2010b) apresenta resultados semelhantes a partir de outro experimento, incluindo agora a leitura de frases e figuras

como estímulos, considerando fatores linguísticos e extralinguísticos. Em relação ao número de sílabas, por exemplo, constatou que monossílabos desfavorecem categoricamente a aplicação do expoente fonológico mais produtivo, ões.

Cristófaro-Silva (2012) também discute a organização do plural de nomes do PB terminados em ditongo nasal na perspectiva de modelos multirrepresentacionais. A autora classifica nomes encerrados pelo ditongo nasal ãos na categoria geral de nomes que expressam plural pelo simples acréscimo de -s (ex. irmãos). Vocábulos que tomam ões ou ães como expoente fonológico de plural (ex. leões, pães), por outro lado, se relacionam pelo fato de apresentarem, segundo a autora, *alteração no radical nominal* e por compartilharem o morfema de plural -is. O padrão do grupo de itens lexicais que toma ões como marca de plural é aplicado em generalizações, de acordo com a análise, como consequência de sua alta frequência de tipo, em consonância com a análise de Huback (2010a).

Essa preferência pelo expoente ões é percebida também por Rizzato (2018), que promove uma análise experimental, aplicada a 79 falantes do PB, envolvendo a seleção de plural de ditongos nasais em 24 frases. O contexto analisado é o do aumentativo -zão – expressão de grau passível de anexação em vocábulos que já contenham traço de número (ex. coraçõezões), em sua interpretação. A opção pela marcação dupla de plural se revela variável em seus dados, ainda que a preferência por ões no âmbito do sufixo seja significativa. Embora sob variação, monossílabos exibem um comportamento distinto também nesta análise, sendo *pãezões* e *cãezões* as formas preferidas, enquanto trissílabos compartilhadores do mesmo plural etimológico optam pela expressão de plural menos redundante (ex. capitãozões).

4 Procedimentos metodológicos

Nesta seção apresentamos a constituição e a organização de duas subamostras extraídas a partir de uma amostra-base de 3.744.513 *types* e 691.758.151 *tokens* disponível para download no site do Corpus Brasileiro – doravante CBras. O CBras é um banco de dados alimentado por diferentes fontes, incluindo fala e escrita. A análise que empreendemos neste estudo tem como base principalmente os *types* oferecidos pela amostra-base; recorremos, contudo, aos *tokens* na discussão sobre frequência lexical.

As subamostras analisadas estão assim constituídas:

(i) Amostra L, contendo 9.245 vocábulos pluralizados correspondentes a bases fechadas ortograficamente por vogal+u/l;

(ii) Amostra N, contendo 5.899 vocábulos pluralizados correspondentes a bases fechadas ortograficamente por ão.

Os dados foram extraídos da amostra-base e organizados com o uso da Plataforma R. A investigação restringiu-se ao que tratamos como *nomes*, numa interpretação ampla desse termo, que abriga substantivos e adjetivos – vocábulos sujeitos à flexão de plural. A organização inicial dos dados consistiu, entre outros aspectos, em juntar formas com pequenas diferenças de grafia, incluindo maiúsculas e minúsculas ou espaçamentos indevidos, eliminar resíduos de palavras estrangeiras ou com sequências incompreensíveis e checar no corpus, através de busca pelo site Linguateca,[3] a classe gramatical de itens homófonos para excluir com segurança não nomes (ex. vão, não). Os dados, contudo, neste caso, não são analisados em sua forma plenamente lematizada, porque optamos pela manutenção da distinção entre lexemas, isto é, palavras simples, derivadas ou compostas foram preservadas separadamente na amostra. No exercício analítico apresentado neste texto, porém, por diversas vezes apresentamos comparações com subconjuntos de dados que fazem referência a lemas, controlados como uma entre as variáveis independentes.

Após a etapa de preparação automática dos dados, cada subamostra foi codificada manualmente levando-se em conta propriedades fonológicas e léxico-morfológicas que pudessem contribuir em alguma medida para contextualizar a ocorrência das diferentes formas de plural em foco.

Ainda que as análises das duas subamostras estudadas sejam inteiramente independentes uma da outra, por serem conduzidas a partir de variáveis análogas, são apresentadas nesta seção e na seção de resultados sempre que possível conjuntamente.

Não é demais registrar, ainda, que, apesar de usarmos neste texto o termo *variável*, comum em estatística, não abordamos o fenômeno em questão como variável no sentido da sociolinguística variacionista, isto é,

[3] https://www.linguateca.pt/acesso/corpus.php?corpus=CBRAS

entendemos que estamos tratando de uma alternância, já que os contextos de seleção de cada variante são na maioria das vezes excludentes. Variação, enquanto *duas ou mais formas de se dizer a mesma coisa*, ocorre em margem muito reduzida na amostra (ex. g[ojs] ~ g[ows] ~ g[oɬs]; capit[õj̃s] ~ capit[ãw̃s] ~ capit[ãj̃s]) e não é nosso foco aqui.

4.1 Variável dependente

As variáveis de interesse em nossa análise são as alternantes de plural atestadas em cada amostra, como exemplificamos inicialmente em (1) e (2). A subdivisão que propomos é tão somente uma alternativa de análise, naturalmente passível de reinterpretação a depender de como se defina o padrão de marcação morfológica de plural nesses casos. Como dissemos, optamos nas duas amostras por representar essas variáveis considerando as rimas das sílabas finais das formas pluralizadas.

Alternantes de plural da Amostra L

A divisão adotada inicialmente para a Amostra L contempla três alternantes de plural: **Vis, Vus e is**. A primeira e a última alternantes, embora se assemelhem no sentido de terminarem em **is**, diferenciam-se quanto à realização obrigatória de um ditongo no primeiro caso (ex. past[ɛjs]) e à não realização ou realização apenas opcional de um ditongo homorgânico, ou ainda de um alongamento da vogal, no segundo (ex. vin[is] ~ vin[ijs] ~ vin[i:s]). A segunda alternante, **Vus**, também forma ditongo, mas com um glide labiovelar (ex. r[ɛws]).

Alternantes de plural da Amostra N

Para a Amostra N, as alternantes em análise são também três, as tipicamente registradas no léxico do português: ões, ãos e ães (ex. bot[õj̃s], m[ãw̃s], c[ãj̃s], respectivamente).

4.2 Variáveis independentes

As variáveis independentes deste estudo, examinadas sempre na perspectiva das alternantes mencionadas, são baseadas em hipóteses da literatura resenhada na seção 3 e em aspectos que consideramos não plenamente contemplados nesses estudos. Os grupos de fatores investigados são a seguir descritos e exemplificados.

Número de sílabas

Nas duas amostras, os vocábulos foram classificados quanto ao número de sílabas da palavra. A primeira classificação incluiu palavras de 1 a 9 sílabas. Uma reorganização, entretanto, foi proposta, seja porque dados de 5 a 9 sílabas se mostraram escassos, seja porque não se observou, numa rodada preliminar, relevância na distinção entre palavras com 4 sílabas e palavras maiores. A classificação analisada, assim, foi **1 sílaba** (ex. réus, cães), **2 sílabas** (ex. quartéis, irmãos), **3 sílabas** (ex. mausoléu, capitães), **4 ou mais sílabas/polissílabas** (ex. governamentais, nacionalizações).

Acento

Palavras terminadas em vogal+u/l e ão apresentam acento final na maioria dos casos e, em proporção consideravelmente reduzida, são acentuadas na penúltima sílaba, justificando a seguinte classificação: **acento final** (ex. quintais, sabões) e **acento pré-final** (ex. móveis, órgãos).

Contexto fonológico

Neste grupo foram classificados os segmentos que precedem a porção fonológica comum às três alternantes estudadas em cada amostra. Assim, na Amostra L, abordamos a vogal do núcleo da sílaba envolvida no plural, a que antecede o glide no ditongo (ex. p[a]us, c[ɛ]us, anz[ɔ] is, vin[i]s, az[u]is). No caso de [i], estão contempladas as possibilidades de não ditongação, de ditongação homorgânica ou de alongamento (ex. vin[i]s ~ vin[ij]s ~ vin[i:]s). Na Amostra N, abordamos a consoante do onset da sílaba que contém o ditongo nasal. Uma classificação inicial considerando todos os segmentos individualmente foi reinterpretada na perspectiva de modo de articulação: **oclusiva** (ex. capi[t]ães), **fricativa** (ex. se[s]ões), **líquida** (ex. sa[l]ões) e **nasal** (ir[m]ãos). Também foram considerados casos de **hiato** (ex. pe.ões).

Afiliação morfológica

Neste grupo de fatores, analisamos a localização morfológica da terminação da base que dá origem aos plurais investigados. Interessa-nos dizer se as porções fonológicas u ou l, na Amostra L, e ão, na Amostra N,

coincidem integralmente com um sufixo (ex. pastor+**il**, mulher+ão), são apenas parte de um sufixo (ex. amá+ve**l,** implementa+ção) ou integram o radical (ex. me**l**, cão). Para isso, propomos uma divisão mais acurada de contextos para cada amostra, que se sujeitou a amalgamações resultando, na Amostra L, em radical mais 3 tipos de sufixos, e, na Amostra N, em radical mais 4 tipos de sufixos, como se detalha na seção 5.4.

Frequência lexical

A frequência lexical dos itens pluralizados em análise é a informada na amostra disponibilizada pelo CBras e foi analisada como uma variável contínua.

Outras variáveis, como padrão de **terminação da base**, **classe gramatical**, **lema** e **frequência do lema**, também foram codificadas e aparecem nesta análise como subsidiárias à descrição das 5 que tomamos como nucleares.

5 Resultados e discussão

Nesta seção fazemos uma apresentação dos resultados da análise das duas subamostras investigadas. Como antecipamos, essa exposição se dá, sempre que possível, de forma combinada, ainda que as análises tenham sido computadas independentemente.

Os resultados aqui apresentados são produto de estatística descritiva fazendo-se uso da Plataforma R. Cada resultado é acompanhado de discussão, considerando-se os achados de trabalhos anteriores e o potencial de cada grupo de fatores para uma futura análise de natureza preditiva.

Começamos por apresentar a frequência geral das alternantes de plural para cada uma das amostras analisadas. Na sequência, tratamos das 5 variáveis que consideramos nucleares na ordem em que foram apresentadas na metodologia. As variáveis subsidiárias são chamadas quando necessárias à discussão dos resultados obtidos para as categorias nucleares.

5.1 Frequência geral

Confirmando o amplamente divulgado na literatura, observamos, na Amostra L, prevalência da alternante **Vis**, seguida por **Vus** e **is**

com índices semelhantes, e, na Amostra N, prevalência da alternante **ões**, seguida de **ãos** e **ães** — esta última, contudo, com ocorrência consideravelmente inferior à alternante **ãos**.

TABELA 1 – Distribuição geral de plurais irregulares – Amostras L e N

Amostra L				Amostra N			
	%	Ocor.			%	Ocor.	
Vis	98,0	9.061	papéis	ões	92,8	5.472	balões
is	1,0	88	funis	ãos	5,2	307	irmãos
Vus	1,0	96	chapéus	ães	2	120	capitães
Total		9.245		Total		5.899	

Por conta da distribuição geral apresentada na Tabela 1, em que se opõem, em cada amostra, uma alternante de emprego superior a 90% a outras duas alternantes com emprego bastante reduzido, nossos resultados serão na maioria das vezes apresentados, a partir daqui, considerando-se um universo de 100% para cada alternante. Isso permitirá focalizar o papel das variáveis examinadas no espectro de cada alternante, dando visibilidade às variantes não prevalentes, e também possibilitará comparar esse efeito entre as alternantes em termos relativos.

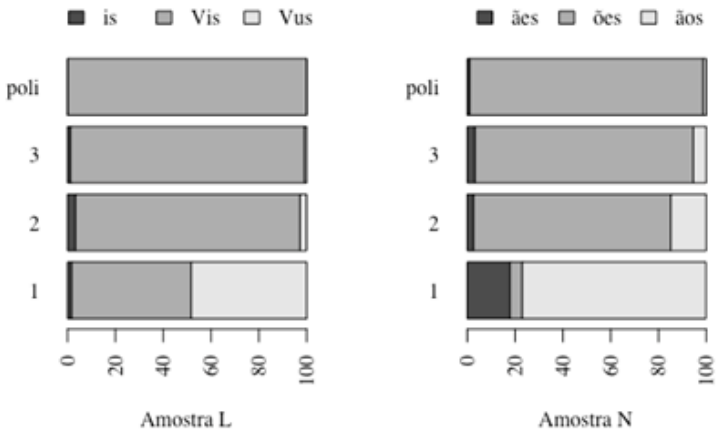
5.2 Número de sílabas

O exame da variável número de sílabas revelou, para as duas amostras investigadas, que, quanto maior a extensão silábica do vocábulo, menor a ocorrência das alternantes não prevalentes.

Na Amostra L, **Vis** é a alternante prevalente em vocábulos de 2 ou mais sílabas. Entre os monossílabos, contudo, divide a liderança com **Vus**, que apresenta 50% de ocorrência. Há apenas um caso, 1,8%, relativo à alternante **is** entre os monossílabos.

Na Amostra N, a alternante **ões** também é prevalente em vocábulos de 2 ou mais sílabas. Entre os monossílabos, contudo, observa-se uma inversão de preferência, com **ãos** como a variante mais frequente, 76,9%, seguida de **ães**, 17,9%, contra reduzidos 5,1% de **ões**. Não devem ser desprezados nesta amostra também os índices encontrados para **ãos** entre dissílabos e trissílabos.

GRÁFICO 1 – % de plurais irregulares e nº de sílabas – Amostras L e N



O comportamento diferenciado de monossílabos relatado por Huback (2010a, 2010b), em especial para plurais de ão, foi atestado em nossas duas amostras. Pode estar em jogo aqui a fidelidade de monossílabos às suas formas de base, conforme defendem Becker, Nevins e Levine (2012) e Becker *et al.* (2018), numa abordagem que pareia monossílabos com sílabas iniciais. Essa tese, contudo, fica na dependência do debate sobre que formas estão disponíveis na base, ou na subjacência, dessas alternantes: se contam com segmentos mais abstratos, como /l/, por exemplo, no caso da Amostra L, e /N/, no caso da Amostra N. Embora tópico de nossa investigação, representações subjacentes não são, contudo, foco deste artigo, razão por que por ora adiamos esse debate. Além disso, considerando que nosso fenômeno se restringe ao contexto de final de palavra, afirmações sobre comportamento análogo de monossílabos a sílabas iniciais seriam, neste recorte, temerárias.

Um exercício necessário, entretanto, é a verificação dos itens lexicais – em nosso caso, lemas – que integram esses padrões excepcionais, a fim de subsidiar a discussão sobre um possível efeito de lexicalização de palavras mais do que de morfemas ou alomorfes neste caso. A hipótese é de que quanto menor a proporção de lemas para ocorrências, maior a restrição lexical, isto é, maior a possibilidade de um item se repetir na amostra. Essa, evidentemente, é uma medida geral, que deve ser relativizada tanto nos casos em que há muito poucas ocorrências

quanto nos casos em que itens particulares apresentam frequências muito desequilibradas em relação aos demais dados.

Na Tabela 2, registramos a distribuição de lemas e monossílabos atestados para as alternantes investigadas em cada amostra devidamente listados/exemplificados (em sua forma pluralizada, por conveniência expositiva). Itens considerados excepcionais estão em destaque. Os números mostram limitação categórica de lemas para a variante **is**, na Amostra L, e para a variante **ões**, na Amostra N. No que diz respeito às demais variantes, as proporções da distribuição lemas/ocorrências, todas entre 20 e 30%, sugerem importante limitação lexical. Esse fato, considerada a frequência individual dos itens, pode ser um indicador de controle do léxico mais do que de seleção alomórfica propriamente dita. Essa discussão será, na perspectiva da frequência lexical geral, retomada na seção 5.6.

TABELA 2 – Plurais de monossílabos – Amostras L e N

	Amostra L				Amostra N		
	%	Lemas/ Ocor.			%	Lemas/ Ocor.	
Vis	29,6	8/27	sais, réis, géis, méis, sóis, grais, **gois**, móis	ões	100	2/2	**chões**, **vões**
Vus	32,1	9/28	graus, maus, réus, céus, paus, naus, véus, vaus, tchaus	ãos	23,3	7/30	mãos, grãos, vãos, sãos, chãos, nãos, pãos
is	100	1/1	vis	ães	28,6	2/7	cães, pães

Entre os dissílabos, trissílabos e polissílabos, apesar da prevalência das variantes **Vis**, na Amostra L, e **ões**, na Amostra N, as alternantes minoritárias também merecem alguma atenção.

Na Amostra L, observa-se, como no caso dos monossílabos, restrição lexical no caso da alternante **Vus**, predominando, entre os dissílabos, os itens *degraus*, *chapéus*, *troféus* e *mingaus*, e, entre os trissílabos, vocábulos como *mausoléus*, *berimbaus* e *bacalhaus*, além de alguns nomes próprios pluralizados, como *nicolaus* ou *venceslaus*. Não há polissílabos no padrão **Vus**. No caso da variante **is**, contudo, o fenômeno parece mais automático, já que não há concentração de lemas, mas uma

relação direta entre a terminação Consoante+il e o plural **is**, tanto em dissílabos, quanto em trissílabos e polissílabos (ex. civis, perfis, infantis, mercantis, estudantis, primaveris). Exploramos novamente essa relação na seção 5.5, quando abordamos a afiliação morfológica das alternantes em questão, incluindo-se os sufixos.

Na Amostra N, em relação ao padrão **ãos**, também há alguma predominância de certos lemas entre dissílabos, com destaque para os itens órgãos, *irmãos*, *cristãos*, órfãos, *bênçãos* e *pagãos*, e entre trissílabos, concentrando maior frequência em itens como *cidadãos*, *artesãos*, *anciãos*, *acórdãos*, *afegãos*, *cortesãos* e *corrimãos*. Os polissílabos em **ãos** da amostra podem ser considerados formas inesperadas (no sentido de serem muito mais recorrentes com a alternante **ões**): *concentrações*, *informações*, *anfitriãos*, *apresentações*, *comunicações*, entre outras de mesmo tipo, todas com baixa frequência. No padrão **ães**, também se atestam vocábulos de todas as extensões. Entre os dissílabos e polissílabos analisados, todos parecem ser exceções a formas mais comumente realizadas com **ões** e, em alguns casos, com **ãos**, como *refrães*, *peães*, *irmães*, *anães*, entre outros. No que se refere aos trissílabos, diferentemente, entre os itens mais frequentes estão alguns que são predominantemente realizados com **ães** na fala culta, como *alemães*, *capitães*, *catalães*, casos muitas vezes de sufixos gentílicos, o que se descreve, como prenunciamos, na seção 5.5, adiante. Os dados de polissílabos, como os de dissílabos, à exceção do vocábulo *tabeliães*, são corriqueiramente produzidos com **ões**, como *cirurgiães*, *informações*, *opiniães*, *alterações* etc.[4]

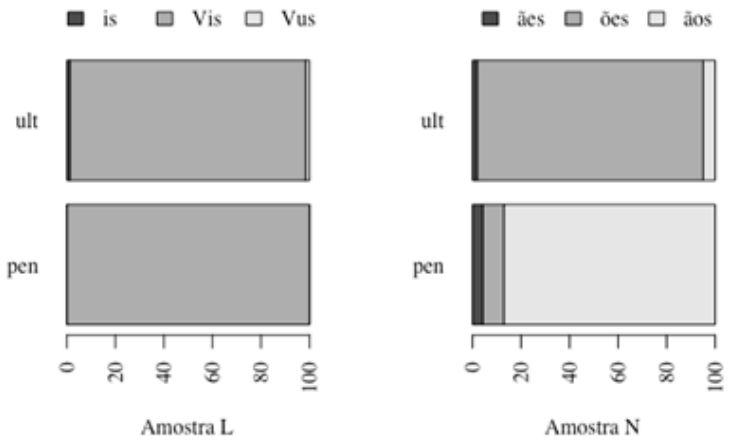
5.3 Acento

Formas terminadas em vogal+u/l e ão ortográficos são predominantemente oxítonas em português. Na Amostra L, 70% dos itens apresentam acento final, contra 30% de paroxítonas. Na Amostra N, quase a totalidade dos itens é de oxítonas, com apenas 0,4% de paroxítonas. Em

[4] Por se tratar de uma grande base de dados, alimentada por diferentes fontes, não se descarta a possibilidade de esses e outros itens excepcionais atestados nesta análise constituírem meros erros de registro. Não nos cabe, contudo, decidir por descartá-los, tanto porque não apresentam marcas incontroversas de lapsos de grafia quanto porque, apesar de atípicos, apresentam-se como possibilidades na língua. A frequência lexical, por outro lado, figura, em nosso entendimento, como regulador – necessário, se não suficiente – na interpretação dessa ambiguidade.

nenhuma das amostras há proparoxítonas. O Gráfico 2 é particularmente informativo em relação aos dados atípicos.

GRÁFICO 2 – % de plurais irregulares e acento – Amostras L e N



Na Amostra L, 100% das paroxítonas fazem plural em **Vis**. Entre os itens de acento final, contudo, ainda que predomine o padrão **Vis**, 1,5% e 1,3% correspondem, respectivamente, aos padrões **Vus** e **is**. A quantificação das ocorrências e a proporção de lemas envolvidos nesses dois padrões excepcionais estão listados na Tabela 3. Os lemas representam, tanto para a alternante **Vus** quanto para **is**, mais de 50% das ocorrências nessas categorias, índices que não permitem conclusões contundentes acerca de restrição lexical.

TABELA 3 – Plurais de oxítonos – Amostra L

	%	Lemas/Ocor.	
Vus	76	73/96	graus, maus, réus, céus, degraus, chapéus
is	53,5	46/86	civis, perfis, infantis, juvenis, barris, estudantis

Na Amostra N, localizamos 23 casos de paroxítonas. Como mostra a Tabela 4, são os mesmos 8 lemas caracteristicamente realizados no plural com a alternante **ãos** que apresentam alguma variação em **ões**

e **ães** – estes últimos com baixa frequência de acordo com os índices do CBras. Se de fato nomes paroxítonos condicionam em alguma medida a seleção da alternante **ãos**, não há que se falar em restrição lexical. O fato, porém, de lidarmos com muito poucas ocorrências neste caso deixa tal conclusão em suspenso, na dependência de análises que considerem o léxico potencial mais do que o institucionalizado.

TABELA 4 – Plurais de paroxítonos – Amostra N

	%	Lemas/Ocor.	
ões	100	2/2	órgões, **acórdões**
ãos	40	8/20	órgãos, órfãos, bênçãos, acórdãos, sótãos, cóvaos, cristóvãos, orégãos
ães	100	1/1	órgães

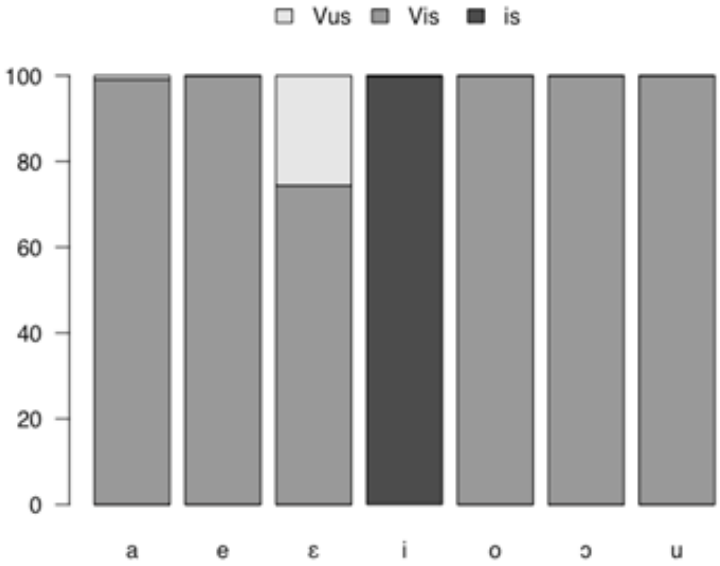
5.4 Contexto fonológico

Nesta subseção tratamos dos segmentos que precedem a porção fonológica comum às três alternantes estudadas em cada amostra. Assim, na Amostra L, abordamos a vogal do núcleo da sílaba envolvida no plural, a que antecede o glide no ditongo ou a que precede s, em casos de suposta não ditongação, como em *funis*. Na Amostra N, abordamos a consoante do onset da sílaba que contém o ditongo nasal, ou sua ausência, em caso de hiato, como em *peões*.

Vogal do núcleo na Amostra L

Na Amostra L, classificamos inicialmente cada uma das vogais que nuclearizam a sílaba envolvida na alternância em questão. A preferência categórica de **Vis** para qualquer vogal é frustrada pela vogal [i], que, por razões estruturais, concentra 100% de suas ocorrências no padrão **is**, e pelas vogais [a] e [ɛ], que apresentam, respectivamente, 0,9% e 25,6% de suas ocorrências no padrão **Vus**.

GRÁFICO 3 – % de plurais irregulares e vogal do núcleo – Amostra L



Os resultados não indicam evidência imediata para agrupar as vogais por algum parâmetro fonético-fonológico (articulatório, por exemplo), razão por que as mantivemos separadas nesta análise. Cabe, contudo, averiguar os itens que justificam esse comportamento não categorial, a fim de checar se não se pode atribuí-lo ao *design* do léxico mais do que propriamente à fonologia. Os dados de [i], categóricos para **is**, serão examinados com mais detalhe quando tratarmos da variável afiliação morfológica, mais especificamente da terminação il. Exploramos, porém, a distribuição dos casos atípicos envolvendo as vogais [a] e [ɛ], como se vê na tabela 5. As proporções relativamente altas de lemas/ocorrências não permitem se concluir que estejamos diante de um efeito de restrição lexical neste caso.

TABELA 5 – Plurais atípicos com núcleos a, ɛ – Amostra L

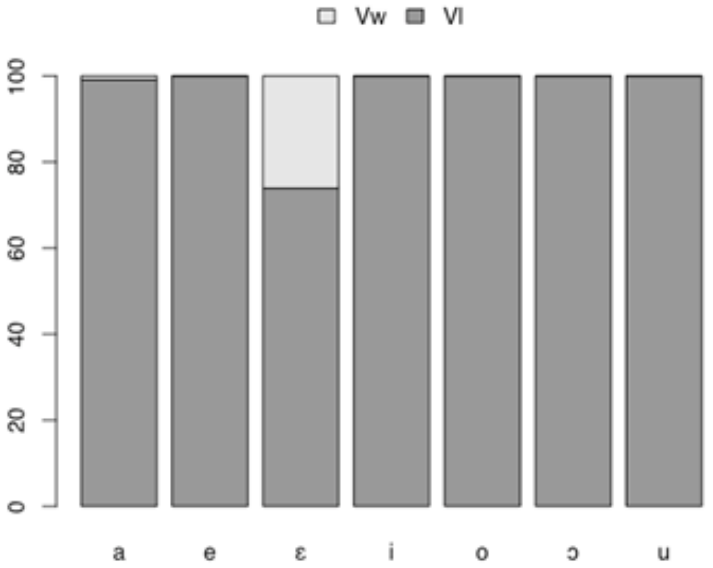
		%	Lemas/Ocor.	
Vus	a	74,5	38/51	maus, degraus, paus, naus, saraus, mingaus
	ɛ	77,8	35/45	réus, céus, chapéus, troféus, véus, mausoléus[5]

Desconsiderando-se, contudo, a explicação de restrição lexical, a interpretação para o resultado da alternante **Vus** deve levar em conta, ainda, a hipótese de distinção subjacente entre bases terminadas em /Vw/, ditongos legítimos, e bases terminadas /Vl/, sujeitas à ditongação por derivação fonológica (BISOL, 1989; CAMARA JR., 1970, entre outros). Essa distinção está presente na ortografia inclusive, contrastando pares como *mau* (adjetivo) vs. *mal* (advérbio), cujos plurais ainda se preservam *maus*/*males* em grande medida na língua culta. Não há garantia de que o dado de escrita, neste caso, reflita a fala, mas também não é possível negar com segurança que alguma representação próxima da escrita possa subjazer, na mente, o dado de fala. O fato é que convenções de escrita podem ser causa e consequência de padrões fonológicos, já que, alimentadas pela fala, são também alimentadoras das representações sonoras mentais de falantes escolarizados (SCHWINDT *et al*., 2007).

A título de exercício, para explorar especificamente os padrões de terminação da base singular, codificamos adicionalmente a Amostra L da seguinte forma: Vl equivale a formas fechadas na escrita por l (ex. canal, pincel) e Vw, por ditongo com u (ex. pau, céu). O cruzamento dessas categorias com as vogais que ocupam o núcleo silábico está representado no Gráfico 4.

[5] Desses itens frequentes de base éu que listamos, apenas troféu apresenta dado de plural com a alternante Vi**s, trofé**is, com frequência 4, contra troféus, com frequência 913.

GRÁFICO 4 – % de terminações da base singular e vogal do núcleo – Amostra L



No padrão de base singular Vl, que responde por 98,9% dos dados, podemos encontrar todas as 7 vogais como núcleos na Amostra L, concentrando-se 94.5% dos dados nas vogais [a] e [e] (ex. avental, móvel). No caso de Vw, só se registram dados com núcleos [a] e [ɛ] (ex. degrau, troféu), com as ocorrências que já relatamos quando tratamos do padrão de plural **Vus**.

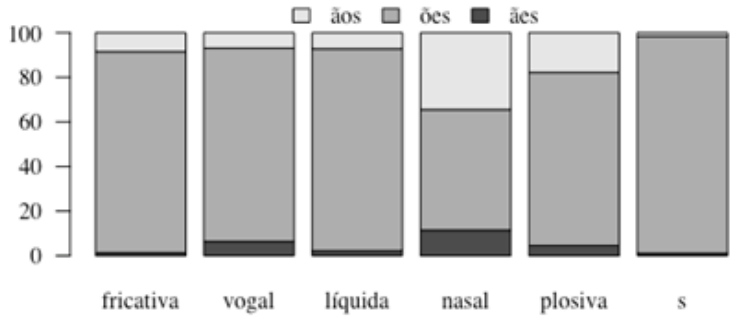
Ainda que não possamos, com base no tipo de dado de que nos utilizamos neste trabalho, concluir sobre a distinção da representação subjacente das alternantes envolvidas, nossos resultados confirmam, por ora, no que tange às vogais nucleares, a correlação entre os padrões Vl_{SING}/Vis_{PL}, Vw_{SING}/Vus_{PL}.

Consoante precedente na Amostra N

Em relação ao contexto fonológico precedente ao ditongo nasal, a partir do exame estatístico de diferentes classificações considerando os mais variados segmentos e encontros consonantais atestados, chegamos

ao resultado apresentado no gráfico 5. Para todas os contextos, prevalece a alternante **ões**. De diferente da categorização tradicional em grandes classes de sons, nossa classificação inclui a consoante [s], por representar 91,8% das fricativas e 69,6% de todos os contextos.

GRÁFICO 5 – % de plurais irregulares e onset – Amostra N



O comportamento diferenciado de [s], que concentra 97,4% de seus dados na alternante **ões**, coincide com a grande incidência do sufixo -ção nos dados. Esse fato é explorado do ponto de vista do sufixo na subseção seguinte. Em relação aos demais contextos não se observam predominâncias que justifiquem em princípio a defesa de uma hipótese assimilatória. Na tabela a seguir, exploramos o comportamento lexical das alternantes não prevalentes em relação à consoante precedente. Os únicos contextos que permitem alguma exploração sobre restrição lexical são, no caso da alternante **ãos**, o das oclusivas e nasais, e, no caso de **ães**, o das nasais. Nesses contextos, a proporção lemas/ocorrências é baixa o suficiente para sugerir alguma medida de restrição lexical.

TABELA 6 – Plurais atípicos e contexto precedente – Amostra N

		%	Lemas/ Ocor.	
ãos	hiato	100	14/14	anciãos, aldeãos, anfitriãos, aviãos
	s	98,4	61/62	infecções, condições, informações, alterações
	fricativo	64,5	20/31	órfãos, artesãos, vãos, cortesãos, chãos
	líquido	62,1	18/29	grãos, refrãos, vilãos, catalãos
	oclusivo	**25,6**	**30/117**	órgão, cidadão, cristão, pagão, acórdão
	nasal	**14,8**	**8/54**	**mãos, irmãos, corrimãos, nãos, alemãos**
ães	hiato	92,3	12/13	guardiães, tabeliães, anciães, cirurgiães
	s	97,8	44/45	infecções, condições, informações, alterações
	líquido	88,9	8/9	capelães, catalães, refrães, castelães, tecelães
	fricativo	80	4/5	escrivães, alazães, decisães, lesães
	oclusivo	63,3	19/30	cães, pães, capitães, charlatães, cristães
	nasal	**22,2**	**4/18**	**alemães, irmães, anães, caimães**

5.5 Afiliação morfológica

Como mencionamos, os vocábulos aqui analisados são substantivos e adjetivos, que rotulamos como *nomes*. Considerando-se os propósitos desta etapa de nossa pesquisa, trabalhamos com uma lista de palavras, não de frases, disponibilizada pelo CBras. Ainda que tenhamos quantificado os vocábulos por classe, muitos itens restaram dúbios, já que adjetivos podem facilmente ser empregados como substantivos em português, assim como substantivos podem eventualmente ser empregados como adjetivos.[6] Decidimos, por outro lado, no âmbito das

[6] Embora a busca do item dentro da frase seja acessível aos usuários do CBras por meio do site Linguateca, não recorremos a esse expediente para checar se nomes de classe dúbia haviam sido empregados como substantivos ou como adjetivos, por consideramos essa informação, operacionalmente custosa, pouco relevante para os fins de nossa pesquisa. Essa checagem restringiu-se à desambiguação, como mencionamos na metodologia, de nomes com outras classes apenas (como verbos ou advérbios, por ex.).

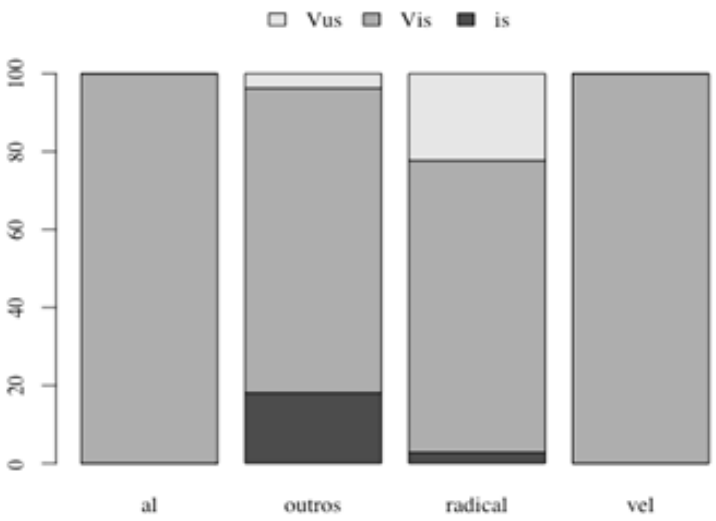
variáveis morfológicas, centrar nossa discussão na afiliação morfológica das alternantes, já que as formas de base para os plurais aqui investigados podem se localizar no radical, coincidir integralmente com sufixos ou, ainda, apenas ser parte de sufixos. Como afixos derivacionais estão relacionados à categorização dos vocábulos, porém, um debate sobre classes pode eventualmente ser desenvolvido a partir do exame desta categoria.

Nesta seção, como na anterior, por conta das peculiaridades de cada amostra no que diz respeito à morfologia, apresentamos seus respectivos resultados separadamente.

Afiliação morfológica na Amostra L

Foram inicialmente examinados 18 contextos, sendo 17 de sufixos potenciais (já que nem todos podem ser seguramente considerados sufixos na sincronia da língua) e formas em que as terminações de base u/l faziam parte do radical (ou raiz acrescida ou não de elementos na borda esquerda). Esses contextos foram reestruturados em apenas 4 categorias, considerando-se que 91,3% dos dados correspondem aos sufixos -al e -vel, distribuindo-se os demais dados entre o radical e outros sufixos comparativamente pouco frequentes.

GRÁFICO 6 – % plurais irregulares e afiliação morfológica da base – Amostra L

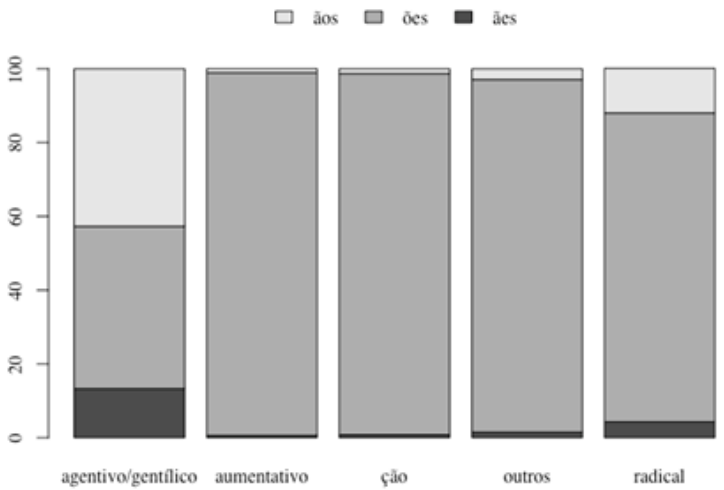


Todos os casos de bases fechadas pelos sufixos -al e -vel fazem plural em **Vis**. Entre outros sufixos e radical, registram-se ocorrências das alternantes **Vus** e **is**. Entre os sufixos que classificamos como *outros* – 4,7% da amostra – estão casos sobretudo de -eu (ex. europeus), ɛu (ex. fogaréus), -ɛl, (ex. carretéis), -ɔl (ex. espanhóis) e -il (ex. febris). A incidência destacada da alternante **is** na categoria *outros* deve-se, em primeiro lugar, às ocorrências do sufixo il, bastante produtivo na formação de adjetivos em português (ex. estudantil, febril, juvenil), mas que nem sempre consegue ser claramente isolado de suas bases (ex. civil, sutil, hostil), e, em segundo lugar, a itens em que il é parte da raiz (ex. perfil, fuzil, refil). A variável *radical* acomoda vocábulos em que não se detecta qualquer processo sufixal (ex. paus, sóis) ou aqueles para os quais a opacidade da fronteira sufixal parece consolidada (ex. ramais, quartéis). Por essa razão, os dados da alternante **Vus** se concentram nessa categoria. Por fim, enfatizamos que o índice para a variável *radical* pode se ampliar se consideramos que parte não desprezível das palavras classificadas em *outros sufixos*, por suspeição da transparência sufixal, poderia estar aptas a migrar para essa categoria.

Afiliação morfológica na Amostra N

A partir de uma divisão inicial em 10 contextos, que procurava dar conta de 8 sufixos e 2 tipos de bases, considerando-se a distribuição dos dados, chegamos a uma classificação que contempla 5 contextos apenas, o radical mais 4 tipos de sufixos. Antes de apreciarmos o gráfico a seguir, que tem por base, como vimos adotando, 100% das ocorrências para cada contexto distribuídas entre as alternantes investigadas, é importante se registrar que 61,4% dos dados da Amostra N dizem respeito ao sufixo -ção (ex. construções), 20,2% ao radical (ex. pães) e 10,9 ao aumentativo (ex. lixões). Os 7,5% de dados restantes distribuem-se entre agentivos/gentílicos (ex. artesãos, catalães) e outros sufixos (ex. colisões).

GRÁFICO 7 – % plurais irregulares e afiliação morfológica da base – Amostra N



Os casos de -ção, de aumentativo e de outros sufixos selecionam a alternante **ões** predominantemente, podendo ser enquadradas suas exceções na pequena margem de variação ou mesmo de erro de registro de escrita a que está sujeito o tipo de dado de que nos utilizamos neste trabalho (ex. durações, informações, tristãos, beatães, decisãos, opiniães).

No caso dos radicais e dos sufixos agentivos/gentílicos, exploramos hipótese diversa: um pouco mais numerosos, os casos atípicos poderiam se referir a restrições lexicais. É o que se explora na tabela 7, que permite identificar indício de restrição lexical no subconjunto dos sufixos agentivos/gentílicos especialmente no que se refere à alternante **ãos**.

TABELA 7 – Plurais atípicos e afiliação morfológica – Amostra N

		%	Lemas/ Ocor.	
ãos	radical	50,7	73/144	órgãos, mãos, grãos, irmãos, órfãos
	agentivo/gentílico	19,4	20/103	cidadãos, cristãos, artesãos, pagãos
ães	radical	80,4	41/51	cães, pães, capitães, tabeliães, refrães
	agentivo/gentílico	46,9	15/32	alemão, guardião, escrivão, capelão

Sobre a relação entre afiliação morfológica e contexto fonológico precedente na Amostra N, cabe apenas um registro sobre a relação -ção/[s]. Trata-se de uma correspondência assimétrica. Obviamente as ocorrências do sufixo -ção (ex. manifestações) coincidem integralmente com as ocorrências de [s]. O contexto precedente [s], porém, corresponde também à maior parte, 37,8%, dos casos de radical (ex. nações), a 4,84% dos casos de sufixo aumentativo (ex. servições) e a 1,66% dos de agentivo/gentílico (todos os casos do lema *saxões*), compreendendo, de modo geral, 8,25% dos dados da Amostra N. Em todos os casos, como vimos, a forma de plural atestada preferencialmente é **ões**. Esses fatos dificultam se afirmar com segurança que o sufixo, e não a consoante [s] (ou sua associação à classe das fricativas), define a seleção da alternante.

5.6 Frequência lexical

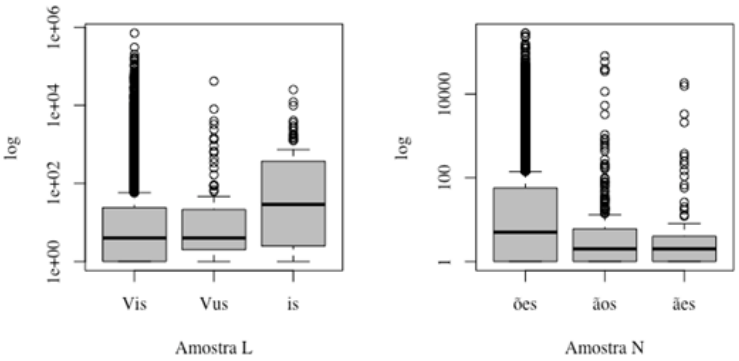
Embora tenhamos discutido em alguma medida o papel de itens lexicais específicos ao longo da apresentação dos resultados das variáveis que investigamos, dedicamos essa seção para uma análise mais geral do efeito da frequência lexical sobre a realização das formas de plural em estudo.

Começamos por listar, com as nuvens de palavras presentes no gráfico 8, os 50 itens mais frequentes de cada amostra.

GRÁFICO 8 – 50 itens de plural irregular mais frequentes – Amostras L e N

No gráfico 9, tratamos da frequência dos itens lexicais com maior detalhamento para cada uma das amostras, fazendo uso de uma escala logarítmica. A motivação para a adoção dessa escala é a discrepância de dados absolutos da variante **Vis** sobre **Vus** e **is**, na Amostra L, e da variante **ões** sobre **ãos** e **ães**, na Amostra N.

GRÁFICO 9 – Plurais irregulares e frequência lexical – Amostras L e N



Analisamos primeiramente os resultados para a Amostra L. Em relação à posição, embora a média de frequência para as três alternantes seja relativamente equilibrada, a despeito da grande diferença de dados, observamos que a mediana se iguala para as alternantes **Vis** e **Vus**, situando-se na frequência 4, mas se eleva a 26,5 para alternante **is**. Apesar disso, é **Vus** que apresenta assimetria positiva, concentrando dados no terceiro quartil. Quanto à dispersão, as variantes apresentam importante diferença de amplitude, que reflete a distribuição da amostra, mas também a grande incidência de *outliers* (dados que se afastam do padrão geral de distribuição) na porção superior dos dados. O intervalo interquartílico entre as variantes **Vis** e **Vus**, apesar da grande distinção no número total de dados, se assemelha, mas a alternante **is** destoa importantemente das demais. Por fim, merecem destaque as caudas do limite inferior: o segundo quartil se inicia no limite mínimo de frequência lexical, 1, para a alternante **Vis**, mas em 2 para as variantes **Vus** e **is**. As caudas superiores se assemelham para as três alternantes, e o limite superior é equilibrado para **Vis** e **Vus**, situando-se entre as frequências 50 e 60, mas

é consideravelmente maior para **is**, alcançando 884. A tabela 8 contribui para a melhor leitura desses dados.

TABELA 8 – Plurais irregulares e frequência lexical em quantis – Amostra L

Alternante	Limite inferior	25%	50%	75%	Limite superior	100%	Média
Vis	1	1	4	24	58,5	710.897	722,2
Vus	1	2	4	21,3	50,3	41.745	692,3
is	1	2	26,5	354,8	884	25.373	900.64

Interpretando esses resultados na perspectiva da frequência lexical, podemos dizer que os números obtidos para a Amostra L conferem pouco destaque às alternantes **Vis** e **Vus** no que concerne à frequência lexical, mas também que não diferenciam essas alternantes entre si de modo importante. Isso, associado ao comportamento diferenciado de **is**, que conta com mais itens frequentes do que suas concorrentes na porção central da amostra, sugere algum controle à generalização de **Vis** como marca de plural de palavras terminadas em u/l, apesar de sua indiscutível prevalência na língua.

Analisamos agora os resultados para a Amostra N. No que concerne à posição, se, por um lado, as médias das duas primeiras alternantes, **ões** e **ãos**, não se mostram tão distanciadas no comparativo com a terceira alternante, **ães**, por outro, a mediana afasta a primeira, com frequência lexical 5, das duas últimas, com frequência lexical 2. As três alternantes apresentam algum grau de assimetria positiva, com mais itens no terceiro quartil, com destaque para **ões**, alternante em que o índice se eleva mais nesta porção. Em relação à dispersão, como na Amostra L, há aqui importante diferença de amplitude, com *outliers* na porção superior para as três alternantes, acompanhando os índices de prevalência dessas variantes no uso. O intervalo interquartílico também é consideravelmente maior para a variante **ões**. O limite inferior é idêntico para as três variantes, coincidindo o início do segundo quartil com a frequência lexical mínima da amostra, 1. As caudas superiores também não se distinguem importantemente, sendo inclusive proporcionalmente idênticas para a primeira e a terceira variantes, respectivamente a mais e a menos prevalentes da Amostra. O limite superior, contudo, é sensivelmente

distinto para as três variantes, acompanhando o caráter decrescente de emprego dessas alternantes. A tabela 9 detalha esses valores.

TABELA 9 – Plurais irregulares e frequência lexical em quantis – Amostra N

Alternante	Limite inferior	25%	50%	75%	Limite superior	100%	Média
ões	1	1	5	57	141	284.580	1.183,9
ãos	1	1	2	6	13,5	80.948	906,7
ães	1	1	2	4	8,5	18.386	342,8

Os resultados obtidos para a Amostra N não sugerem predominância de itens mais frequentes para as variantes marginais, **ãos** e **ães**. Isso somado ao fato de que a variante prevalente no uso, **ões**, concentra também itens lexicais mais frequentes contribui para a ideia de generalização dessa variante como marca de plural de palavras terminadas em ão na língua.

6 Considerações finais

Neste texto apresentamos resultados de um estudo descritivo sobre a expressão fonológica de formas de plural irregular em português, no âmbito da palavra, com base em dados do Corpus Brasileiro. Duas amostras foram consideradas: Amostra L, relativa ao plural de nomes terminados, no singular, em vogal+u/l ortográficos, a que correspondem as alternantes **Vis**, **Vus** e **is**, e Amostra N, relativa ao plural de nomes terminados em ão ortográfico, a que correspondem as alternantes **ões**, **ãos** e **ães**. Variáveis fonológicas e léxico-morfológicas foram quantificadas em relação a cada uma dessas alternantes.

O estudo confirmou a prevalência, amplamente relatada na literatura, das variantes **Vis** e **ões**, respectivamente, para as Amostras L e N. Dedicamos, por isso, nossa maior atenção às variantes menos frequentes. Em relação a essas variantes, merecem destaque os aspectos a seguir resumidos.

(i) Em relação ao número de sílabas, monossílabos mostram comportamento diferenciado nas duas amostras, como relatado por Huback (2010a, 2010b); Becker *et al.* (2018), Rizzato (2018), entre

outros. Na Amostra L, o destaque é para **Vus**, que se aproxima muito em emprego da variante predominante, **Vis**. Na Amostra N, destaca-se **ãos**, com grande vantagem em relação às demais alternantes, seguido de **ães**. O contraste entre ocorrências e lemas sugere que os padrões excepcionais encontrados entre os monossílabos são, em grande medida, restritos lexicalmente.

(ii) Quanto ao acento, a grande maioria dos dados é de oxítonas, o que apenas confirma uma hipótese geral sobre a preferência do padrão acentual do português, já que estamos diante de sílabas supostamente pesadas (admitindo-se uma consoante ou mesmo um glide em coda). Não há proparoxítonas. As reduzidas paroxítonas atestadas, no caso da Amostra L, seguem o padrão predominante, selecionando **Vis**, e se referem, majoritariamente a formas sufixadas (sincrônica ou diacronicamente). No caso da Amostra N, as paroxítonas selecionam predominantemente **ãos**, em ocorrências bastante restritas lexicalmente.

(iii) Em relação ao contexto fonológico, examinamos, no caso da Amostra L, a vogal que nucleariza a sílaba envolvida em cada alternante. Para todas as vogais há predomínio da alternante **Vis**, com exceção de [i], que realiza **is**, por restrição estrutural, já que todas as formas dizem respeito à terminação il. As vogais [a] e [ɛ] estão presentes em dados com a alternante **Vus**, sem indícios de restrição lexical, dada a alta proporção lema/ocorrência nessas categorias. Esse comportamento das vogais se confirma quando se contrastam padrões de terminação das formas pluralizadas a padrões de terminação das bases. No caso da Amostra N, porque a vogal nuclear, [ã], é comum à forma singular das três alternantes, examinamos a consoante que precede esta vogal. O destaque é para [s], que corresponde a quase 70% dos dados, distribuído entre as três variantes, com privilégio para **ões**. As demais consoantes distribuem-se de modo relativamente equilibrado. Identifica-se controle lexical apenas no caso de oclusivas e nasais precedendo **ãos** e de nasais precedendo **ães**.

(iv) No que concerne à afiliação morfológica das alternantes, na Amostra L, 91,3% dos dados são de palavras terminadas nos sufixos -al ou -vel, todos realizando a variante **Vis**. Os demais dados distribuem-se entre radical e outros sufixos. Desses, destacam-se os dados do sufixo -il, todos realizando **is**, e dos ditongos localizados na raiz, que realizam **Vus**. Na Amostra N, 61,4% dizem respeito ao sufixo -ção e 20,2%, ao radical. Os demais itens são fechados por sufixos aumentativos, por agentivos/gentílicos e outros sufixos minoritários. Em todos os casos predomina

a alternante **ões**, à exceção dos agentivos/gentílicos, grupo em que se destacam as alternantes **ãos** e **ães**, com alguma restrição lexical. As alternantes **ãos** e **ães**, embora em número mais reduzido e sem aparente controle lexical, também se destacam no contexto dos radicais.

(v) Por fim, quanto à frequência lexical geral das amostras estudadas, podemos dizer, em relação à Amostra L, que o comportamento similar das alternantes **Vis** e **Vus**, apesar de sua considerável diferença em termos de ocorrências, combinado ao comportamento próprio de **is** podem indicar alguma resistência para a generalização de **Vis** como marca de plural de palavras terminadas em u/l. No caso da Amostra N, a variante mais recorrente, **ões**, é também a que concentra maior frequência lexical, sem destaque para o comportamento de **ãos** e **ães**. Esses fatos contribuem para a ideia de generalização dessa variante como marca de plural de palavras terminadas em ão na língua.

Este estudo integra um projeto maior, que trata da representação de plurais irregulares em português brasileiro. Sua contribuição descritiva se dá pelo mapeamento de itens do léxico a partir de um banco com grande volume de dados de uso falado e escrito da língua. Esse mapeamento é ponto de partida da análise experimental em curso, em que discutimos, em perspectiva inferencial, o uso dessas alternantes e, sobretudo, suas representações de base.

Agradecimento

Agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processos PQ 310921/2018-0 e PIBIC 154093/2020-3, e à Fundação de Amparo à Pesquisa do Estado do Rio Grande do Sul (FAPERGS), processo PROBIC 20/2551-0000315-9, pelo fomento em forma de bolsas. Agradecemos aos colegas Pedro Surreaux, Rodrigo Mahfuz e Júlia Ricardo.

Contribuição dos Autores

Este artigo corresponde a um recorte de pesquisa sobre representação de plurais no português brasileiro, concebida e coordenada pelo primeiro autor. Os três autores participaram das etapas de levantamento e análise dos dados, de discussão dos resultados e de redação deste texto.

Referências

ABAURRE GNERRE, M. B. M. Alguns casos de formação de plural em português: uma abordagem natural. *Cadernos De Estudos Linguísticos*, Campinas, v. *5*, p. 127-156, 1983.

BAUER, L. *A Glossary of Morphology*. Edinburgh: Edinburgh University Press, 2004.

BECKER, M. *et al*. The Acquisition Path of [w]-final Plurals in Brazilian Portuguese. *Journal of Portuguese Linguistics*, Lisboa, v. 17, n. 4, p. 1-17, 2018. DOI: https://doi.org/10.5334/jpl.189. Disponível em: https://jpl.letras.ulisboa.pt/articles/10.5334/jpl.189/. Acesso em: 11 set. 2020.

BECKER, M.; NEVINS, A.; LEVINE, J. Asymmetries in Generalizing to and from Initial Syllables. *Language,* Washington, DC, v. 88, n. 2, p. 231-268, 2012. DOI: https://doi.org/10.1353/lan.2012.0049. Disponível em: https://becker.phonologist.org/projects/english/becker_nevins_levine_english_2012.pdf. Acesso em: 11 set. 2020.

BISOL, L. O ditongo na perspectiva da fonologia atual. *DELTA,* São Paulo, v. 5, n. 2, p. 185-224, 1989.

BISOL, L. A nasalidade, um velho tema. *DELTA*, São Paulo, v.14, nº especial, p. 27-46, 1998. DOI: https://doi.org/10.1590/S0102-44501998000300004. Disponível em: https://revistas.pucsp.br/delta/article/view/43390/28850. Acesso em: 11 set. 2020.

BISOL, L. A nasalidade fonológica no português e suas restrições. *Diadorim*, Rio de Janeiro, v. 18, p. 116-126, 2016. DOI: https://doi.org/10.35520/diadorim.2016.v18n0a4050. Disponível em: https://revistas.ufrj.br/index.php/diadorim/article/view/4050. Acesso em: 11 set. 2020.

BISOL, L. Sufixos de duas faces. *Revista da Abralin*, Aracaju, v. 19, n. 1, p. 1-12, 2020. DOI: HTTPS://doi.org/10.25189/rabralin.v19i1.1380. Disponível em: https://revista.abralin.org/index.php/abralin/article/view/1380. Acesso em: 11 set. 2020.

BONET, E.; LLORET, M. R.; MASCARÓ, J. The Prenominal Allomorphy Syndrome. *In*: _____. (org.). *Understanding Allomorphy.* Perspectives from Optimality Theory. Bristol: Equinox Publishing, 2015. v. 5, p. 1-44.

BYBEE, J. *Phonology and Language Use.* Cambridge: Cambridge University Press, 2001. DOI: https://doi.org/10.1017/CBO9780511612886

CAMARA JR., J. M. *Problemas de Lingüística Descritiva*. Petrópolis: Editora Vozes, 1969.

CAMARA JR., J. M. *Estrutura da Língua Portuguesa.* 35. ed. Rio de Janeiro: Editora Vozes, 1970.

CRISTÓFARO-SILVA, T. Organização fonológica de marcas de plural no português brasileiro: uma abordagem multirrepresentacional. *Revista da Abralin*, Curitiba, v. 11, p. 273-305, 2012. DOI: 10.5380/rabl.v11i1.32468. Disponível em: https://revistas.ufpr.br/abralin/article/view/32468. Acesso em: 11 set. 2020.

GOMES, C. A., PRADO, L. O. do; AMARAL, T. L. A. Aspectos cognitivos e sociais da variação linguística na alternância de formas de plural de nomes do PB. *In*: ORSINI, M.; CAVALCANTE, S. R.; MARINS, J. (org.). *Contribuições à descrição e ao ensino do português brasileiro:* da fonética ao discurso, com parada obrigatória na sintaxe (título provisório). Rio de Janeiro: EDUFRJ, 2021. No prelo.

GUIMARÃES, M.; NEVINS, A. Probing the Representation of Nasal Vowel in Brazilian Portuguese with Language Games. *ORGANON,* Porto Alegre, v. 28, n. 54. p. 155-178, 2013. DOI: https://doi.org/10.22456/2238-8915.38298. Disponível em: https://seer.ufrgs.br/organon/article/view/38298. Acesso em: 11 set. 2020.

HUBACK, A. P. *Efeitos de frequência nas representações mentais.* 2007. 318 p. Tese. Faculdade de Letras, Universidade Federal de Minas Gerais, 2007.

HUBACK, A. P. Plurais irregulares do português brasileiro: efeitos de frequência. *Revista da Abralin*, Curitiba, v. 9, n. 1, p. 11-40, 2010a. DOI: https://doi.org/10.5380/rabl.v9i1.52337. Disponível em: https://revistas.ufpr.br/abralin/article/view/52337/32236. Acesso em: 11 set. 2020.

HUBACK, A. P. Plurais em -ão do português brasileiro: efeitos de frequência. *Revista Linguística*, Rio de Janeiro, v. 6, n. 1, p. 9-26, 2010b. DOI: https://doi.org/10.31513/linguistica.2010.v6n1a4436

RIZZATO, É. *Intera*ção do plural de -ão e do aumentativo -zão na formação de compostos no português brasileiro. 2018. 94 f. Dissertação (Mestrado em Linguística) – Instituto de Estudos da Linguagem, Universidade Estadual de Campinas, Campinas, 2018.

SCHWINDT, L. C. S. *et al*. A influência da variável escolaridade em fenômenos fonológicos variáveis: efeitos retroalimentadores da escrita. *Revista Virtual de Estudos da Linguagem – ReVEL*, [*S.l.*], v. 5, n. 9, p. 1-12, 2007. Disponível em: https://www.lume.ufrgs.br/bitstream/handle/10183/184784/000640837.pdf?sequence=1. Acesso em: 11 set. 2020.

WETZELS, L. The Lexical Representation of Nasality in Brazilian Portuguese. *Probus*, [*S.l.*], v. 9, p. 203-232, 1997.

WETZELS, L. Comentários sobre a estrutura fonológica dos ditongos nasais no Português do Brasil. *Revista de Letras*, Fortaleza, v. 1, n. 22, p. 25-30, 2000. DOI: https://doi.org/10.1515/prbs.1997.9.2.203. Disponível em: http://www.revistadeletras.ufc.br/rl22Art03.pdf. Acesso em: 11 set. 2020.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1325-1379, 2021

Uma proposta de coextensividade entre termo técnico, grupo nominal e item lexical no português brasileiro: um estudo com base em ferramentas da linguística de corpus sob o arcabouço de teoria sistêmico-funcional

A proposal of coextensiveness between technical term, nominal group, and lexical item in Brazilian Portuguese: a study based on corpus linguistics' software within the framework of systemic-functional theory

Júlia Santos Nunes Rodrigues
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais / Brasil
juliasnrodrigues@ufmg.br
http://orcid.org/0000-0002-7673-1833

Kícila Ferreguetti
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais / Brasil
Kfo2008@ufmg.br
http://orcid.org/0000-0002-1919-0073

Adriana S. Pagano
Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais / Brasil
apagano@ufmg.br
http://orcid.org/0000-0002-3150-3503

Resumo: Objetivo: Sob a perspectiva do trabalho de Pearson (1998) e utilizando as ferramentas da linguística de corpus disponíveis para o português brasileiro, a pesquisa apresentada neste artigo busca verificar em que medida a coextensividade entre termo técnico, grupo nominal e item lexical pode ser considerada válida para o português brasileiro. Método: Para a verificação desta coextensividade, um corpus de artigos acadêmicos sobre o domínio experiencial do autocuidado em Diabetes Mellitus foi compilado eletronicamente. Esse corpus foi inserido no software concordanciador AntConc (ANTHONY, 2019) e três palavras-chave foram extraídas com base no

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1325-1379

corpus de referência CALIBRA (FIGUEREDO; PAGANO; FERREGUETTI, 2014). O contexto de cada uma dessas palavras foi analisado por meio da ferramenta clusters/n-grams do AntConc, considerando os clusters/n-grams com o número mínimo de dez ocorrências. Resultados: A investigação desses clusters/n-grams formados à direita e à esquerda de cada uma das palavras-chave selecionadas mostrou que a coextensividade entre termo técnico, grupo nominal e item lexical nem sempre pode ser identificada, embora a utilização da ferramenta de cluster/n-grams pode ser considerada eficaz para buscar por itens lexicais que estão em coextensividade à ordem do grupo nominal, em razão da existência de pelo menos um grupo nominal em cada cluster/n-gram analisado. Conclusão: Ainda que os programas utilizados em Pearson (1998) não estejam totalmente difundidos para o português brasileiro, a abordagem sistêmico-funcional para o grupo nominal e para o item lexical em conjunto com as ferramentas do software concordanciador utilizado se mostraram eficientes para a análise proposta neste artigo.

Palavras-chave: item lexical; grupo nominal; termo técnico; linguística de corpus; sistêmico-funcional; co-extensividade.

Abstract: Objective: Drawing on Pearson (1998) and using corpus linguistics tools available for Brazilian Portuguese, we report on a study aimed at exploring to what extent the concepts of technical term, nominal group and lexical item are coextensive in Brazilian Portuguese. Method: A corpus of academic articles on the experiential domain of Diabetes Mellitus self-care was compiled and queried in AntConc, a concordancing software (ANTHONY, 2019). Using as a reference corpus CALIBRA (FIGUEREDO; PAGANO; FERREGUETTI, 2014), three keywords were extracted analysed with AntConc tool clusters/n-grams, considering clusters/n-grams with a minimum number of ten occurrences. Results: Analysis of clusters/n-grams to the right and left of each of the selected keywords showed that technical term, nominal group and lexical item cannot always coextensive. The use of cluster/n-grams tool can be considered effective to search for lexical items that are coextensive to the order of the nominal group, due to the existence of at least one nominal group in each cluster/n-gram analyzed. Conclusion: Although the programs used by Pearson (1998) are not fully available to Brazilian Portuguese, a systemic-functional approach to nominal group and lexical item together with the tools of the concordancing software used in this paper proved to be efficient for the analysis herein proposed.

Keywords: lexical item; nominal group; technical term; corpus linguistics; systemic functional theory; coextensiveness.

Recebido em 10 de outubro de 2020

Aceito em 07 de janeiro de 2021

1 Introdução

O registro de uso de corpus é antigo. Na Grécia Antiga, foi criado o Corpus Helenístico de Alexandre, o Grande. Na Antiguidade e Idade Média, os corpora com citações da Bíblia foram desenvolvidos. No século XX, muitos pesquisadores utilizavam os corpora para trabalhos de descrição da linguagem. No entanto, apesar desses registros, o uso de corpora nessas fases era restrito ao aprendizado de línguas, sendo que todo o processo de elaboração dos corpora era feito de forma manual em razão da inexistência de recursos tecnológicos (SARDINHA, 2004, p. 3).

Os corpora como conhecemos hoje, com milhões de palavras, com textos compilados, majoritariamente, de forma automática, construídos para suprir diversas demandas linguísticas e sendo viabilizados por softwares de diferentes origens tem início com o lançamento do corpus Brown (Brown Corpus of Standard American English) no início da década de 1960 (SARDINHA, 2004). Desde essa época até os dias atuais, a Linguística de Corpus tem evoluído muito, principalmente a partir do uso de computadores pessoais nos anos de 1980.

Hoje em dia já existem corpora compilados para uma grande variedade de línguas, corpora utilizados para diferentes finalidades, como, tradução, criação de dicionários e gramáticas, processamento de linguagem natural, terminologia, etc. Há também softwares livres, como o AntConc (ANTHONY, 2019), por exemplo, que auxiliam pesquisadores no desenvolvimento de estudos que utilizam corpora. Contudo, a evolução tecnológica no âmbito da Linguística de Corpus não pode ser vista de forma homogênea para todas as línguas. No âmbito da língua inglesa, por exemplo, sobretudo no contexto britânico, a Linguística de Corpus tem uma disponibilidade maior de recursos tecnológicos em razão de investimentos financeiros em pesquisas dessa área, bem como pelo fato dos estudos de corpora, como conhecemos hoje, terem iniciado no contexto dessa língua. Esse investimento em tecnologia pode ser observado com clareza no trabalho de Pearson (1998), que serviu de base para o estudo do presente artigo, haja vista o desenvolvimento de um anotador morfossintatico (CLG tagger) e de um programa de padrão de correspondência para aquele trabalho.

Apesar dessas limitações, este artigo tem como principal objetivo apresentar soluções que podem ser utilizadas como forma de diminuir o abismo existente entre os recursos tecnológicos disponíveis para o

contexto da língua inglesa e ainda incipientes para o português brasileiro. Essas soluções baseiam-se ora em ferramentas da própria Linguística de Corpus ora no suporte de outras teorias linguísticas, no caso, a teoria sistêmico-funcional (FERREGUETTI, 2018; FIGUEREDO, 2007; HALLIDAY; MATTHIESSEN, 2014).

Nesse sentido, a pesquisa detalhada no presente artigo faz uso de um corpus de artigos acadêmicos sobre autocuidado em Diabetes Mellitus, escritos em português brasileiro, para extração de termos técnicos dessa área do conhecimento, por meio das ferramentas de lista de palavras-chave (Keyword List), clusters e n-grams (Clusters/N-grams) e lista de concordância (Concordance) do software concordanciador AntConc (ANTHONY, 2019). As análises do trabalho em questão são pautadas na teoria sistêmico-funcional principalmente na propriedade estabelecida por Halliday (2002, p. 59-60) denominada co-extensividade (coextensiness), a qual pode ocorrer entre o item lexical e a escala de ordens da gramática – morfema, palavra, grupo/frase preposicional e oração – sobretudo no que se refere às ordens da palavra e do grupo/frase preposicional (FERREGUETTI, 2018; FIGUEREDO, 2007).

Partindo dessa propriedade, espera-se verificar em que medida a correspondência entre termo técnico, grupo nominal e item lexical pode ser considerada válida para o português brasileiro. Em paralelo à verificação dessa hipótese, soluções de como os recursos tecnológicos desenvolvidos para/pela a linguística de corpus alinhados ao arcabouço linguístico da teoria sistêmico-funcional são testadas a fim de suprimir recursos da linguística de corpus difundidos para a língua inglesa (cf. PEARSON, 1998) e pouco (ou nada) disseminados para o português brasileiro.

2 Fundamentação teórica

2.1 A Linguística de Corpus e o estudo de termos técnicos

Dentre as diferentes vertentes linguísticas, a Linguística de Corpus e a Terminologia são conhecidas como subáreas da linguística cujos objetivos estão relacionados ao estudo de termos técnicos sob diferentes abordagens. Na obra "Linguística de Corpus", Sardinha (2004) afirma que:

> A Linguística de Corpus ocupa-se da coleta e da exploração de corpora, ou conjunto de dados linguísticos textuais coletados criteriosamente, com o próposito de servirem para a pesquisa de uma língua ou variedade linguística. Como tal, dedica-se à exploração da linguagem por meio de evidências empíricas, extraídas por computador (SARDINHA, 2004, p. 3).

Sob essa perspectiva, a Linguística de Corpus lida com termos técnicos em seu contexto e cotexto de uso, ocupando-se com o desenvolvimento de ferramentas que consigam examinar uma grande quantidade de textos de maneira automática (ALMEIDA; CORREIA, 2008).

Por outro lado, a Terminologia, a partir do século XVII, pode ser entendida sob dois vieses diferentes: i. "Como conjunto de termos de uma área técnica ou científica" ou ii. "Como disciplina de natureza linguística que estuda esse conjunto de termos" (ALMEIDA, 2004, p. 31). Cada vez mais as teorias de Terminologia entendem o texto especializado de uma determinada área do conhecimento como seu principal ponto de estudo, sendo o termo técnico compreendido como:

> uma condição especial da palavra, um signo linguístico dotado de significado e significante, e atrelado a uma determinada unidade e corpo de conhecimentos historicamente estabelecidos. Desse modo, terminologias deixam de ser unidades "estranhas" ou "artificiais", índices de uma língua à parte da língua-sistema, e passam a ser vistas como palavras que têm ou adquirem um estatuto peculiar em uma dada situação de comunicação (FINATTO, 2007, p. 224).

Apesar da Linguística de Corpus e da Terminologia serem subáreas da linguística, cada uma apresenta uma série de particularidades que inviabiliza a sobreposição de uma subárea sobre a outra. Uma das implicações diz respeito ao termo técnico que para a Linguística de Corpus precisa apresentar uma determinada frequência para ser considerado um possível termo técnico, bem como necessita ter seu contexto analisado para certificar-se que se trata de um termo técnico, ao passo que para a Terminologia a frequência de um dado termo não compreende um critério de seleção para que esse termo seja entendido como técnico (FINATTO, 2007). Outra implicação compreende a elaboração dos corpora compilados para os estudos terminológicos, visto

que a extração de termos considerados candidatos a técnicos é diretamente influenciada pela forma como os textos dos corpora são manipulados pelos pesquisadores.

Do ponto de vista da complementariedade, no entanto, os recentes avanços tecnológicos relacionados à Linguística de Corpus se mostram relevantes também para o progresso da Terminologia enquanto disciplina. O desenvolvimento e o aprimoramento de ferramentas computacionais podem viabilizar a gestão e manipulação de grandes bases textuais, beneficiando as pesquisas terminológicas baseadas em corpora (ALMEIDA; CORREIA, 2008).

Ainda no tange a complementariedade entre essas duas subáreas, está a distinção conceitual entre palavra e termo, uma vez que os manuais de Terminologia designam termo como aquilo que geralmente é caracterizado como palavra. Contudo, por meio da análise de corpora, o pesquisador consegue ter acesso ao contexto e cotexto em que determinado termo aparece, o que pode favorecer a seleção desse termo como candidato a termo técnico (ALMEIDA; CORREIA, 2008).

A subseção seguinte apresenta o trabalho que norteou a pesquisa abordada no presente artigo. Esse estudo deixa claro como os avanços tecnológicos da Linguística de Corpus podem beneficiar diretamente as pesquisas na área de Terminologia.

2.2 O trabalho de Pearson (1998)

A pesquisa desenvolvida por Jennifer Pearson publicada no livro *Terms in Context* de 1998 teve como um dos principais objetivos estabelecer parâmetros metodológicos que fossem capazes de auxiliarem terminologistas, lexicográfos e linguistas de corpus a lidar com termos técnicos de uma forma mais automatizada. Para além disso, a autora explica conceitos técnicos inerentes à Linguística de Corpus, estabelece as diferenças entre palavra e termo técnico e apresenta um panorama de metodologias utilizadas por outras pesquisas que também enfocaram na definição e extração de termos.

Três corpora com domínios distintos foram utilizados nesse trabalho. O primeiro chamado de "Nature corpus", com 230.000 palavras, lida com o uso especializado da linguagem, visto que envolve artigos acadêmicos publicados no periódico Nature ao longo do ano de 1989. O segundo denominado "International Telecommunications Union (ITU)

corpus", com 4.7 milhões de palavras, engloba os textos do manual da União Internacional de Telecomunicações da Europa (International Telecommunications Union CCITT Handbook) conhecido como "The Blue Book", escrito por membros do ITU para aprendizes da área. O terceiro nomeado como "GCSE corpus", com um milhão de palavras, compreende uma série de livros didáticos sobre história, geografia, biologia, química, sociologia e política, cujo intuito é promover o ensino de disciplinas do currículo escolar.

Encerrado o processo de compilação desses corpora, um anotador morfossintático, CLG tagger, desenvolvido pelo grupo de Linguística de Corpus da Universidade de Birmingham, foi utilizado para gerar uma anotação automática das classes de palavra de cada um dos três corpora.

Em seguida, uma busca manual por sinais linguísticos, como, "i.e." ou "e.g." foi realizada em cada um dos corpora. Isso se deu pela hipótese levantada sobre a possibilidade de coocorrência de termos técnicos com alguns sinais linguísticos específicos.

Essa busca manual por padrões de coocorrência de termos técnicos e sinais linguísticos resultou em conjuntos de padrões que indicavam como os termos técnicos eram formados para cada um dos três corpora. Esses padrões eram formados por meio de uma sequência de etiquetas, como, por exemplo, adj + noun + noun. Esses conjuntos de padrões foram inseridos em um programa de padrão de correspondência que fora treinado para selecionar nos textos de cada corpus palavras que estavam em concordância com aqueles padrões obtidos manualmente, ou seja, o programa de padrão de correspondência rastreava os textos de cada corpus e extraía todos os conjuntos de palavras que se encaixavam nos *inputs* dados pelos pesquisadores a partir dos padrões obtidos manualmente. Então, esses padrões gerados automaticamente pelo programa de padrão de correspondência eram concebidos pelos pesquisadores como possíveis termos técnicos.

O próximo passo dessa pesquisa foi a investigação desses possíveis termos técnicos com o propósito de refinar quais deles poderiam ser considerados termos técnicos de fato. Para isso, os pesquisadores responsáveis lançaram mão das seguintes estratégias:

Referência genérica: a presença ou ausência de referência endofórica no cotexto em que determinado termo aparece pode levar esse termo a ser considerado genérico ou individual, neste último caso, o termo poderia ser realizado pelo próprio nome, não por um termo qualquer.

Termo sinalizado: o possível termo técnico não pode ser precedido por uma série de determinantes.

Termo não-sinalizado: o possível termo técnico pode não ser precedido por um determinante nem ser precedido apenas por artigos indefinidos. No entanto, somente essa estratégia não é suficiente para definir se um dado termo é técnico de fato. Tal estratégia só é considerada suficiente se o termo em questão está em coocorrência com algum dos sinais linguísticos a seguir:

i. "por exemplo", "ou seja";
ii. "chamado de", "conhecido como", "denominado por", "também chamado de", "geralmente conhecido como" e
iii. Obedecer à construção: "possível termo técnico" seguido de "o termo" ou "esse processo" ou "esse método" ou "esse instrumento", etc.

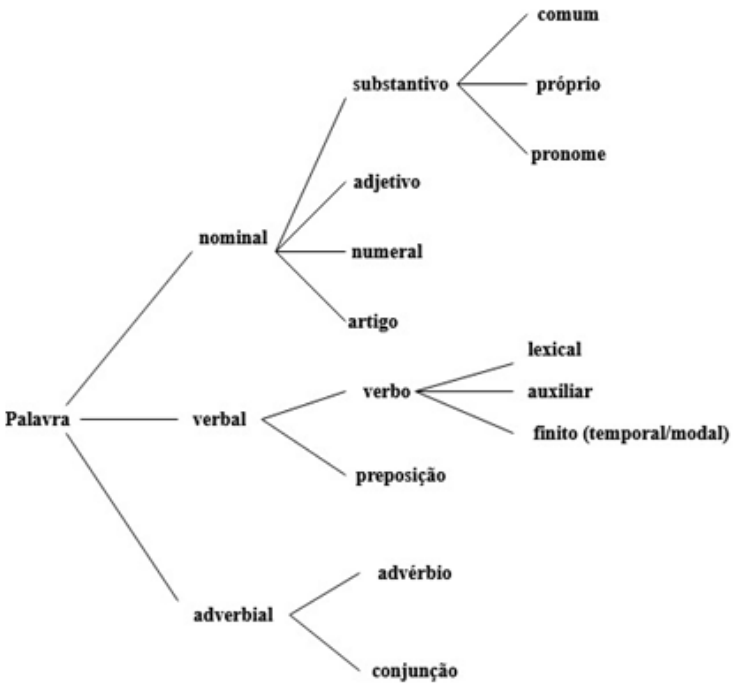
Por fim, Pearson detalha os resultados encontrados para cada um dos três corpora, bem como aponta dificuldades e limitações relacionadas à pesquisa de extração de termos técnicos em corpora especializados.

2.3 O termo técnico realizado por grupo nominal sob a perspectiva da TSF

No âmbito da teoria sistêmico-funcional (TSF), a gramática é organizada em uma escala de ordens hierárquicas em que cada ordem é formada pelas ordens que estão imediatamente abaixo. O português brasileiro conta com quatro ordens: oração, grupo, palavra e morfema. Nesse sentido, a oração é constituída por grupos, que por sua vez, são constituídos por palavras, que são constituídas por morfemas (cf. FERREGUETTI, 2018; FIGUEREDO, 2007).

Ainda com relação ao sistema linguístico do português, em um primeiro nível de delicadeza (delicacy), há grupos de cinco tipos: nominal, verbal, adverbial, conjuntiva e preposicional (FIGUEREDO, 2007). Isso significa dizer que para cada grupo as palavras que os constituem são, geralmente, da classe correspondente, ou seja, as palavras que constituem o grupo nominal, por exemplo, são, na sua maioria, da classe de palavras nominal. A Figura 1 apresenta a distribuição de classes de palavra de acordo com as concepções da TSF.

FIGURA 1 – Distribuição das classes de palavra segundo a TSF



Fonte: Traduzido e Adaptado de Halliday e Matthiessen (2014, p. 75).

A Figura 1 identifica que as três principais classes de palavra são nominal, verbal e adverbial. As palavras nominais compreendem os substantivos (comum, próprio e/ou pronome), adjetivos, numerais e/ou artigos. As palavras verbais englobam os verbos (lexical, auxiliar e/ou finito) e as preposições. E as palavras adverbiais envolvem os advérbios e as conjunções.

No que tange o grupo nominal, objeto de discussão desta seção, Halliday e Matthiessen (1999) afirmam que o grupo nominal apresenta duas funções primárias denominadas como Qualidade e Ente. Essas funções são responsáveis pela ideia de permanência que é intrínseca ao grupo nominal. Em outras palavras, os elementos permanentes são capazes de se repetirem ao longo do texto, por isso são mais duradouros e podem participar de eventos distintos. Em contrapartida, os elementos transitórios frequentemente realizados pelo grupo verbal representam os eventos do texto (FIGUEREDO, 2007).

No que diz respeito à estrutura do grupo nominal em português brasileiro, as funções de Ente e de Qualidade são as duas principais. O Ente representa, linguisticamente, os seres do mundo, é o núcleo semântico do grupo nominal e pode ser realizado pelos substantivos (comum, próprio ou pronome); e a Qualidade é responsável por definir qual o subconjunto de seres o escritor ou falante daquele texto se refere. Para além dessas duas funções, o grupo nominal pode contar ainda com as funções do Dêitico, Numerativo, Epíteto, Classificador e Qualificador. O Dêitico determina um subconjunto do Ente, sendo que essa determinação pode ser em relação à definição, especificidade e/ou localização. O Numerativo aponta alguma característica numérica ao subconjunto do Ente. O Epíteto indica alguma qualidade do Ente, a qual pode englobar traços objetivos do subconjunto do Ente, bem como alguma avaliação do falante. O Classificador envolve a relação de hiponímia e é responsável por delimitar o Ente em relação a uma subclasse. O Qualificador tem a função de caracterizar o Ente do grupo nominal por meio de frase preposicional,[1] oração encaixada ou oração não-finita (FIGUEREDO, 2007; HALLIDAY; MATTHIESSEN, 2014). O Quadro 1 mostra as classes de palavra mais prováveis para cada uma das funções do grupo nominal.

QUADRO 1 – Funções do grupo nominal em PB e as respectivas classes de palavra mais frequentes

Função	**Classes de palavra mais provável**		**Outras possibilidades de ocorrência**	
Ente	substantivo	pronome pessoal	verbo (fenômeno)	
Dêitico	artigo	pronome		
Numerativo	numeral		pronome	
Epíteto	adjetivo		verbo	
Classificador	adjetivo	substantivo	pronome	numeral

Fonte: Figueredo (2007, p. 226).

[1] De acordo com a Teoria Sistêmico-funcional, a frase preposicional corresponde à ordem localizada entre a palavra e oração na escala de ordem da gramática. Ela é constituída por uma preposição + um grupo nominal (cf. FERREGUETTI, 2018; HALLIDAY; MATTHIESSEN, 2014).

Os dados do Quadro 1 revelam que para todas as funções do grupo nominal as palavras relacionadas à classe nominal são mais prováveis de aparecer nos grupos nominais em português brasileiro. É importante ressaltar que os verbos são as palavras localizadas fora da classe nominal que apresentam ocorrências no grupo nominal. Isso pode ser explicado pelas metáforas gramaticais, frequentes em determinados tipos de texto (cf. *grammatical metaphor* – HALLIDAY; MATTHIESSEN, 2014).

Tendo detalhado os principais pontos teóricos relacionados ao grupo nominal em português brasileiro, a correspondência entre grupo nominal e termo técnico pode ser entendida por meio da perspectiva trinocular (de baixo, de cima e ao redor) inerente à teoria sistêmico-funcional. A análise do termo técnico pela perspectiva de baixo revela que termos técnicos são constituídos por palavras da classe nominal e são frequentemente realizados por grupos nominais, que operam na ordem da oração como Participantes. A análise pela perspectiva de cima indica que os termos técnicos funcionam semanticamente como participantes. E a análise pela perspectiva ao redor aponta que os termos técnicos podem ser co-extensivos (coextensiveness) ao grupo nominal, ou seja, os termos técnicos podem compartilhar de funções inerentes ao grupo nominal, como, Ente e Qualidade, por exemplo (cf. HALLIDAY, 1961).

3 Metodologia

A metodologia deste artigo é dividida em duas partes. A primeira parte aborda a compilação do corpus utilizado, bem como descreve o corpus que serviu como corpus de referência para a presente pesquisa. A segunda explica como se deu a adaptação do trabalho de Pearson (1998) para o português brasileiro, uma vez que o presente estudo foi baseado na obra Terms in Context de Jennifer Pearson.

3.1 O *corpus*

Os textos selecionados para constituírem o corpus utilizado nesta pesquisa compreendem artigos acadêmicos que retratam pesquisas desenvolvidas no âmbito do Diabetes Mellitus tipo II. O domínio experiencial (HALLIDAY; MATTHIESSEN, 1999; HAO, 2015) desses

textos é, portanto, essa condição crônica, especificamente, o autocuidado em Diabetes Mellitus.[2]

Esses artigos acadêmicos foram extraídos de publicações da área das Ciências da Saúde a partir da palavra-chave "autocuidado em diabetes mellitus" digitada na aba de busca do Google Acadêmico.[3] Foram coletados 40 artigos originalmente escritos em português brasileiro, publicados entre 2010 e 2019, gerando, ao todo, 133.232 *tokens*, sendo que a seleção por artigos desse período de tempo é justificada pela necessidade de textos atuais, capazes de retratar padrões produzidos na última década.

Cada artigo foi salvo separadamente em um arquivo no bloco de notas e nominado com o Título da Revista_ano de publicação, sendo que, para dois ou mais artigos de um determinado periódico publicados no mesmo ano, a nomeação se deu da seguinte forma Título da Revista_ano de publicaçãob.

É importante destacar que os *abstracts*, quadros, gráficos, anexos, figuras, tabelas e referências bibliográficas de todos os artigos selecionados para o corpus foram excluídos quando cada artigo foi transferido para o respectivo arquivo do bloco de notas, dado que a utilização desse material poderia gerar um corpus poluído, sem contribuições significativas para os resultados do presente estudo.

3.2 O corpus de referência

Para que a lista de palavras-chave pudesse ser gerada, recorreu-se à parte escrita monológica do corpus CALIBRA (Catálogo da Língua Brasileira) como corpus de referência. A seleção por essa parte do CALIBRA é justificada pelo fato do corpus utilizado para a análise da

---

[2] É importante mencionar que este artigo faz parte de uma pesquisa de doutorado em desenvolvimento pela primeira autora deste artigo, a partir do apoio financeiro da Fundação de Amparo à Pesquisa do Estado de Minas Gerais – FAPEMIG. Tal pesquisa está localizada na área de Estudos Linguísticos no escopo do Programa de Pós-graduação em Estudos Linguísticos da Faculdade de Letras da UFMG e também no âmbito do Projeto Empoder@ – Protótipo conceitual e metodológico para avaliação de intervenções orientadas ao autocuidado em diabetes, uma parceria entre o Laboratório Experimental de Tradução (LETRA) da FALE/UFMG, a Escola de Enfermagem da UFMG e o Departamento de Estatística do ICEx/UFMG.

[3] Disponível em: https://scholar.google.com.br/?hl=pt.

presente pesquisa ser formado por artigos acadêmicos, um tipo de texto escrito e monológico, tal como os textos do corpus de referência. Essa similaridade entre os textos do corpus de análise e os textos do corpus de referência, mesmo que mínima, pode garantir a coerência dos resultados obtidos para o presente artigo.

O CALIBRA compreende um corpus com cerca de um milhão de palavras (*tokens*), compilado de acordo com a tipologia do contexto de cultura (HALLIDAY, 1978). Essa tipologia é determinada por meio de cinco variáveis: i. Especialização (especializado/não especializado); ii. Papel da língua na situação (constitutivo/auxiliar); iii. Modo de produção (escrito/falado); iv. Modo de interação (monólogo/falado); v. Processo sociossemiótico (explorar/compartilhar/explicar/relatar/recriar/fazer/recomendar/habilitar). O Quadro 2 resume todos os tipos de texto presentes no CALIBRA e destaca aqueles que foram utilizados como corpus de referência nesta pesquisa.

QUADRO 2 – CALIBRA – Distribuição dos tipos de texto segundo o contexto de cultura

	PRODUÇÃO		escrito		falado	
	INTERAÇÃO		diálogo	monólogo		diálogo
ESPECIALIZAÇÃO	PAPEL	PROCESSO				
especializada	constitutivo	EXPLICAR	"yahoo respostas"	livro texto	palestra	debate
não-especializada		RELATAR	questionário	reportagem	depoimento	entrevista
		RECRIAR	quadrinhos	conto	causo	teatro de improviso
		COMPARTILHAR	e-chat	blog (diário)	vlog (diário)	bate-papo
	auxiliar	FAZER	carta comercial	receita	instruções	co-operação
especializada	constitutivo	RECOMENDAR	auto-ajuda	anúncios	orações	consulta médica
		HABILITAR	perguntas mais frequentes	panfletos	orientações	perguntas e respostas
		EXPLORAR	carta ao editor	artigo acadêmico	discurso	discussão

Fonte: Adaptado de Figueredo, Pagano e Ferreguetti (2014).

O Quadro 2 mostra que a parte do corpus de referência utilizada para a geração da lista de palavras-chave conta com textos dos tipos: livro texto, reportagem, conto, blog (diário), receita, anúncios, panfletos e artigo acadêmico. Todos originalmente escritos em português e monológicos.

Em relação ao número de textos e de *tokens* da parte do CALIBRA utilizada como corpus de referência para este artigo, a Tabela 1 informa todas as quantidades.

TABELA 1 – O corpus de pesquisa: número de textos e de tokens

Processo sociossemiótico	**Número de textos**	**Número de *tokens***
Compartilhar	19	9.608
Explicar	24	28.841
Explorar	18	18.552
Fazer	56	32.646
Habilitar	53	58.875
Recomendar	25	36.912
Recriar	39	33.708
Relatar	30	33.142
TOTAL	264	252.284

Fonte: Elaborada para fins deste artigo.

3.3 A adaptação do trabalho de Pearson (1998) para o português brasileiro

Como detalhado anteriormente (cf. Referencial Teórico), o estudo publicado em Pearson (1998) contou com alguns recursos que ainda são pouco difundidos para o português brasileiro, como, por exemplo, o anotador morfossintático automático e o programa de padrão de correspondência. Em função disso, os parâmetros metodológicos estabelecidos para a presente pesquisa se baseiam em alguns dos passos metodológicos definidos em Pearson (1998), excluindo as fases dependentes dos recursos tecnológicos desenvolvidos para aquele trabalho. Portanto, não se pode afirmar que a metodologia da presente pesquisa replica, *ipsis litteris*, aquela descrita no estudo de Pearson (1998), uma vez que várias modificações metodológicas tiveram que ser implementadas.

Após a compilação do corpus de pesquisa, detalhado na primeira subseção desta metodologia, todos os textos desse corpus foram inseridos no *software* concordanciador AntConc. Em seguida houve a exclusão de todos os itens gramaticais, bem como os metadados que serviam para documentação do corpus. Isso se deu pelo ajuste das configurações de "Tool Preferences > WordList" do AntConc, a partir da inserção de uma lista de itens gramaticais importada para o concordanciador, permitindo que uma lista de palavras-chave fosse gerada sem o ruído causado pela presença desses itens gramaticais.

Com a lista de palavras-chave gerada, as três palavras dessa lista que tiveram maior número de ocorrências foram selecionadas para terem seus cotextos investigados. Como o uso do programa de padrão de correspondência em português brasileiro pode ser considerado restrito a algumas áreas do conhecimento, a ferramenta de geração de clusters/N-grams do AntConc foi determinante para que se pudesse verificar a existência de grupos nominais formados à esquerda e à direita da palavra-chave em questão. A busca pelos clusters/N-grams se deu sempre primeiro à esquerda e, depois, à direita da palavra-chave, sendo a extensão de cinco palavras para cada direção, contando com a palavra-chave em questão. Todos os clusters/N-grams formados à esquerda e à direita da palavra-chave que tivessem o mínimo de dez ocorrências deveriam ter seus respectivos cotextos examinados, por meio da ferramenta de concordância do AntConc, que permite verificar quais palavras estão em coocorrência com determinado cluster/N-gram e/ou palavra de busca. Quando a busca pela extensão de cinco palavras à esquerda e/ou à direita da palavra-chave não resultasse em clusters/N-grams com o mínimo de dez ocorrências, a extensão deveria ser reduzida de uma em uma palavra, isto é, quatro palavras, três palavras, até que clusters/N-grams com o número mínimo de dez ocorrências fossem obtidos.

Após a extração de cada cluster/N-gram à esquerda e à direita da palavra-chave que tivesse o número mínimo de dez ocorrências, deu-se início ao processo de análise desses clusters/N-grams. Essa análise compreende a busca por possíveis grupos nominais presentes nesses clusters/N-grams, bem como a anotação das funções exercidas por cada palavra que compõe o grupo nominal em análise.

É importante mencionar que esse processo de busca e análise de grupos nominais presentes em clusters/N-grams pode ser visto como uma forma semiautomática de pesquisa por termos técnicos, visto que em português brasileiro termos técnicos são comumente coextensivos ao grupo

nominal (cf. FIGUEREDO, 2007). Além disso, os artigos acadêmicos, que constituem o corpus de pesquisa utilizado no presente estudo, compreendem um tipo de texto em que as chances de aparecer termos técnicos são altas dado o grau de especialidade da linguagem utilizada nesse tipo de texto (HALLIDAY, 1978). Esse uso especializado da linguagem pode favorecer também o aparecimento de termos técnicos realizados por mais de uma palavra, isto é, grupos nominais que contam com mais de uma função, por exemplo, Ente e Qualificador (cf. FIGUEREDO, 2007).

O último passo metodológico do presente estudo foi verificar dentre os grupos nominais investigados anteriormente quais poderiam ser caracterizados como termos técnicos. Essa verificação foi feita pela análise da coocorrência desses grupos nominais com algum (alguns) dos sinais linguísticos apontados no trabalho de Pearson (1998) (cf. Referencial Teórico), bem como pela análise de alguma particularidade do corpus de pesquisa utilizado e/ou algum sinal linguístico que não foi mencionado no estudo de Pearson (1998), mas que se mostrou relevante para a presente pesquisa.

4 Resultados

Como detalhado nas seções de metodologia e de referencial teórico, o presente trabalho se baseia na pesquisa apresentada em Pearson (1998), a qual propõe uma metodologia para auxiliar terminologistas, lexicógrafos e aqueles que se dedicam à linguística de corpus a identificar termos técnicos por meio de ferramentas da linguística de corpus utilizando três corpora diferentes (cf. seção 3 Metodologia).

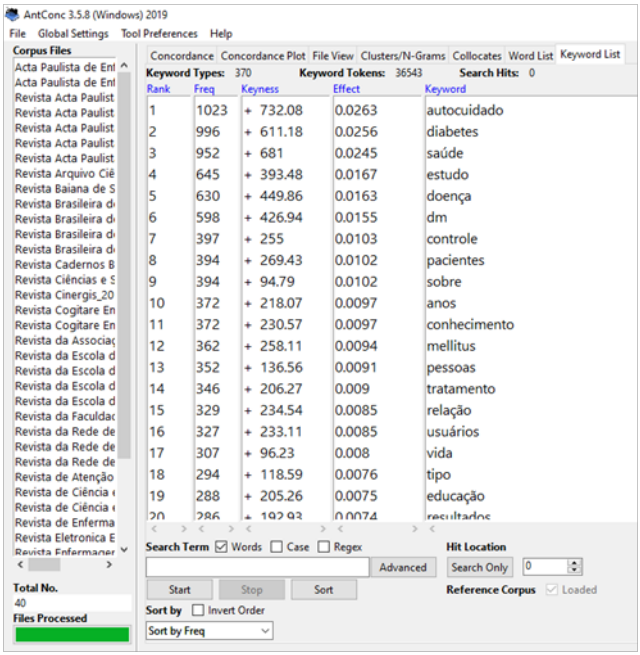
Na obra, a autora relata que a pesquisa se inicia quando os textos dos corpora são importados para um anotador morfossintático (*P.O.S. tagger*), desenvolvido especialmente para aquele trabalho, a fim de isolar os itens da classe de palavras nominal[4] e, posteriormente, verificar se eles formavam padrões, isto é, se aqueles itens nominais extraídos do anotador morfossintático eram frequentemente encontrados juntos nos corpora. A hipótese apresentada nesse estudo sugere que esses padrões

[4] Para a teoria Sistêmico-funcional (HALLIDAY; MATTHIESSEN, 2014, p. 75), há três classes de palavras – nominal (nominal), verbal (verbal) e adverbial (adverbial) – a classe nominal, sobre a qual o presente artigo enfoca, engloba os adjetivos (adjective); numerais (numeral), determinantes (determiner) e os substantivos (noun) que, por sua vez, podem ser comum (common), próprio (proper) ou pronomes (pronoun).

de coocorrência entre itens da classe de palavras nominal poderiam identificar termos técnicos comuns para as áreas de conhecimento dos corpora utilizados naquela pesquisa.

Entretanto, como ainda não há um *software* livre que faça esse tipo de processo de anotação morfossintática para o português brasileiro, sem que haja o desenvolvimento de um *script* em linguagem de programação e a utilização de um corpus de treinamento que auxilie na exatidão dos resultados, o presente estudo optou por fazer uma busca semiautomática dos itens da classe de palavras nominal no corpus utilizado neste trabalho (cf. Metodologia). Para isso, o corpus de artigos acadêmicos sobre Diabetes Mellitus foi importado para o *software* concordanciador AntConc. Após a extração dos itens gramaticais do corpus (cf. Metodologia), uma lista de palavras-chave foi gerada a partir do *upload* da parte escrita monológica do CALIBRA, a qual foi utilizada nesta pesquisa como corpus de referência. A seguir, a parte inicial da lista de palavras-chave é apresentada.

FIGURA 2 – Captura de tela com o início da lista de palavras-chave proveniente do corpus de artigos acadêmicos



Fonte: Elaborada para fins deste artigo.

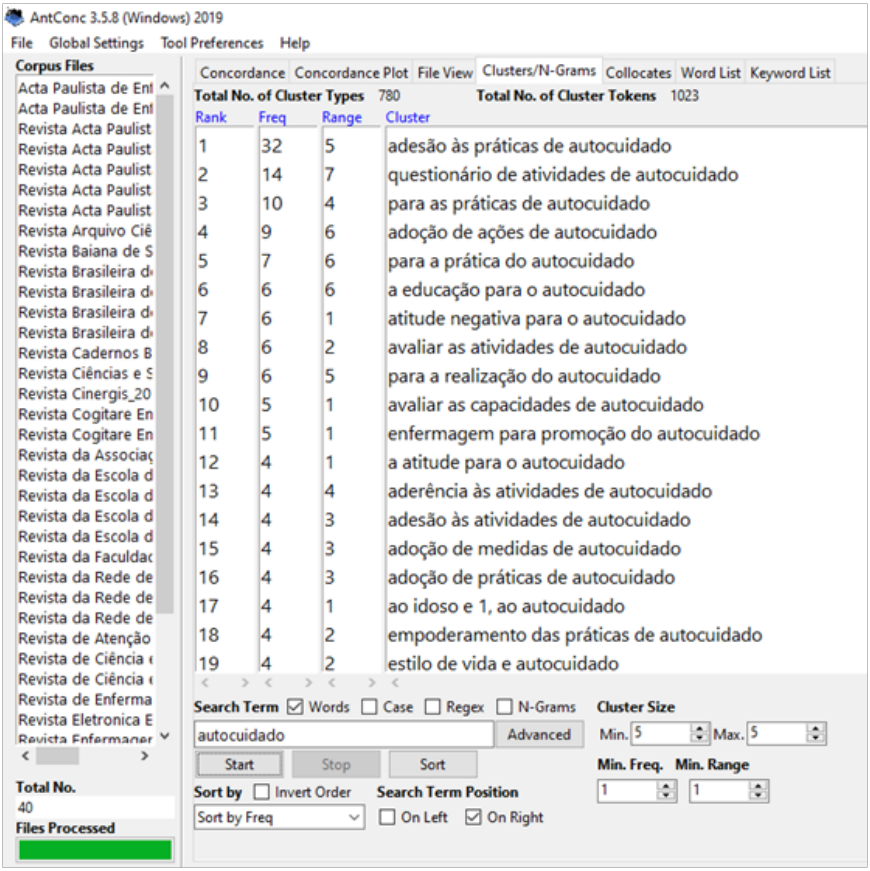
A Figura 2 mostra uma lista com as palavras-chave mais frequentes a partir da comparação que o *software* concordanciador faz com o corpus de referência utilizado. Os índices da coluna de frequência (Freq) detalham em ordem decrescente o número de vezes que determinada palavra-chave apareceu no corpus, comprovando que as palavras autocuidado (1023 ocorrências), diabetes (996 ocorrências) e saúde (952 ocorrências) são as três palavras-chave mais frequentes e por isso foram selecionadas para serem investigadas neste artigo.

Seguindo os passos metodológicos descritos em Pearson (1998), o próximo passo compreende a análise das coocorrências, ou seja, quais são as palavras da classe nominal que frequentemente coocorrem com outras palavras dessa mesma classe, sugerindo a possibilidade de termos técnicos daquela área de conhecimento. Como no presente artigo não foi utilizado um anotador morfossintático para selecionar somente as palavras da classe nominal, a ferramenta de clusters/N-grams do *software* concordanciador AntConc serviu para examinar qual(is) palavra(s) da classe nominal coocorre(iam) com cada uma das três palavras-chave selecionadas anteriormente.

Antes de apresentar os clusters/N-grams mais frequentes para cada uma das palavras-chave, é importante relembrar que a busca por clusters/N-grams objetiva encontrar termos técnicos da área de conhecimento dos textos do corpus de maneira semiautomática, adaptando a metodologia proposta por Pearson (1998) para o contexto do português brasileiro. Nesse sentido, grupos nominais devem estar instanciados nos clusters/N-grams gerados para as palavras-chave – autocuidado, diabetes e saúde (cf. Referencial Teórico).

Finalizado os esclarecimentos acerca da correspondência entre termo técnico, grupo nominal e clusters/N-grams, inicia-se o detalhamento dos clusters/N-grams formados à esquerda da palavra-chave com maior número de ocorrências – "autocuidado".

FIGURA 3 – Captura de tela dos clusters/N-grams formados à esquerda da palavra-chave "autocuidado"

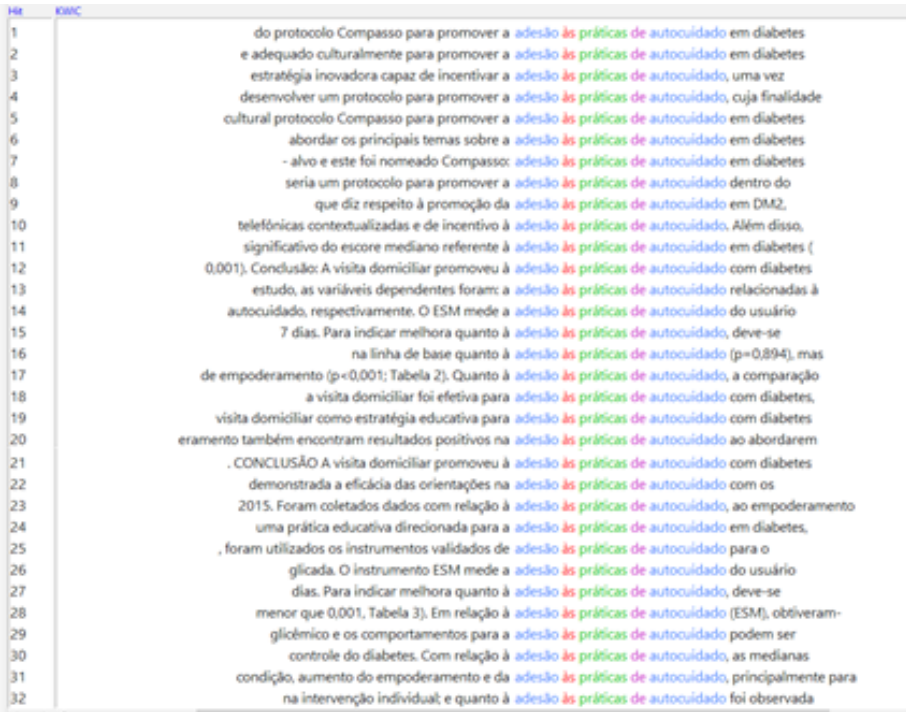


Fonte: Elaborada para fins deste artigo.

Na Figura 3, tem-se os clusters/N-grams formados com a palavra-chave autocuidado, sendo que esses clusters/N-grams abrangem cinco palavras à esquerda do item de busca, uma vez que o tamanho do cluster, marcado na parte inferior à direita da Figura 3 em *Cluster Size*, é cinco, tanto para o tamanho mínimo quanto para o tamanho máximo. Seguindo a metodologia do presente trabalho, os clusters/N-grams com o mínimo de dez ocorrências são selecionados para serem analisados de maneira mais profunda (cf. Metodologia). Portanto o cotexto em que os clusters/N-grams "adesão às práticas de autocuidado", "questionário de

atividades de autocuidado" e "para as práticas de autocuidado" aparecem são explorados a seguir.

FIGURA 4 – Captura de tela do primeiro cluster/N-gram formado à esquerda da palavra-chave "autocuidado": "adesão às práticas de autocuidado"



Fonte: Elaborada para fins deste artigo.

O primeiro cluster/N-gram formado com cinco palavras à esquerda de "autocuidado" é "adesão às práticas de autocuidado", como mostra a Figura 4. Nesse cluster/N-gram, há dois grupos nominais, a saber:

QUADRO 3 – Os dois grupos nominais presentes em "adesão às práticas de autocuidado" e as respectivas funções exercidas na ordem do grupo nominal

1º grupo nominal	adesão	às práticas de autocuidado	
	Ente	**Qualificador**	
2º grupo nominal	às	práticas	de autocuidado
	Dêitico	**Ente**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

Contudo, o primeiro grupo nominal surge de uma metaforização[5] em que "aderir" sofreu um processo de metaforização gramatical para "adesão", havendo a nominalização de "aderir" para "adesão", comum em textos em que o uso da linguagem especializada é predominante. Por essa perspectiva prevista pela teoria sistêmico-funcional, esse primeiro grupo nominal não existiria, já que esse cluster/N-gram seria uma oração: "aderir às práticas de autocuidado". Em razão disso, "as práticas de autocuidado" passa a ser entendido como o único grupo nominal desse cluster/N-gram.

O segundo cluster/N-gram formado a partir da palavra-chave "autocuidado" é "questionário de atividades de autocuidado". A Figura 5 traz as linhas de concordância para esse cluster/N-gram.

[5] Segundo os pressupostos da teoria sistêmico-funcional, a metáfora gramatical é frequentemente encontrada em textos científicos, como os artigos acadêmicos, uma vez que é nesse tipo de texto em que a linguagem especializada pode ser identificada. Uma das formas possíveis de se analisar e compreender a metáfora gramatical é por meio das mudanças de ordem, isto é, uma oração é metaforizada e torna-se um grupo nominal (cf. HALLIDAY; MATTHIESSEN, 2014).

FIGURA 5 – Captura de tela do segundo cluster/N-gram formado à esquerda da palavra-chave "autocuidado": "questionário de atividades de autocuidado"

Fonte: Elaborada para fins deste artigo.

Assim como o primeiro cluster/N-gram formado à esquerda da palavra-chave "autocuidado", este segundo cluster/N-gram também apresenta dois grupos nominais.

QUADRO 4 – Os grupos nominais presentes em "questionário de atividades de autocuidado" e as respectivas funções exercidas na ordem do grupo nominal

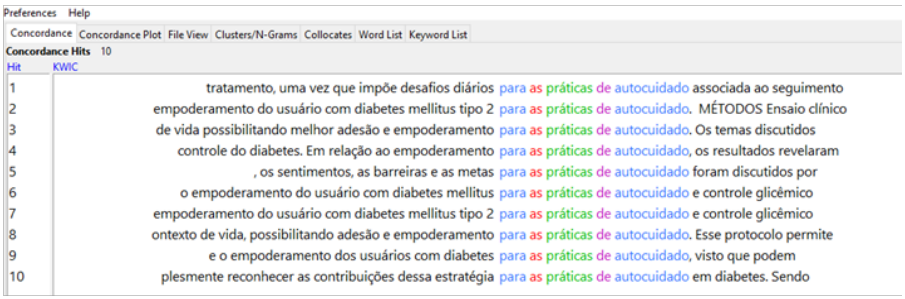
1º grupo nominal	Questionário	de atividades de autocuidado
	Ente	**Qualificador**
2º grupo nominal	Atividades	de autocuidado
	Ente	**Qualificador**

Fonte: Elaborado para fins deste artigo.

Embora haja dois grupos nominais no cluster/N-gram "questionário de atividades de autocuidado", pela análise do cotexto das linhas de concordância da Figura 5, é possível afirmar que o primeiro grupo nominal, que engloba todo cluster/N-gram, é o único que pode ser considerado, uma vez que ele corresponde ao nome do questionário utilizado nas pesquisas reportadas nos artigos selecionados para o corpus. Em outras palavras, o segundo grupo nominal "atividades de autocuidado" apresenta um significado diferente daquele sinalizado pela análise das linhas de concordância do cluster/N-gram em questão, cujo objetivo deve ser revelar o nome do instrumento usado.

O terceiro e último cluster/N-gram formado à esquerda da palavra-chave "autocuidado" é "para as práticas de autocuidado". A Figura 6, a seguir, mostra as linhas de concordância desse cluster/N-gram.

FIGURA 6 – Captura de tela do terceiro cluster/N-gram formado à esquerda da palavra-chave "autocuidado": "para as práticas de autocuidado"



Fonte: Elaborada para fins deste artigo.

Esse cluster/N-gram é formado pela preposição "para", a qual não exerce função no grupo nominal presente no cluster/N-gram, por isso tal palavra não aparece no Quadro 5, que aborda as funções presentes no grupo nominal em questão.

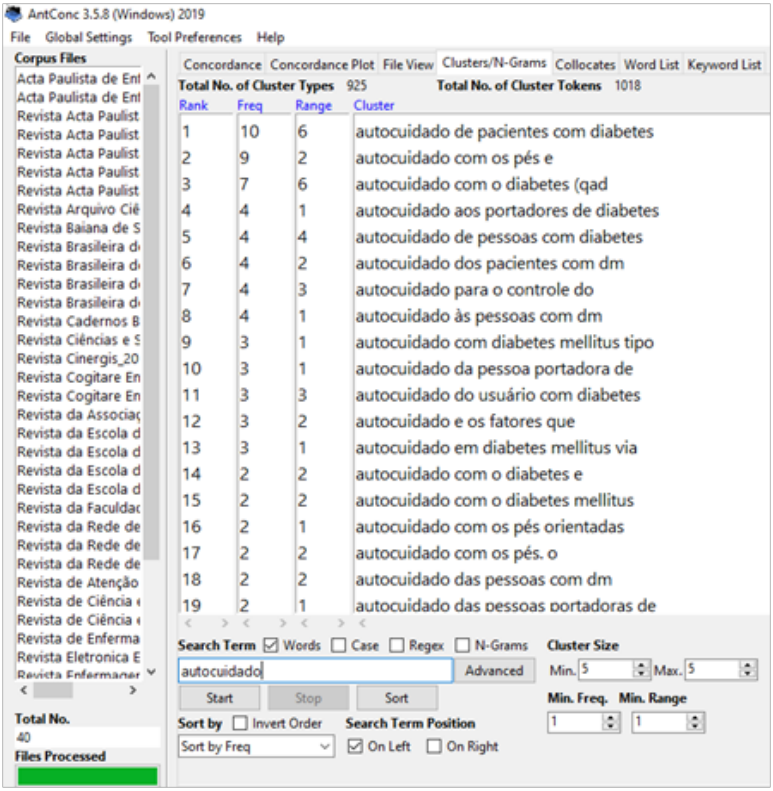
QUADRO 5 – O grupo nominal presente em "para as práticas de autocuidado" e as respectivas funções exercidas na ordem do grupo nominal

Grupo nominal	as	práticas	de autocuidado
	Dêitico	**Ente**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

Encerrado o detalhamento dos cluster/N-grams formados à esquerda da palavra-chave "autocuidado" que tiveram o número mínimo de dez ocorrências no corpus, torna-se necessário apresentar os clusters/N-grams formados à direita dessa palavra-chave. Nesta segunda configuração, os clusters/N-grams também seguem o padrão de busca de cinco palavras à direita de "autocuidado" como mostra a Figura 7 a seguir.

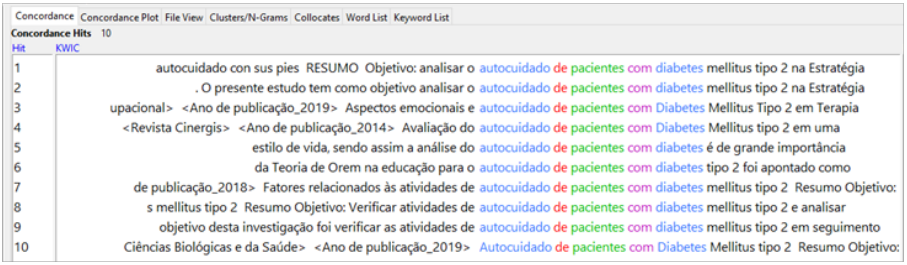
FIGURA 7 – Captura de tela dos clusters/N-grams formados à direita da palavra-chave "autocuidado"



Fonte: Elaborada para fins deste artigo.

A lista de clusters/N-grams formados à direita da palavra-chave "autocuidado", apresentada na Figura 7, revela que somente um cluster/N-gram teve o mínimo de dez ocorrências – "autocuidado de pacientes com diabetes". Adiante, as linhas de concordância, que mostram os cotextos em que esse cluster/N-gram aparece, estão retratadas na Figura 8.

FIGURA 8 – Captura de tela do primeiro cluster/N-gram formado à direita da palavra-chave "autocuidado": "autocuidado de pacientes com diabetes"



Fonte: Elaborada para fins deste artigo.

Esse cluster/N-gram possui dois grupos nominais, como mostra as classificações destacadas no Quadro 6 a seguir.

QUADRO 6 – Os grupos nominais presentes em "autocuidado de pacientes com diabetes" e as respectivas funções exercidas na ordem do grupo nominal

1º grupo nominal	autocuidado	de pacientes com diabetes
	Ente	**Qualificador**
2º grupo nominal	pacientes	com diabetes
	Ente	**Qualificador**

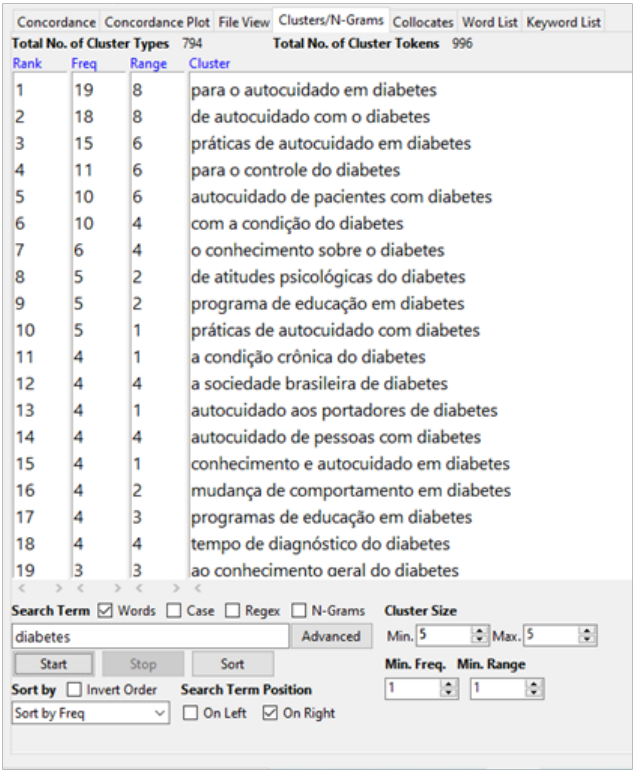
Fonte: Elaborado para fins deste artigo.

Os dois grupos nominais encontrados no cluster "autocuidado de pacientes com diabetes" apresentam significados diferentes. No primeiro grupo nominal o que está qualificado é o Ente "autocuidado" pela frase preposicional "de pacientes com diabetes". No segundo grupo nominal o que está qualificado é o Ente "pacientes" pela frase preposicional "com diabetes". Como a palavra-chave que norteou a busca pelos clusters/N-grams foi "autocuidado" não seria coerente considerar o segundo grupo nominal de forma desmembrada do primeiro, em razão das diferentes instanciações para a função do Ente. Nesse caso, portanto, há a correspondência exata entre o cluster/N-gram e o grupo nominal.

A segunda palavra-chave mais frequente no corpus é "diabetes", como mostra a Figura 1. Assim, os resultados apresentados a seguir dizem respeito aos dados encontrados para tal palavra. A Figura 9 traz

os clusters/N-grams formados à esquerda de "diabetes", considerando o intervalo de cinco palavras.

FIGURA 9 – Captura de tela dos clusters/N-grams formados à esquerda da palavra-chave "diabetes"

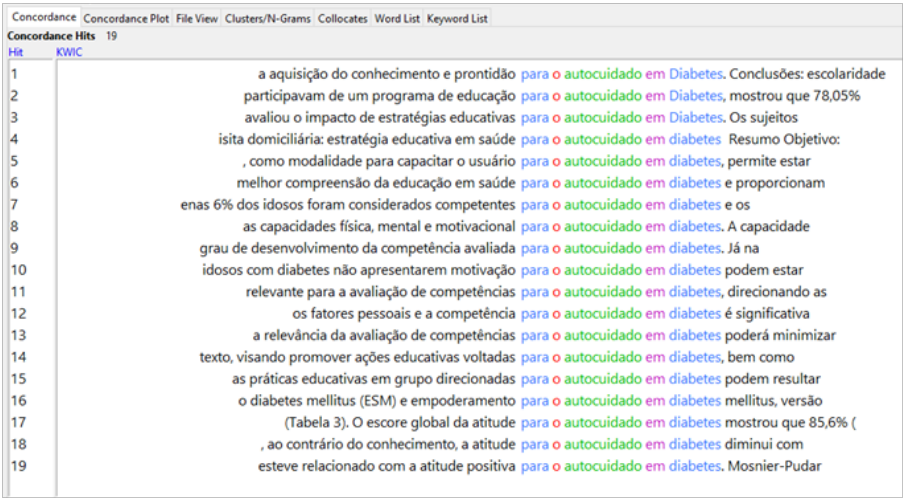


Fonte: Elaborada para fins deste artigo.

Seguindo a metodologia previamente desenvolvida para este artigo, os clusters/N-grams que têm no mínimo dez ocorrências são examinados, são eles: "para o autocuidado em diabetes", "de autocuidado com o diabetes", "práticas de autocuidado em diabetes", "para o controle do diabetes", "autocuidado de pacientes com diabetes" e "com a condição do diabetes".

As linhas de concordância do primeiro cluster/N-gram formado à esquerda de "diabetes" estão na Figura 10 adiante.

FIGURA 10 – Captura de tela do primeiro cluster/N-gram formado à esquerda da palavra-chave "diabetes": "para o autocuidado em diabetes"



Fonte: Elaborada para fins deste artigo.

A primeira palavra do cluster/N-gram em questão é a preposição "para", a qual não apresenta função no grupo nominal "o autocuidado em diabetes" que está no cluster/N-gram, destacado na Figura 10. O Quadro 7 destaca as funções deste grupo nominal.

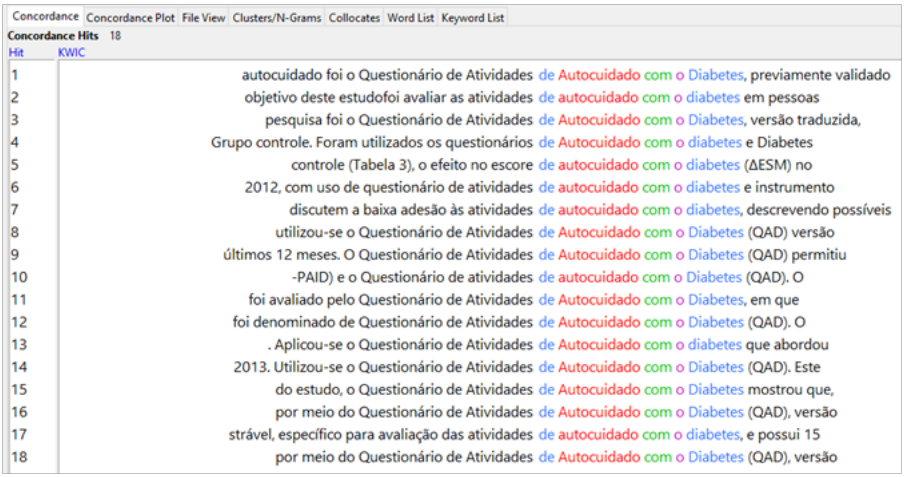
QUADRO 7 – O grupo nominal presente em "para o autocuidado em diabetes" e as respectivas funções exercidas na ordem do grupo nominal

Grupo nominal	o	autocuidado	em diabetes
	Dêitico	**Ente**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

O segundo cluster/N-gram formado à esquerda da palavra-chave "diabetes" compreende a expressão "de autocuidado com o diabetes". Para viabilizar o entendimento de tal cluster/N-gram, as linhas de concordância em que ele aparece no corpus estão destacadas na Figura 11.

FIGURA 11 – Captura de tela do segundo cluster/N-gram formado à esquerda da palavra-chave "diabetes": "de autocuidado com o diabetes"



Fonte: Elaborada para fins deste artigo.

O cotexto onde o cluster/N-gram em questão aparece evidencia que o grupo nominal localizado nesse cluster/N-gram faz parte de um grupo nominal maior, "Questionário de atividades de autocuidado com o diabetes", que, por sua vez, teve a outra porção analisada no segundo cluster/N-gram mais frequente para a palavra-chave "autocuidado". Por esse motivo, a preposição "de" presente no cluster "de autocuidado com o diabetes" funciona como preposição na frase preposicional "de autocuidado", que qualifica "atividades". O Quadro 8 mostra, portanto, as funções presentes no grupo nominal "autocuidado com o diabetes".

QUADRO 8 – O grupo nominal presente em "de autocuidado com o diabetes" e as respectivas funções exercidas na ordem do grupo nominal

Grupo nominal	autocuidado	com o diabetes
	Ente	**Qualificador**

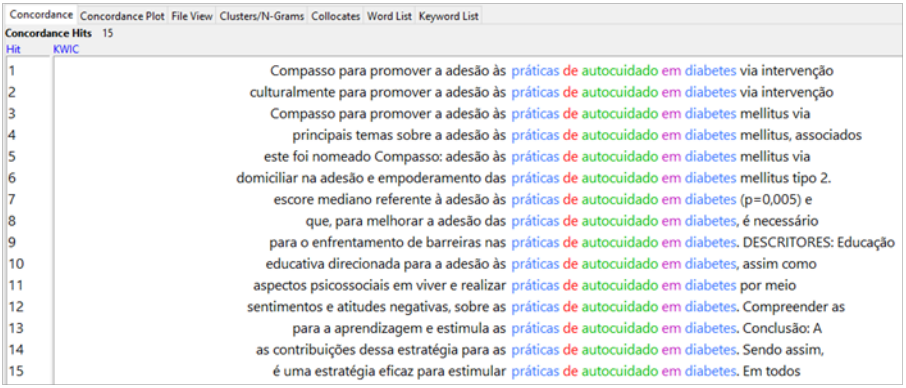
Fonte: Elaborado para fins deste artigo.

As funções, destacadas no Quadro 8, que constituem o grupo nominal presente no cluster/N-gram em questão, demonstram que a frase preposicional "com o diabetes" exerce a função de qualificar o ente

"autocuidado", ressaltando o tipo de autocuidado, no caso, aquele que se refere à condição crônica diabetes.

O terceiro cluster/N-gram mais frequente formado à esquerda da palavra-chave "diabetes" diz respeito à "práticas de autocuidado em diabetes". As linhas de concordância em que esse cluster/N-gram aparece estão detalhadas na Figura 12.

FIGURA 12 – Captura de tela do terceiro cluster/N-gram formado à esquerda da palavra-chave "diabetes": "práticas de autocuidado em diabetes"



Fonte: Elaborada para fins deste artigo.

Esse terceiro cluster/N-gram conta com dois grupos nominais cujas funções estão no Quadro 9 a seguir.

QUADRO 9 – Os grupos nominais presentes em "práticas de autocuidado em diabetes" e as respectivas funções exercidas na ordem do grupo

1º grupo nominal	práticas	de autocuidado em diabetes
	Ente	**Qualificador**
2º grupo nominal	autocuidado	em diabetes
	Ente	**Qualificador**

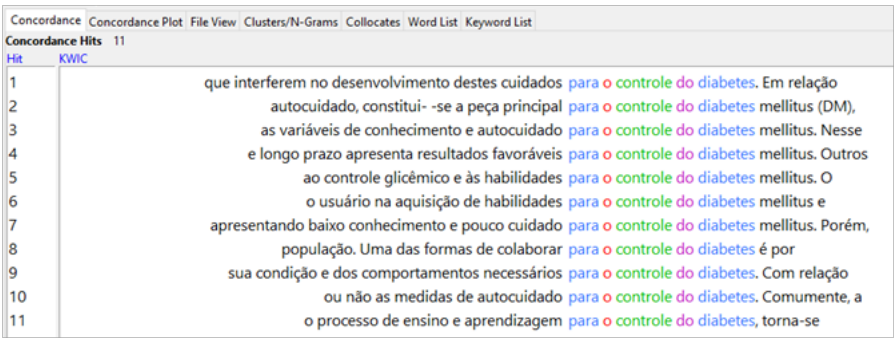
Fonte: Elaborado para fins deste artigo.

O primeiro grupo nominal presente no Quadro 9 engloba o cluster/N-gram em questão por completo. Neste grupo nominal, a frase preposicional "de autocuidado em diabetes" qualifica o Ente "práticas",

delimitando quais os tipos de prática são frequentemente abordados no corpus utilizado. O segundo grupo nominal "autocuidado em diabetes" tem como Ente "autocuidado" e como qualificador "em diabetes", este responsável por indicar qual o tipo de autocuidado os textos do corpus tratam. Fica evidente na análise das funções dos dois grupos nominais que seus significados são diferentes, enquanto no primeiro grupo nominal o que está qualificado são as "práticas", no segundo o que está qualificado é o "autocuidado". Entretanto, como um dos principais objetivos deste artigo é investigar a correspondência entre cluster/N-grams, grupo nominal e termo técnico, o primeiro grupo – práticas de autocuidado em diabetes – tende a se aproximar com maior exatidão dessa correspondência, por isso, a partir deste momento, apenas esse grupo é considerado como grupo nominal presente no terceiro cluster/N-gram formado à esquerda da palavra-chave "diabetes". Para além disso, o pressuposto de que os clusters/N-grams devem ter no mínimo dez ocorrências reforça também a escolha pelo primeiro grupo nominal, já que não se sabe qual a frequência do segundo grupo nominal no corpus.

O quarto cluster/N-gram formado à esquerda da palavra-chave "diabetes" trata-se da expressão "para o controle do diabetes". A Figura 13 mostra as linhas de concordância onde este cluster/N-gram pode ser encontrado no corpus.

FIGURA 13 – Captura de tela do quarto cluster/N-gram formado à esquerda da palavra-chave "diabetes": "para o controle do diabetes"



Fonte: Elaborada para fins deste artigo.

A análise dos cotextos em que este cluster/N-gram aparece possibilita afirmar que o "para" funciona como uma conjunção na ordem

da oração. Segundo o trabalho de Figueredo (2007), essa função não está diretamente relacionada ao grupo nominal em português brasileiro, ou seja, palavras da classe adverbial exercem prototipicamente funções no grupo adverbial, ao passo que palavras da classe nominal tem funções prototípicas no grupo nominal. Por essa razão, o grupo nominal presente neste cluster/N-gram envolve apenas a porção: "o controle do diabetes". O Quadro 10 salienta as funções desse grupo nominal.

QUADRO 10 – O grupo nominal presente em "para o controle do diabetes" e as respectivas funções exercidas na ordem do grupo nominal

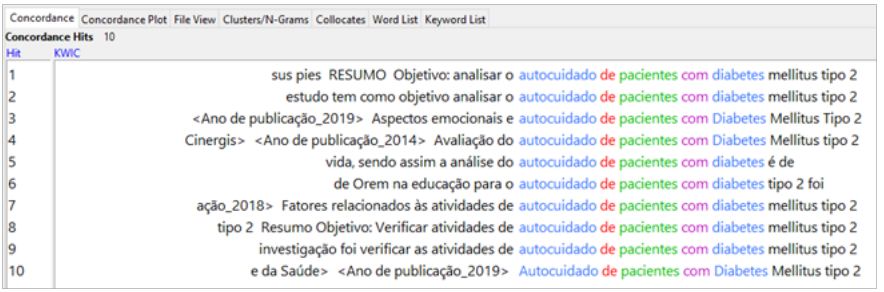
Grupo nominal	o	controle	do diabetes
	Dêitico	**Ente**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

Em conjunto com a preposição "de" a palavra-chave "diabetes" opera como um grupo nominal formado por uma frase preposicional, cuja função é qualificar o Ente "controle", caracterizando o tipo de controle, no caso, aquele relacionado ao Diabetes Mellitus.

O quinto cluster/N-gram formado à esquerda de "diabetes" é "autocuidado de pacientes com diabetes". A seguir, a Figura 14 revela as linhas de concordância com os cotextos em que este cluster/N-gram aparece.

FIGURA 14 – Captura de tela do quinto cluster/N-gram formado à esquerda da palavra-chave "diabetes": "autocuidado de pacientes com diabetes"



Fonte: Elaborada para fins deste artigo.

Tal como os exemplos anteriores, o cluster/N-gram em questão "autocuidado de pacientes com diabetes" conta com dois grupos nominais, como mostra o Quadro 11.

QUADRO 11 – Os grupos nominais presentes em "autocuidado de pacientes com diabetes" e as respectivas funções exercidas na ordem do grupo nominal

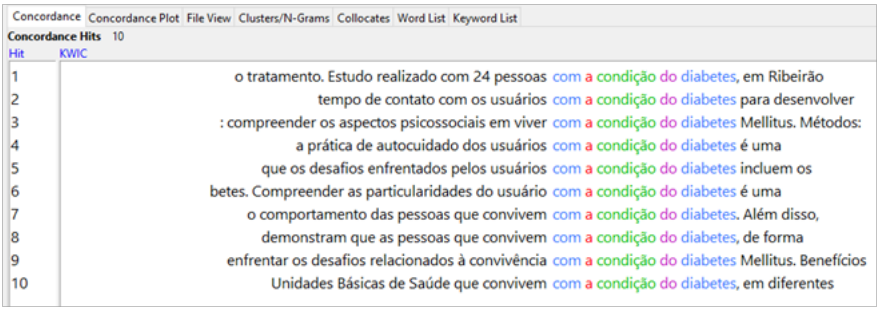
1º grupo nominal	autocuidado	de pacientes com diabetes
	Ente	**Qualificador**
2º grupo nominal	pacientes	com diabetes
	Ente	**Qualificador**

Fonte: Elaborado para fins deste artigo.

O primeiro grupo nominal "autocuidado de pacientes com diabetes" envolve todo o cluster/N-gram que está sendo examinado, enquanto que o segundo grupo nominal está em relação de dependência (hipotaxe) com o primeiro nominal. Isso salienta que os dois grupos nominais apresentam significados diferentes. No primeiro grupo nominal, o Ente é "autocuidado", e a qualificação se dá por "de pacientes com diabetes", que classifica o tipo de autocuidado abordado nos artigos acadêmicos do corpus. No segundo grupo nominal, o Ente é "pacientes", e a qualificação é "com diabetes", identificando que os tipos de pacientes mencionados no corpus são aqueles que tem diabetes. Tendo em mente o fato de que o primeiro grupo nominal "autocuidado de pacientes com diabetes" envolve todo o cluster/N-gram em questão, e, portanto, tem o mínimo de dez ocorrências requisitadas na metodologia deste artigo, bem como apresenta significado mais abrangente e coerente ao cotexto em que aparece nos textos do corpus, este grupo nominal passa a ser considerado o único válido para este cluster/N-gram.

O sexto e último cluster/N-gram formado à esquerda da palavra-chave diabetes, considerando a extensão de cinco palavras e com o mínimo de dez ocorrências, é "com a condição do diabetes". A Figura 15 traz as linhas de concordância que mostram os cotextos em que este cluster/N-gram aparece no corpus.

FIGURA 15 – Captura de tela do sexto cluster/N-gram formado à esquerda da palavra-chave "diabetes": "com a condição do diabetes"



Fonte: Elaborada para fins deste artigo.

Investigando as linhas de concordância da Figura 15, observa-se que o cluster/N-gram em questão começa com a preposição "com", a qual em conjunto com o restante do cluster/N-gram instanciado pelo grupo nominal "a condição do diabetes", tem a função de qualificar os Entes que estão fora do escopo do cluster/N-gram, como, por exemplo, "24 pessoas", "os usuários" e "as pessoas". Nesse sentido, se a preposição "com" fosse considerada como parte do grupo nominal presente no cluster/N-gram, a extensão deste cluster/N-gram precisaria ser modificada, o que causaria incoerência com a metodologia estabelecida para o presente estudo, bem como poderia causar alterações nos resultados encontrados até então. Por esses motivos, o grupo nominal presente no cluster/N-gram "com a condição do diabetes" diz respeito à "a condição do diabetes". As funções deste grupo nominal estão destacadas no Quadro 12 a seguir.

QUADRO 12 – O grupo nominal presente em "com a condição do diabetes" e as respectivas funções exercidas na ordem do grupo nominal

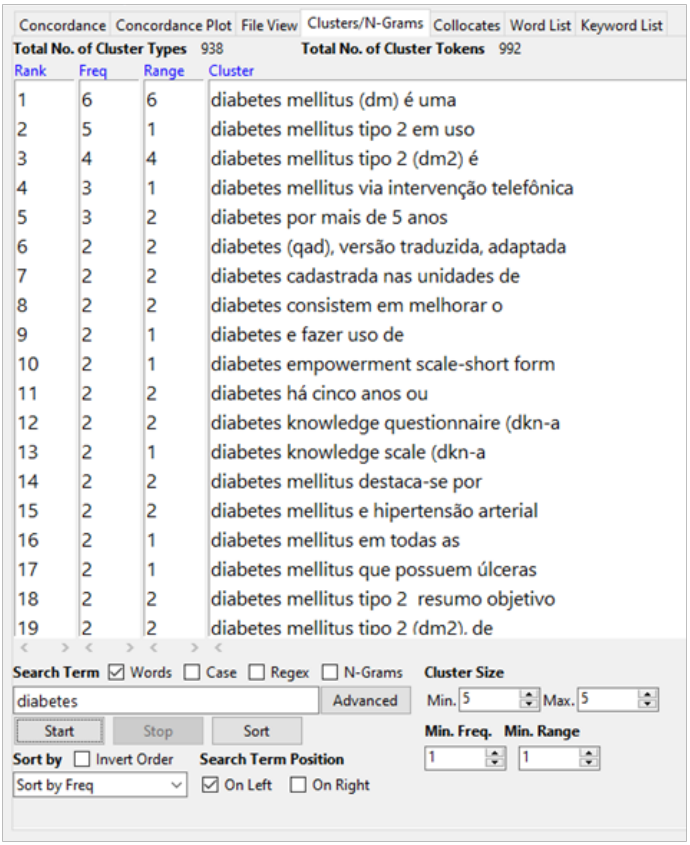
Grupo nominal	a	condição	do diabetes
	Dêitico	**Ente**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

Neste grupo nominal, o qualificador "do diabetes" tem a função de identificar o tipo de "condição" que foi mencionada com maior frequência nos textos do corpus utilizado na presente pesquisa. "Condição", por sua vez, funciona como o Ente do grupo nominal. Ainda neste grupo nominal, aparece o determinante "a" que opera como dêitico.

Encerrada a análise dos clusters/N-grams formados à esquerda da palavra-chave "diabetes" que obtiveram o número mínimo de dez ocorrências, a Figura 16 apresenta os clusters/N-grams formados à direita desta palavra-chave.

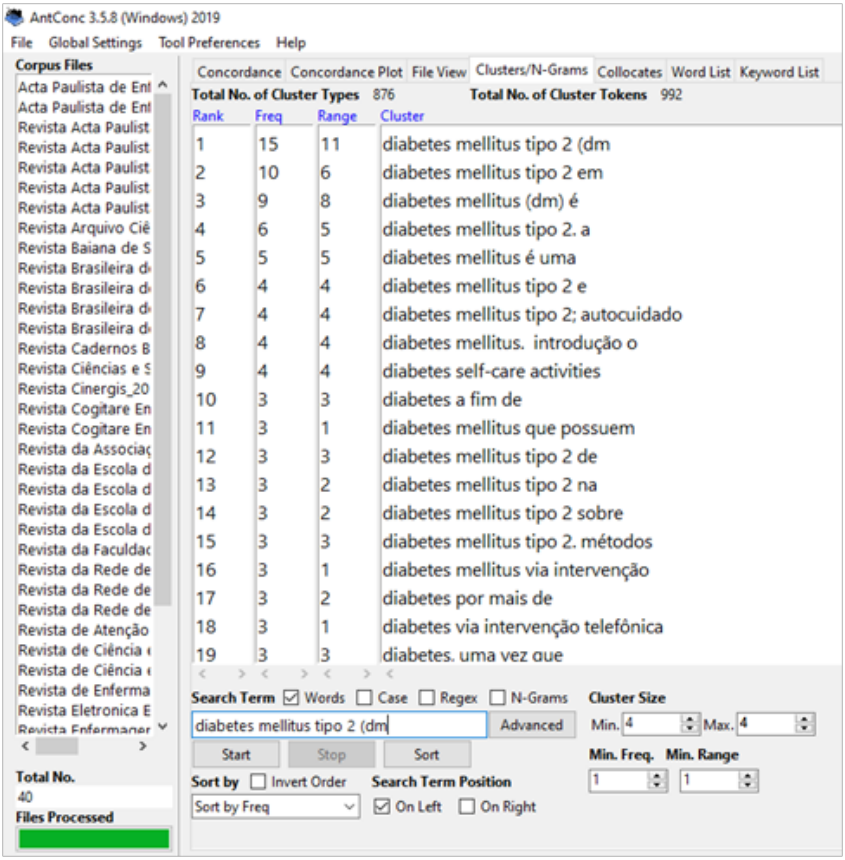
FIGURA 16 – Captura de tela dos clusters/N-grams formados à direita da palavra-chave "diabetes" com extensão de cinco palavras



Fonte: Elaborada para fins deste artigo.

Observando a coluna de frequências (Freq) da Figura 16, é possível afirmar que não há nenhum cluster/N-gram com o mínimo de dez ocorrências, condição estabelecida na metodologia do presente artigo. Por isso, o tamanho mínimo e máximo do cluster foi reduzido para quatro, como mostra a Figura 17 a seguir.

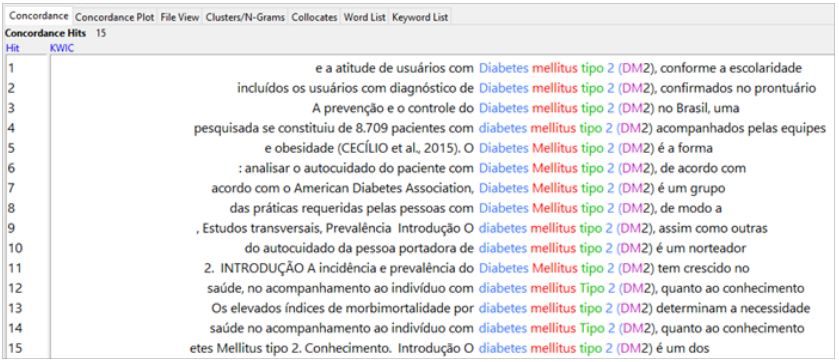
FIGURA 17 – Captura de tela dos clusters/N-grams formados à direita da palavra-chave "diabetes" com extensão de quatro palavras



Fonte: Elaborada para fins deste artigo.

Reduzindo a extensão do cluster/N-gram para quatro palavras à direita da palavra-chave "diabetes", dois clusters/N-grams aparecem com o mínimo de dez ocorrências. O primeiro diz respeito à "diabetes mellitus tipo 2 (dm". Adiante, as linhas de concordância, que demonstram onde este cluster/N-gram ocorre nos textos do corpus, estão representadas na Figura 18.

FIGURA 18 – Captura de tela do primeiro cluster/N-gram formado à direita da palavra-chave "diabetes": "diabetes mellitus tipo 2 (DM"



Fonte: Elaborada para fins deste artigo.

O cotexto deste cluster/N-gram envolve o nome da condição crônica que todos os textos do corpus utilizado abordam, seguido da sigla que identifica a condição aos pares, uma vez que o público-alvo principal dos artigos acadêmico são pesquisadores, assim como os autores desse tipo de texto que, na maior parte das vezes, também ocupam esse papel de pesquisador. O Quadro 13 expõe as funções deste grupo nominal.

QUADRO 13 – O grupo nominal presente em "diabetes mellitus tipo 2 (DM" e as respectivas funções exercidas na ordem do grupo nominal

Grupo nominal	Diabetes	Mellitus	Tipo 2
	Ente	**Classificador**	**Classificador**

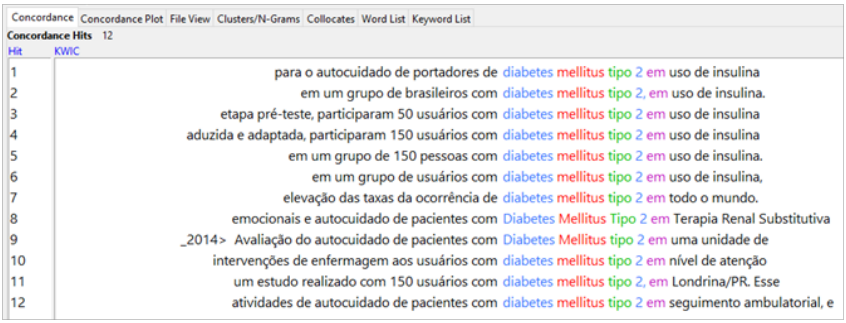
Fonte: Elaborado para fins deste artigo.

Seguindo a iniciativa salientada no trabalho de Bowker e Pearson (2002) sobre o fato de se considerar variações de um item como um único termo, em que Diabetes Mellitus tipo 2, Diabetes Mellitus e DM2 corresponderiam a um mesmo termo, ou seja, o significado de todas essas variações é o mesmo, todas dizem respeito à condição crônica Diabetes Mellitus, embora cada ocorrência possa ser contabilizada. Considerando esta perspectiva, o presente artigo computa as ocorrências de DM e Diabetes Mellitus tipo 2, presentes no cluster/N-gram em questão, de forma separada, mas considera que ambos os termos tem significados análogos. Nesse sentido, "Diabetes" funciona como Ente do grupo nominal.

"Mellitus" e "Tipo 2" como Classificadores de "Diabetes", inserindo essa condição na classe dos mellitus e do tipo 2, não do tipo 1, por exemplo.

O segundo cluster/N-gram formado à direita da palavra-chave "Diabetes" é "diabetes mellitus tipo 2 em". A Figura 19 exibe as linhas de concordância em que esse cluster/N-gram aparece.

FIGURA 19 – Captura de tela do segundo cluster/N-gram formado à direita da palavra-chave "diabetes": "diabetes mellitus tipo 2 em"



Fonte: Elaborada para fins deste artigo.

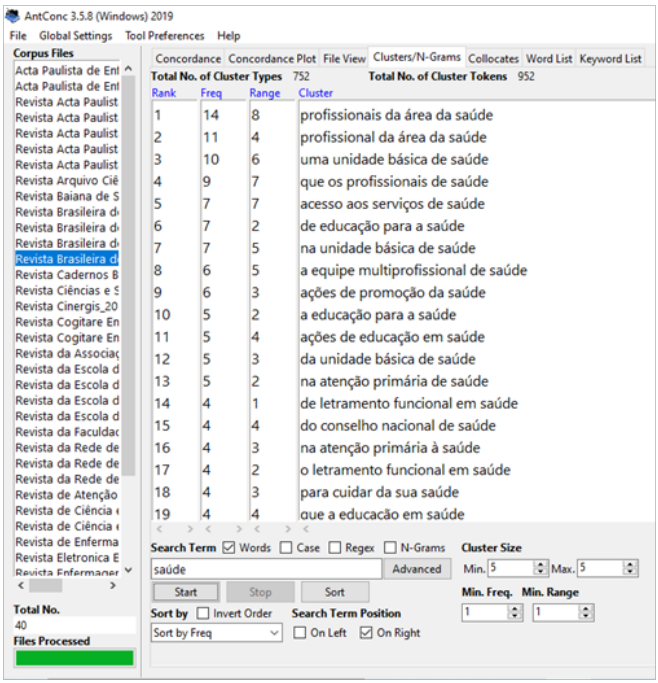
A análise das linhas de concordância aponta que o grupo nominal "diabetes mellitus tipo 2" presente no cluster/N-gram em questão faz parte de grupos nominais maiores, como: "50 usuários com diabetes mellitus tipo 2 em uso de insulina" e "avaliação do autocuidado de pacientes com diabetes mellitus tipo 2 em uma unidade". Em função disso, a preposição "em" que aparece ao final deste segundo cluster/N-gram opera como a preposição presente nas frases preposicionais (preposição + grupo nominal) que sucedem este cluster/N-gram. A depender da linha de concordância examinada, essas frases preposicionais podem indicar o lugar onde os pacientes/usuários com Diabetes Mellitus abordados nos artigos acadêmicos do corpus realizam seus tratamentos ou podem informar sobre a maneira pela qual esses pacientes/usuários recebem tratamento, no caso, pelo uso de insulina.

Ainda com relação ao grupo nominal "diabetes mellitus tipo 2" presente no cluster/N-gram em questão, a análise das linhas de concordância da Figura 19 indica que esse grupo nominal opera junto com as preposições "com/de", constituindo uma frase preposicional, a qual qualifica o Ente que a antecede. Esse Ente é realizado por "usuários", "pessoas", "portadores", "pacientes", "grupo" ou "taxas", dependendo de qual linha de concordância está em pauta.

Tendo em mente essas particularidades do cluster/N-gram "diabetes mellitus tipo 2 em" sobre o fato do grupo nominal que aparece nesse cluster/N-gram apresentar função tanto com as palavras que o antecede quanto com aquelas que o sucede inviabiliza a análise desse grupo nominal de forma isolada. Em outras palavras, para compreender o grupo nominal "diabetes mellitus tipo 2" nos casos destacados pelas linhas de concordância da Figura 19, é primordial entender as funções desse grupo em relação ao que vem antes e depois dele. No entanto, como a metodologia do presente artigo estabelece que a extensão do cluster/N-gram não pode ultrapassar as cinco palavras, os dados obtidos para este grupo nominal, e, consequentemente, para o cluster/N-gram em que ele está presente não estão incluídos nos resultados deste trabalho.

A terceira e última palavra-chave selecionada para o presente estudo diz respeito ao item "saúde". A Figura 20 destaca os clusters/N-grams formados à esquerda desta palavra.

FIGURA 20 – Captura de tela dos clusters/N-grams formados à esquerda da palavra-chave "saúde"

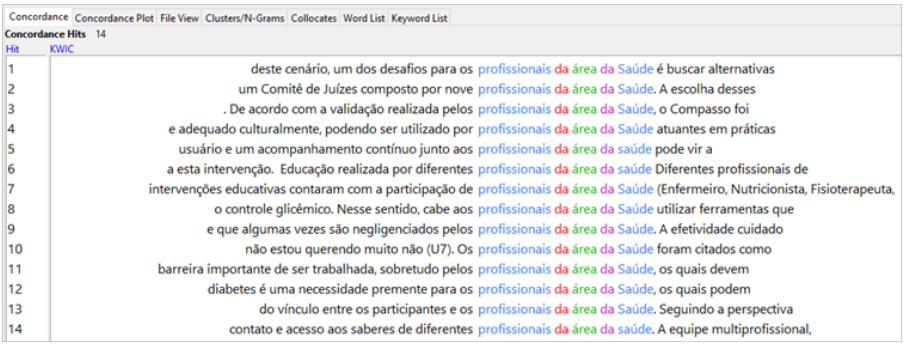


Fonte: Elaborada para fins deste artigo.

Levando em consideração os passos metodológicos definidos para a presente pesquisa, os cluster/N-grams que contam com o mínimo de dez ocorrências tem seus resultados discutidos. Para a palavra-chave "saúde", os clusters/N-grams que estão em concordância com esse padrão de ocorrências são: "profissionais da área da saúde", "profissional da área da saúde" e "uma unidade básica de saúde".

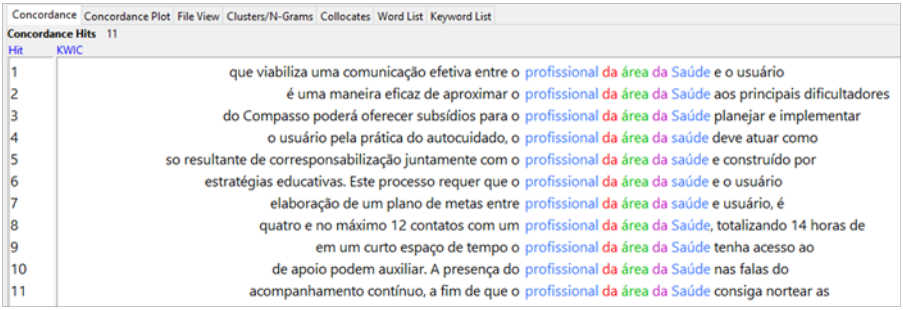
Considerando a possibilidade reportada no trabalho de Bowker e Pearson (2002) sobre as variações de um mesmo termo serem associadas a um significado em comum, o primeiro e segundo clusters/N-grams – "profissionais da área da saúde" e "profissional da área da saúde" – podem ser contabilizados de maneira separada, mas compreendidos sob um mesmo significado. A variação entre estes clusters/N-grams corresponde apenas à flexão de número entre "profissionais", que aparece no plural no primeiro cluster, e "profissional", que aparece no singular no segundo cluster. As Figuras 21 e 22 abordam as linhas de concordâncias, onde estão localizados estes clusters/N-grams.

FIGURA 21 – Captura de tela do primeiro cluster/N-gram formado à esquerda da palavra-chave "saúde": "profissionais da área da Saúde"



Fonte: Elaborada para fins deste artigo.

FIGURA 22 – Captura de tela do primeiro cluster/N-gram formado à esquerda da palavra-chave "saúde": "profissional da área da Saúde"



Fonte: Elaborada para fins deste artigo.

A análise das linhas de concordância das Figuras 21 e 22, respectivamente, indica que o pressuposto revelado por Bowker e Pearson (2002) sobre a variação e o significado de palavras análogas encontradas no mesmo corpus pode ser considerado válido para os clusters um e dois formados à esquerda da palavra-chave "saúde", uma vez que a variação de número em profissionais e profissional não gerou alterações expressivas de significado entre os dois clusters/N-grams. Por esse motivo, o Quadro 14 destaca o grupo nominal que aparece nestes clusters/N-grams, bem como as funções presentes neste grupo.

QUADRO 14 – Os grupos nominais presentes em "profissionais/profissional da área da saúde" e as respectivas funções exercidas na ordem do grupo nominal

1º grupo nominal	profissionais/profissional	da área da saúde
	Ente	**Qualificador**
2º grupo nominal	área	da saúde
	Ente	**Qualificador**

Fonte: Elaborado para fins deste artigo.

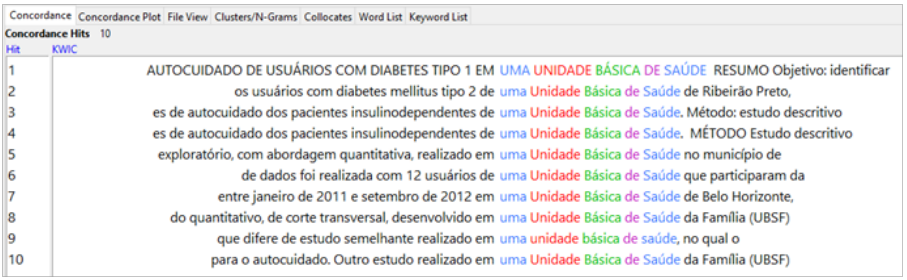
Antes de discutir as funções presentes neste grupo nominal, faz-se importante ressaltar que há um segundo grupo nominal dentro de "profissionais/profissional da área da saúde". Este grupo nominal corresponde a "área da saúde", em que "área" funciona como Ente e "da saúde" como Qualificador, este último responsável por caracterizar a área

mencionada nos textos do corpus. Entretanto, como esse grupo nominal "área da saúde" está em relação de hipotaxe (dependência) com o grupo nominal "profissionais/profissional da área da saúde" e não há como prever se a frequência mínima desse grupo nominal dependente é dez, como acontece com o grupo nominal "profissionais/profissional da área da saúde", esse segundo grupo nominal não será explorado neste artigo.

Considerando apenas o grupo nominal "profissionais/profissional da área da saúde", o Quadro 14 mostra que a frase preposicional "da área da saúde" funciona como Qualificador dos Entes profissionais e profissional, caracterizando-os dentro da área da saúde, não da educação ou advocacia, por exemplo.

O terceiro cluster/N-gram formado à esquerda da palavra-chave "saúde" refere-se a "uma unidade básica de saúde". Por meio das linhas de concordância, a Figura 23 demonstra o cotexto em que este cluster/N-gram aparece no corpus.

FIGURA 23 – Captura de tela do terceiro cluster/N-gram formado à esquerda da palavra-chave "saúde": "uma unidade básica de saúde"



Fonte: Elaborada para fins deste artigo.

A análise das linhas de concordância da Figura 23 revela que todo o cluster/N-gram "uma unidade básica de saúde" engloba um grupo nominal, cujas funções estão sinalizadas no Quadro 15.

QUADRO 15 – O grupo nominal presente em "uma unidade básica de saúde" e as respectivas funções exercidas na ordem do grupo nominal

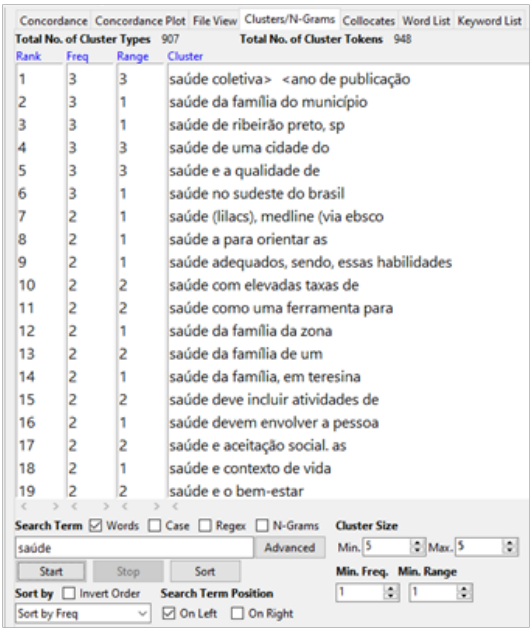
Grupo nominal	uma	unidade	básica	de saúde
	Dêitico	**Ente**	**Qualificador**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

O Quadro 15 identifica que o item “unidade” funciona como Ente do grupo nominal em questão. “Básica” e “de saúde” funcionam como Qualificadores e são responsáveis por inserir o Ente “unidade” em uma classe de unidades específica, no caso, aquelas que são básicas e destinadas a lidar com a saúde. Por fim, o determinante não-seletivo e não-específico “um” opera como Dêitico, demonstrando que nos artigos acadêmicos selecionados para o corpus da presente pesquisa as unidades básicas de saúde foram mencionadas de forma ampla, sem o enfoque no nome ou nos detalhes que poderiam identificar quais unidades básicas de saúde os estudos se referiam.

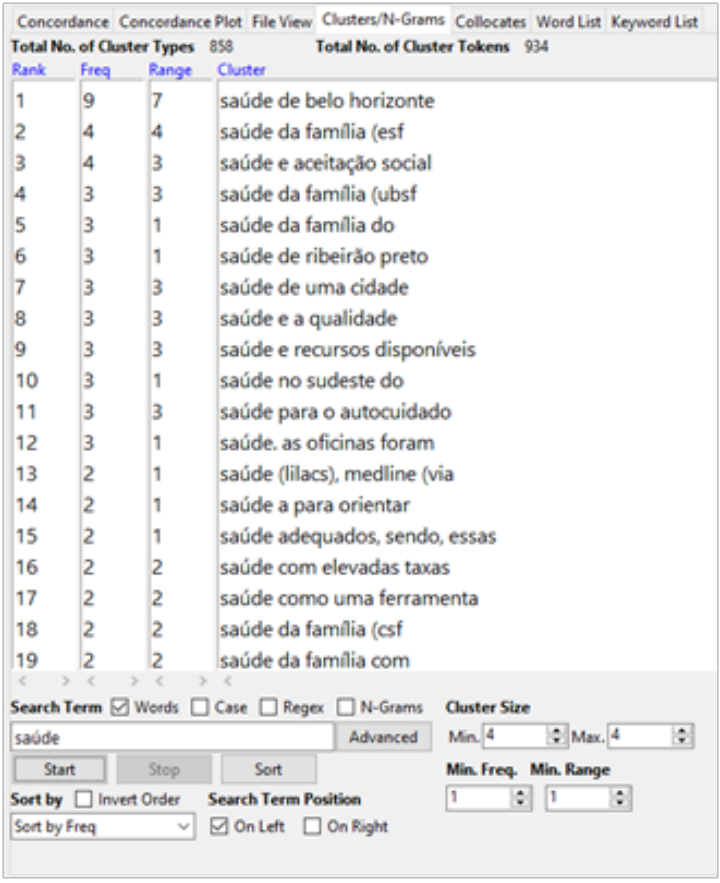
No que diz respeito às buscas pelos clusters/N-grams formados à direita da palavra-chave “saúde”, as Figuras 24 e 25 deixam evidente que as buscas com a extensão de cinco palavras e com a extensão de quatro palavras, respectivamente, não apresentaram os resultados esperados, uma vez que não houve clusters/N-grams com o mínimo de dez ocorrências, quantidade determinada na metodologia deste artigo.

FIGURA 24 – Captura de tela dos clusters/N-grams formados à direita da palavra-chave “saúde” com extensão de cinco palavras



Fonte: Elaborada para fins deste artigo.

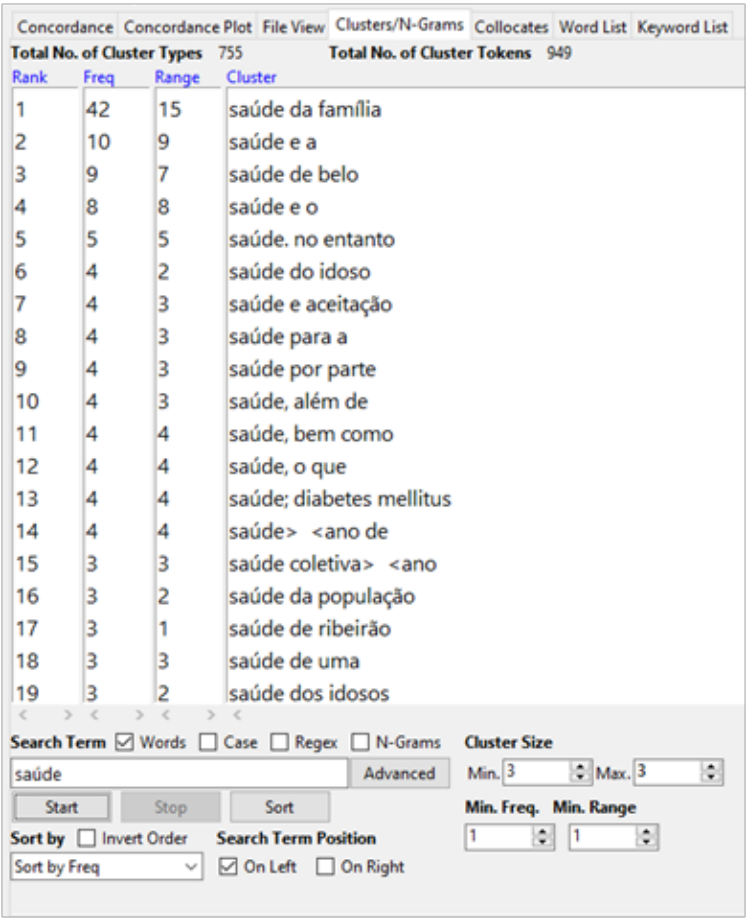
FIGURA 25 – Captura de tela dos clusters/N-grams formados à direita da palavra-chave "saúde" com extensão de quatro palavras



Fonte: Elaborada para fins deste artigo.

Para evitar incoerência com os pressupostos metodológicos estabelecidos para o presente estudo, a busca por clusters/N-grams formados à direita da palavra-chave "saúde" se deu com a extensão de três palavras que resultou em dois clusters/N-grams, com o mínimo de dez ocorrências cada, como mostra a Figura 26 a seguir.

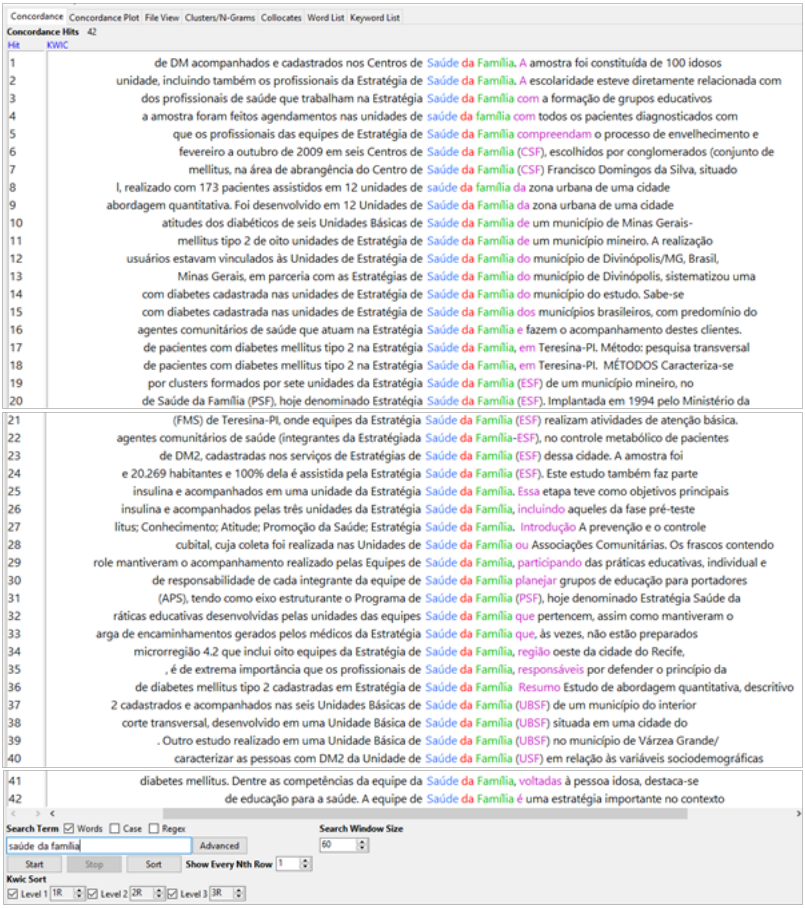
FIGURA 26 – Captura de tela dos clusters/N-grams formados à direita da palavra-chave "saúde"



Fonte: Elaborada para fins deste artigo.

O primeiro cluster/N-gram formado à direita da palavra-chave "saúde", que apresentou o mínimo de dez ocorrências, corresponde a "saúde da família". Adiante, a Figura 27 traz as linhas de concordância em que tal cluster/N-gram aparece.

FIGURA 27 – Captura de tela do primeiro cluster/N-gram formado à direita da palavra-chave "saúde": "saúde da família"



Fonte: Elaborada para fins deste artigo.

A análise das linhas de concordância detalhadas na Figura 27 aponta que o cluster/N-gram "saúde da família" compreende um grupo nominal que está em relação de dependência com outro(s) grupo(s) nominal(is) que varia(m) conforme a linha de concordância examinada, a saber: "os profissionais das equipes de estratégia da saúde da família", "unidades de estratégia de saúde da família", "centro de saúde da família", etc. Entretanto, o cluster/N-gram em questão não abarca essa dependência entre os grupos nominais, porque a busca se deu pela extensão de três

palavras à direita da palavra-chave "saúde", e a dependência é encontrada à esquerda dessa palavra-chave.

Observando as linhas de concordância da Figura 27 de maneira aprofundada, é possível afirmar que das 42 ocorrências mais da metade (23) aborda a "saúde da família" como uma estratégia de assistência aos usuários do sistema único de saúde do Brasil, que ora pode ser viabilizada por uma equipe/profissionais, ora por uma unidade/centro. Em razão disso, o Quadro 16 detalha as funções que podem ser encontradas nos grupos nominais que consideram a "saúde da família" como equipe ou como unidade, respectivamente.

QUADRO 16 – Os grupos nominais possíveis segundo a análise do cluster/N-gram "saúde da família" e as respectivas funções exercidas na ordem do grupo nominal

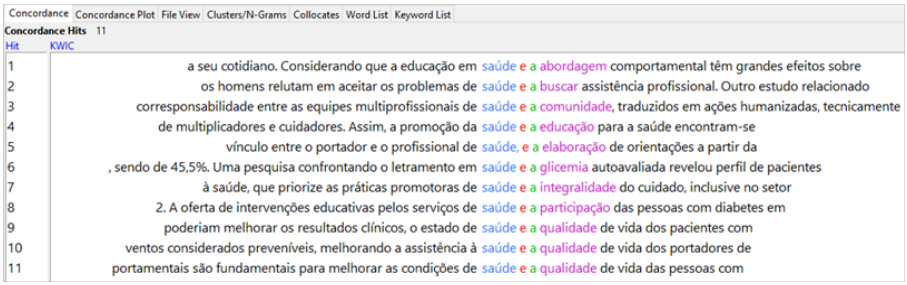
1º grupo nominal	equipes	de saúde	da família
	Ente	**Qualificador**	**Qualificador**
2º grupo nominal	unidades	de saúde	da família
	Ente	**Qualificador**	**Qualificador**

Fonte: Elaborado para fins deste artigo.

Os dados do Quadro 16 indicam que a principal diferença entre os grupos nominais "equipes de saúde da família" e "unidades de saúde da família" diz respeito à realização do Ente, que no primeiro refere-se às "equipes"; e no segundo às "unidades". Em comum, os dois grupos nominais têm as duas qualificações realizadas por frases preposicionais "de saúde" e "da família", respectivamente, ambas responsáveis por identificar o tipo das equipes ou das unidades mencionadas nos textos do corpus utilizado nesta pesquisa.

O segundo e último cluster/N-gram formado à direita da palavra-chave "saúde" com o mínimo de dez ocorrências é "saúde e a". A princípio, tal cluster/N-gram não apresenta sentido completo, pois termina com "a", uma palavra da classe nominal > determinante que precisa coocorrer com outra(s) para fazer sentido numa oração e/ou grupo nominal, por exemplo. A seguir, a Figura 28 destaca as linhas de concordância em que este cluster/N-gram aparece.

FIGURA 28 – Captura de tela do segundo cluster/N-gram formado à direita da palavra-chave "saúde": "saúde e a"



Fonte: Elaborada para fins deste artigo.

O exame das linhas de concordância da Figura 28 indica que não há um padrão de palavras da classe nominal que se repete após o determinante "a". Além disso, o grupo nominal "saúde", presente no cluster/N-gram em questão, parece funcionar em conjunto com as preposições de/em, formando uma frase preposicional, responsável por qualificar diferentes Entes, a depender da linha de concordância analisada. Como não há, portanto, um padrão de coocorrências em relação ao cluster/N-gram "saúde e a" para que algum grupo nominal possa ser extraído e investigado, os resultados encontrados para tal cluster/N-gram não foram contemplados no presente artigo.

A seguir, o Quadro 17 resume todos os grupos nominais investigados neste artigo, derivados das buscas por clusters/N-grams à esquerda e à direita das palavras-chave selecionadas.

QUADRO 17 – Resumo dos clusters/N-grams e grupos nominais analisados

Palavra-chave	**Clusters/N-grams**	**Grupo nominal**
autocuidado	adesão às práticas de autocuidado	as práticas de autocuidado
	Questionário de atividades de autocuidado	Questionário de atividades de autocuidado
	para as práticas de autocuidado	as práticas de autocuidado
	autocuidado de pacientes com diabetes	autocuidado de pacientes com diabetes

diabetes	para o autocuidado em diabetes	o autocuidado em diabetes	
	de autocuidado com o diabetes	autocuidado com o diabetes	
	práticas de autocuidado em diabetes	práticas de autocuidado em diabetes	
	para o controle do diabetes	o controle do diabetes	
	autocuidado de pacientes com diabetes	autocuidado de pacientes com diabetes	
	com a condição do diabetes	a condição do diabetes	
	diabetes mellitus tipo 2 (DM	diabetes mellitus tipo 2	
	diabetes mellitus tipo 2 em	diabetes mellitus tipo 2	
saúde	profissionais da área da saúde	profissionais da área da saúde	
	profissional da área da saúde	profissional da área da saúde	
	uma unidade básica de saúde	uma unidade básica de saúde	
	saúde da família	equipes de saúde da família	unidades de saúde da família
	saúde e a	-	

Fonte: Elaborado para fins deste artigo.

De acordo com os passos metodológicos adaptados do trabalho de Pearson (1998) para a presente pesquisa, a análise dos dados do Quadro 17 aponta que nem todos os grupos nominais examinados podem ser entendidos como termos técnicos, visto que: i. Alguns compreendem nomes de instrumentos que foram utilizados nas pesquisas reportadas nos artigos acadêmicos, como, "Questionário de atividades de autocuidado" e "de autocuidado com o diabetes", por exemplo, os quais correspondem ao título de um questionário (Questionário de atividades de autocuidado com o diabetes); ii. Outros indicam o estabelecimento onde ocorreu parte da pesquisa, como, por exemplo, "uma unidade básica de saúde"; iii. Outros ainda dizem respeito ao tipo de profissional e/ou equipe que contribuiu para a realização da pesquisa, como, "profissionais da área da saúde", "profissional da área da saúde" e "equipes de saúde da família" ou caracterizam o tipo de sujeito que participou da pesquisa, como, por exemplo, usuários/pessoas "com a condição do diabetes".

Por outro lado, há grupos nominais que podem ser compreendidos como termos técnicos sem que haja a necessidade de uma consulta aprofundada ao cotexto onde tal grupo nominal aparece, como, por exemplo, "diabetes mellitus tipo 2". Isso pode ser explicado pelo fato deste grupo nominal se referir ao nome de uma condição crônica, sendo, consequentemente, um termo técnico da área das Ciências da Saúde.

Por último, tem-se os grupos nominais que precisam de uma investigação mais criteriosa para serem validados como termos técnicos. Essa investigação compreende uma sequência de passos estabelecidos a partir da adaptação para o português brasileiro da metodologia proposta no estudo de Pearson (1998). Dentre os grupos nominais detalhados no Quadro 17, os que podem ser caracterizados como termos técnicos, após a análise dos cotextos onde aparecem, são: "adesão às práticas de autocuidado/em diabetes", "o autocuidado em diabetes" e "autocuidado de pacientes com diabetes". A análise mostrou que esses grupos nominais podem ser caracterizados como termos técnicos, porque referem-se a um conjunto de estratégias e/ou competências utilizadas pelos profissionais da Saúde para verificar o empoderamento e os índices de autocuidado dos usuários com a condição crônica do Diabetes Mellitus. O Quadro 18 mostra alguns exemplos que apontam para essa espécie de relação de hiponímia prevista em Pearson (1998) como forma de identificar termos técnicos em corpus (cf. item iii da Subseção 2.2 O trabalho de Pearson (1998) na seção 2 Fundamentação Teórica).

QUADRO 18 – Termo técnico coextensivo a grupo nominal: extração e exemplificação das relações de Pearson (1998)

Termo técnico coextensivo a um grupo nominal	**Exemplos extraídos do corpus** **Exemplos de relação extraídas pelo contexto**
adesão às práticas de autocuidado/práticas de autocuidado em diabetes	O instrumento ESM mede a adesão às práticas de autocuidado do usuário com diabetes. Tem o escore total de oito pontos e abrange questões referentes às atividades de autocuidado, relacionadas à alimentação e à atividade física dos últimos sete dias. Para indicar melhora quanto à adesão às práticas de autocuidado, deve-se obter um escore mínimo de cinco pontos.
	instrumento (ESM) para medir – adesão às práticas de autocuidado
	Com relação à adesão às práticas de autocuidado, as medianas de pontuação no GI aumentaram após o processo educativo, e a comparação entre os grupos intervenção e controle em relação a essa variável evidenciou diferença estatisticamente significativa (p=0,026), indicando uma melhora nas práticas de autocuidado.
	adesão às práticas de autocuidado – essa variável
	intervenção e 111 do Grupo controle. Foram utilizados os questionários de Autocuidado com o diabetes e Diabetes Empowerment Scale-Short Form para comparação entre grupos na linha de base, assim como entre o antes e depois intragrupo. O nível de significância foi 0,05. Resultados: O grupo intervenção apresentou aumento estatisticamente significativo do escore mediano referente à adesão às práticas de autocuidado em diabetes (p=0,005) e à escala de empoderamento (p<0,001). Conclusão: A visita domiciliar promoveu à adesão às práticas de autocuidado com diabetes mellitus tipo 2.
	adesão às práticas de autocuidado em diabetes – escore mediano

autocuidado de pacientes com diabetes	Objetivo: analisar o autocuidado de pacientes com diabetes mellitus tipo 2 na Estratégia Saúde da Família, em Teresina-PI.
	autocuidado de pacientes com diabetes – como único objetivo da pesquisa
	Com exceção da idade e do histórico familiar de DM2, quaisquer outras variáveis podem ser controladas através da mudança no estilo de vida, sendo assim a análise do autocuidado de pacientes com diabetes é de grande importância para regular problemas relacionados com a doença. A educação do paciente na contribuição
	autocuidado de pacientes com diabetes – como outras variáveis

Fonte: Elaborado para fins deste artigo.

O primeiro termo técnico coextensivo ao grupo nominal destacado no Quadro 18 compreende "adesão às práticas de autocuidado/em diabetes". A segunda coluna desse quadro indica as relações estabelecidas entre esse grupo nominal e o contexto em que ele aparece no corpus, as quais tornam-se relevantes para a caracterização desse grupo nominal como um termo técnico de fato. A primeira relação refere-se ao fato de que a "adesão às práticas de autocuidado/em diabetes" compreende algo passível de ser contabilizado de acordo com as atitudes tomadas pelos usuários com essa condição crônica. Isso pode ser feito por meio de um instrumento, bem como pela análise do escore mediano da pesquisa. Além disso, a "adesão às práticas de autocuidado" é referenciada como "essa variável", aspecto apontado por Pearson (1998) como indicativo de termo técnico (cf. item iii da Subseção 2.2 O trabalho de Pearson (1998) na seção 2 Fundamentação Teórica).

Ainda com relação aos dados apresentados no Quadro 18, o segundo termo técnico coextensivo ao grupo nominal diz respeito ao "autocuidado de pacientes com diabetes". Com base nas premissas adotadas no estudo de Pearson (1998) e adaptadas para o presente artigo, esse grupo nominal pode ser caracterizado como termo técnico por duas razões principais: i. Por compreender o único objeto de análise de uma das pesquisas reportadas nos artigos acadêmicos que compõem o corpus da presente pesquisa e ii. Por estabelecer uma relação de hiponímia com "outras variáveis", demonstrando que o "autocuidado de pacientes com diabetes" pode ser definido como uma variável.

Por fim, a análise do contexto do grupo nominal "o autocuidado em diabetes", presente no Quadro 17 como candidato a termo técnico,

não apontou resultados significativos que indicassem que se tratava de um termo técnico de fato. Contudo, é importante salientar que tal análise, assim como as demais, foi feita com base nos resultados descritos no estudo de Pearson (1998) e adaptados para o português brasileiro neste artigo, o que pode influenciar nos resultados obtidos no presente trabalho.

5 Conclusão

Este artigo buscou mostrar como a linguística de corpus, no que concerne o uso das ferramentas disponíveis no *software* concordanciador AntConc, pode funcionar como um recurso acessível para extração de candidatos a termos técnicos em textos especializados, mesmo quando não se tem para o português brasileiro todos os artifícios mencionados no estudo de Pearson (1998), como, por exemplo, o programa de padrão de correspondência e o anotador morfossintático (CLG tagger) desenvolvidos para aquela pesquisa. Contudo, é relevante destacar que o conhecimento e a aplicação dos pressupostos teóricos da linguística sistêmico-funcional acerca do grupo nominal em português brasileiro foram imprescindíveis para suprimir a ausência desses artifícios detalhados em Pearson (1998).

Considerando a ferramenta de geração de clusters/N-grams do AntConc por meio da seleção de palavras-chave de uma lista gerada pela ferramenta de Keywords do mesmo *software* concordanciador, os resultados obtidos indicam que a coextensividade existente entre cluster/N-gram, termo técnico, grupo nominal e item lexical nem sempre funciona de maneira exata, apesar da presença de pelo menos um grupo nominal dentro de todos os clusters/N-gram gerados. Mas, por se tratar de uma forma semiautomática de extração de termos técnicos coextensivos ao grupo nominal, a utilização dessa ferramenta pode ser avaliada como um recurso útil para a busca por termos técnicos em textos especializados, como artigos acadêmicos, por exemplo.

A análise dos grupos nominais que apareceram nos clusters/N-grams revelou que o tipo de texto, artigo acadêmico, bem como o domínio dos textos, diabetes mellitus, selecionados para constituírem o corpus utilizado no presente estudo podem ter influenciado nos tipos de grupos nominais encontrados. Muitos dos grupos nominais examinados faziam parte de outro grupo nominal maior, como, por exemplo, o grupo nominal "Questionário de atividades de autocuidado com o diabetes" formado

pelos clusters/N-gram "Questionário de atividades de autocuidado" gerado para a palavra-chave "autocuidado" e "de autocuidado com o diabetes" gerado para a palavra-chave "diabetes". A maior parte dos grupos nominais analisados apresentou também pelo menos uma frase preposicional com função de Qualificador, como "autocuidado de pacientes com diabetes". Houve também grupos nominais que foram resultado de uma metáfora gramatical (nominalização), por exemplo, "adesão às práticas de autocuidado em diabetes". Essas características – grupos nominais com muitas funções, constituídos por mais de uma palavra e/ou metafóricos – são frequentemente encontradas nos textos em que a linguagem empregada é especializada, cujo autor e leitor são pares e ambos possuem conhecimento experto na área de domínio do texto.

Para além das adaptações do trabalho Pearson (1998) que se fizeram necessárias no escopo da linguística de corpus, resultados diferentes foram observados no que compreende as distinções entre os sistemas linguísticos do inglês, abordados na pesquisa de Pearson (1998), e do português brasileiro, abordado no presente artigo. A principal diferença refere-se aos determinantes com função de dêitico. Em Pearson (1998), a autora menciona que os candidatos a termos técnicos que estavam sinalizados, ou seja, apresentavam algum artigo definido exercendo a função de dêitico não poderiam ser classificados como candidatos a termo técnico. No presente artigo, entretanto, esse aspecto não pode ser considerado relevante, visto que a presença de dêiticos realizados por artigos definidos não se mostrou um impasse para que determinados grupos nominais pudessem estar em coextensividade aos termos técnicos.

Por outro lado, uma espécie de relação de hiponímia mencionada no estudo de Pearson (1998) como uma das maneiras de caracterizar determinado candidato a termo técnico como técnico de fato (cf. Metodologia – item iii) também foi observada neste artigo, sobretudo, nos contextos dos grupos nominais "adesão às práticas de autocuidado/ em diabetes" e "autocuidado de pacientes com diabetes" que apresentaram essa relação com "essa variável" e com "outras variáveis", respectivamente, e puderam, posteriormente, serem caracterizados como termos técnicos. É válido mencionar que essa espécie de relação de hiponímia sugerida por Pearson (1998) como uma maneira de caracterizar e encontrar um termo técnico em um texto especializado também já foi

reportada no trabalho de Figueredo *et al*. (2019) como uma das maneiras possíveis de caracterizar um item lexical em português brasileiro.

Ainda no campo dos fatores concordantes entre o estudo de Pearson (1998) e a pesquisa detalhada neste artigo encontram-se dois aspectos que podem ser considerados limitantes, mas passíveis de serem investigados em pesquisas futuras. O primeiro deles está relacionado à ideia de consultas a especialistas na área dos textos que constituem o corpus de pesquisa, no caso um profissional da área da Saúde. Esses especialistas funcionariam como uma fonte extralinguística capaz de corroborar se determinado termo poderia ser considerado um termo técnico de fato. O segundo aspecto corresponde ao tamanho do corpus de pesquisa utilizado que pode interferir nos resultados encontrados, ou seja, com um corpus de maiores proporções, a quantidade de termos candidatos a termos técnicos também poderia ser maior, embora a análise manual proposta no presente trabalho poderia tornar-se inviável pela demanda extra de tempo a ser despendido.

Declaração de autoria

Júlia Santos Nunes Rodrigues: compilação do corpus utilizado na pesquisa, anotação da pesquisa, escrita das seções metodologia e resultados, escrita resumo e abstract e revisão do artigo.
Kícila Ferreguetti: auxílio na compilação do corpus de referência, escrita da seção de fundamentação teórica, formatação do artigo nos moldes da revista, incluindo a formatação das referências e revisão do artigo.
Adriana S. Pagano: supervisão da pesquisa, escrita das seções introdução e conclusão e revisão do artigo.

Referências

ALMEIDA, G. M. D. B.; CORREIA, M. Terminologia e corpus: relações, métodos e recursos. *In*: TAGNIN, S. E. O.; VALE, O. A. (org.). *Avanços da Linguística de Corpus no Brasil*. São Paulo: Humanitas, 2008. p. 67-90.

ALMEIDA, L. B. Identidade científica da Terminologia. *In*: _____. *Curso básico de Terminologia*. São Paulo: Edusp, 2004. p. 25-96.

ANTHONY, L. AntConc Homepage, *Laurence Anthony Website,* Tokyo, Version 3.5.8, 2019. Disponivel em: https://www.laurenceanthony.net/software/antconc/. Acesso em: 22, Fevereiro, 2019.

BOWKER, L.; PEARSON, J. *Working with Specialized Language.* A Pratical Guide do Using Corpora. 1. ed. London; New York: Routledge, 2002. DOI: https://doi.org/10.4324/9780203469255

FERREGUETTI, K. A *Frase preposicional com função de qualificador no grupo nominal*: um estudo de equivalentes textuais no par linguístico inglês e português brasileiro. 2018. 154f. Tese (Doutorado em Estudos Linguísticos) – Faculdade de Letras, Universidade Federal de Minas Gerais, 2018.

FIGUEREDO, G. *Uma descrição sistêmico-funcional do grupo nominal em português brasileiro*. 2007. 292f. Dissertação (Mestrado em Estudos Linguísticos) - Faculdade de Letras, Universidade Federal de Minas Gerais, 2007.

FIGUEREDO, G. P. *et al*. O léxico como um recurso linguístico para a produção de significado no texto: um estudo de caso com protocolos de investigação. *Estudos da Língua(gem),* Vitória da Conquista, v. 17, n. 3, p. 37-59, 2019. DOI: https://doi.org/10.22481/el.v17i3.5928

FIGUEREDO, G. P.; PAGANO, A. S.; FERREGUETTI, K. Os sistemas textuais de focalização na organização funcional da gramática do Português Brasileiro. *D.E.L.T.A.,* São Paulo, v. 30, n. 2, p. 309-352, 2014. DOI: https://doi.org/10.1590/0102-445080334301692532

FINATTO, M. J. B. Exploração terminológica com apoio informatizado: diálogos entre terminologia e linguística de corpus. *In*: LORENTE, M. *et al.* (org.). *Estudis de linguística i de linguística aplicada en honor de M. Teresa Cabré Castellví*. Barcelona: Institut Universitari de Linguística Aplicada de la Universitat Pompeu Fabra, 2007. p. 221-230.

HALLIDAY, M. A. K. Categories of the Theory of Grammar. *In*: _____. *Collected Works of M.A.K. Halliday.* London; New York: Continuum, 1961. p. 37-88.

HALLIDAY, M. A. K. *Language as Social Semiotic*. The Social Interpretation of Language and Meaning. 1. ed. London: Edward Arnold, 1978.

HALLIDAY, M. A. K. *On Grammar*. 1. ed. London; New York: Continuum, 2002.

HALLIDAY, M. A. K.; MATTHIESSEN, C. *Construing Experience as Meaning:* A Language Based Approach to Cognition. London: Cassell, 1999.

HALLIDAY, M. A. K.; MATTHIESSEN, C. M. I. M. *An Introduction to Functional Grammar*. 3. ed. London: Routledge, 2014. DOI: https://doi.org/10.4324/9780203431269

HAO, J. *Construing biology*: An Ideational Perspective. 2015. These (PhD of Linguistics) – Department of Linguistics University of Sydney, Sydney, Sydney, 2015.

PEARSON, J. *Terms in Context*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 1998.

SARDINHA, B. Visão geral da Linguística de Corpus. *In*: SARDINHA, T. B. *Linguística de Corpus.* São Paulo: Editora Manole, 2004. p. 1-42.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1381-1414, 2021

The Pragmatics of Aeronautical English: an investigation through Corpus Linguistics

A Pragmática do inglês aeronáutico: uma investigação pela Linguística de Corpus

Malila Carvalho de Almeida Prado
Fujian University of Technology (FJUT), Fuzhou, Fujian / China
malilaprado@hotmail.com
https://orcid.org/0000-0001-6281-6759

Abstract: The ICAO Language Proficiency Rating Scale offers parameters for aeronautical English teaching and assessment focused on oral skills. It assists governments worldwide in assessing pilots and air traffic controllers' English proficiency, licensing them for international operations. This paper addresses two of the six linguistic areas listed in the Rating Scale, namely fluency and interaction, to understand what conversational elements are present in pilot-controller communications with a view to informing pedagogical material. The analysis is based on a corpus of pilot-controller radio communications in abnormal situations, revealing a more spontaneous code as opposed to the documented Standard Phraseology mandated for routine situations. Corpus Linguistics is the methodology chosen for this investigation, concentrated on the top frequent three-word clusters extracted from the corpus. Investigation of these clusters reveals that fluency and interaction are interconnected and should be considered in a broader perspective that takes into account language in use. To illustrate, 'we'd like' and 'if you can' are commonly employed as requests in this specific register. The paper concludes by suggesting that learners' awareness of pragmatic aspects of language is pivotal in the aviation English classroom.

Keywords: Plain Aviation English; fluency; interaction; Corpus Linguistics; Pragmatics.

Resumo: A Escala de Proficiência Linguística da ICAO oferece parâmetros para o ensino e a avaliação do inglês aeronáutico focado nas habilidades orais. Serve para os governos em todo o mundo avaliarem a proficiência em inglês de pilotos e controladores de tráfego aéreo, licenciando-os para operações internacionais. Este estudo aborda

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1381-1414

duas das seis áreas linguísticas elencadas na Escala, quais sejam, fluência e interação, para compreender quais elementos conversacionais estão presentes nas comunicações entre pilotos e controladores com o objetivo de subsidiar materiais pedagógicos. A análise se baseia em um *corpus* de comunicações via rádio entre pilotos e controladores em situações anormais, revelando um código mais espontâneo, diferentemente da Fraseologia Padrão oficial mandatória nas situações rotineiras. A Linguística de *Corpus* é a metodologia utilizada nesta investigação, concentrada nos mais frequentes blocos de linguagem de três palavras evidenciados no *corpus* de estudo. A investigação desses blocos de linguagem revela que fluência e interação são interconectadas e deveriam ser consideradas a partir da perspectiva da língua em uso. Para ilustrar, 'we'd like' e 'if you can' são normalmente empregados como solicitações. Conclui-se sugerindo que a conscientização dos aprendizes sobre aspectos pragmáticos da língua é fundamental na sala de aula do inglês aeronáutico.

Palavras-chave: Plain Aviation English; fluência; interação; Linguística de *Corpus*; Pragmática.

Submitted on October 10th, 2020

Accepted on November 23th, 2020

1 Introduction

Even following the spread of the communicative approach and the stimulus in promoting authentic language in the language classroom, research shows a different scenario (RÜHLEMANN, 2008). This may be a result of a lack of understanding of the characteristics of language use, in particular of the importance usually given to language form rather than language use (MCCARTHY; CLANCY, 2018). On one hand, authenticity in the classroom is sometimes criticized over certain features found pedagogically difficult to deal with, such as hesitation, false starts, and speed of delivery (cf. WIDDOWSON, 1998). On the other hand, promoting strategies that help the learner tackle authentic language use may contribute to the learning process from the start (FIELD, 2009).

In language testing, particularly in the field of English for Specific Purposes (ESP), Douglas (1999) argues that real-life tasks should be implemented in language proficiency tests as a means of truly and fairly analyzing the candidates' production. This has shown to be highly relevant in aviation English studies such as Kim (2018), which compares the language production of both novice and experienced air

traffic controllers and pilots: the more experienced the professional, the better the performance when assessed in real-life tasks.

Increasing attention has been drawn to aviation English since pilots and air traffic controllers were required to show sufficient English language proficiency to operate internationally. This proficiency requirement is described in the Manual of Implementation of the Language Proficiency Requirements (ICAO, 2004, 2010), which also specifies the Language Proficiency Rating Scale (Scale henceforth) that guides raters responsible for granting licenses to the above-mentioned professionals. The Scale is divided into six language areas: pronunciation, structure, vocabulary, fluency, comprehension, and interaction distributed across six different levels of proficiency.

Some studies have criticized the Scale by questioning its authenticity, particularly when considering radio communications held between pilots and controllers in abnormal situations, an avowed interest of the International Civil Aviation Organization (ICAO), as noted in the second edition of the Manual (ICAO, 2010). ICAO documents recommend that Standard Phraseology, a specialized and rehearsed register, be used in all routine situations of a flight. However, when abnormalities occur, such as engine failures or bird strikes, pilots and controllers need to resort to what is referred as "Plain Aviation English", a more spontaneous language placed between the documented Standard Phraseology and everyday conversations (BIESWANGER, 2016, p. 83). Both Standard Phraseology and Plain Aviation English belong to the realm of aeronautical English and are equivalent to the language used by pilots and controllers on the radio; all other portions of language (produced by crew members, mechanics, flight attendants) go under the umbrella of Aviation English (TOSQUI-LUCKS; SILVA, 2020). For the purposes of this paper, I aim to study the Plain Aviation English, that is, a sub-register of aeronautical English.

In aviation, any minor problem can become a disaster (cf. FRIGINAL, MATHEWS; ROBERTS, 2020; WEIR, 1999), and all areas of communication therefore deserve attention. Many studies, including those listed in Doc 9835 (ICAO, 2010), draw on accidents to which miscommunications were a contributory cause (FRIGINAL, MATHEWS; ROBERTS, 2020). Nevertheless, Mathews (2012, 2020) claims that there may be more incidents and accidents to which language is a contributing factor than we are aware of, given that accident investigations often fail

to consider linguistic expertise, stressing that the knowledge and tools applied in the investigation of operational and mechanical complications are far more meticulous than those used in human factor issues, particularly in communication.

Most research has tended to focus on aeronautical English as this is the main interest of ICAO Language Proficiency Requirements (LPRs). Some studies have pointed out the lack of attention or even vagueness in the description of language areas (cf. ESTIVAL; FARRIS; MOLESWORTH, 2016; GARCIA, 2015). Others address the need for more research into communicative elements excluded from the Scale, such as interactional competence (MONTEIRO, 2019), cross-cultural competence (BOROWSKA, 2017), and ELF communicative strategies (ISHIHARA; PRADO, in press). Beyond merely pointing out problems with the Scale, these studies suggest linguistic manifestations that may equip Scale users, such as the lack of correspondence between the Scale and real-life scenarios may cause misunderstandings among Scale users and, more dangerously, misconceptions (cf. PFEIFFER, 2009).

Mathews (2020) points out that despite the criticism to which the LPRs have been subjected, this was a useful starting point because it brought about not only testing and teaching programs worldwide but also academic research. As one of the designers of ICAO documents, Mathews emphasizes that academic and industrial collaborations are key to advances in this area.

Bearing in mind that, in aeronautical English, Plain Aviation English should resemble the Standard Phraseology in aspects such as clarity and objectivity, and seeking to examine how the description of fluency and interaction present in the Scale compares to the specific verbal-only communication in moments considered non-routine, I compiled a corpus described in Section 4 of this paper of pilot-controller radio communications in abnormal situations to allow for an investigation of such elements. This corpus allows for an examination of the data so as to consider two questions, which are (1) what linguistic elements correspond to interaction and fluency in radio communications in abnormal situations?; and (2) what elements can compose an aeronautical English teaching curriculum?

This paper is structured as follows: I first address research on conversational elements of aeronautical English. Next, I discuss characteristics of oral language as well as findings from studies of spoken

corpora. The methodology and the corpus used for this investigation are then presented, followed by the analysis of the data highlighted. I conclude by raising the importance to intercultural pragmatics and communicative strategies in the pedagogy of Aviation English and offer suggestions for how to approach these in the aeronautical English classroom.

2 Oral elements in Aeronautical English

The two linguistic areas, fluency and interaction, this paper intends to address are now described, starting from fluency:

> Produces stretches of language at an appropriate tempo. There may be occasional loss of fluency on transition from rehearsed or formulaic speech to spontaneous interaction, but this does not prevent effective communication. Can make limited use of discourse markers or connectors. Fillers are not distracting.

This is the rationale for a level 4 candidate, that is, a candidate who is granted the language proficiency license for international operations. In fluency, keywords such "tempo", "discourse markers", "fillers" can be spotlighted and paralleled to the viewpoint common at the time of the publication of Doc 9835 (ICAO, 2004), which refers to hesitation as an "occasional loss of fluency". However, the last years have seen a change in this perspective, as studies on spoken corpora have shown that hesitation, especially the filled pause (e.g. uh, um, er), is an important item used as a strategy to request assistance from the interlocutor or to signal a change of ideas, for example, and, as such, should be considered a word or a linguistic event rather than an indication of a loss for words (GÖTZ, 2013). This updated definition of fluency brings it closer to Monteiro's study (2019) of interactional competence in aviation English testing.

In interaction, the Scale states the following for level 4:

> Responses are usually immediate, appropriate and informative. Initiates and maintains exchanges even when dealing with an unexpected turn of events. Deals adequately with apparent misunderstandings by checking, confirming or clarifying.

The concern over prompt responses as well as their quality calls attention to another feature that has been questioned elsewhere: that of placing the burden of the communication on a person only, rather than considering it as a two-way endeavor (MCNAMARA, 2011). The transition from Standard Phraseology to Plain Aviation English is also taken into account – as it is stated in fluency as well. Communicative strategies are listed as "checking", "confirming" and "clarifying", which are related to communication repairs. However, such strategies can be used in routine situations and it is worth investigating what other strategies can be employed in abnormal situations (e.g. MONTEIRO, 2019).

Mell (2004) analyzed a corpus built in France from transactions between French controllers and international traffic. He verified that more than 75% of the language used in radio communications in routine situations regards the management of the communication itself through functions such as the "expression of satisfaction or complaint, reprimand, concern or reassurance, apologies, […] opening or closing, self-correction, readback, acknowledgement, checking, repetition, confirmation, clarification, or relaying" (MELL, 2004, p. 13). A more thorough list compiled by Mell in his thesis defended in 1991 can be found in one of the annexes to the ICAO Manual (ICAO, 2004, 2010). Lopez (2013) followed in Mell's footsteps and investigated an updated version of the corpus. She too concluded that social conventions play an important role in radio communications and language and that even in a restricted environment such as aviation, language cannot be controlled.

Nevertheless, Garcia and Fox (2020) argue that listening – or comprehension, as expressed in the Scale – should have an exclusive scale given that it is a much more complex activity. Regarding pronunciation, specifically speech rate, ICAO's Standard Phraseology recommends that a maximum of 100 words per minute be used (ICAO, 2007). However, a study of this speech rate revealed that it is too slow for radio environments and in fact compromises understanding (BIESWANGER, 2013). In a comparison between a radio communication corpus and a professional radio broadcaster's corpus, Trippe and Baese-Berk (2019) concluded that pilots and controllers tend to have a faster speech rate.

Kim (2018) conducted a study with six professionals, analyzing their perceptions of a real communication between a Russian pilot and a Korean controller. The professionals recognized that despite clear

linguistic limitations, the pilot acted professionally and handled the communication effectively, whereas the controller, even though he demonstrated a higher linguistic level, did not show the same experience in dealing with the problem, thus overloading the pilot. Along with work by Moder and Halleck (2009), Knock (2014), and Emery (2014), this line of research heeds technical knowledge combined with language proficiency. In addition, McNamara (2011) argues that air-ground communications are held between two participants at least as opposed to being an individual responsibility. Thus, training pilots and controllers to communicate effectively on the radio "should emphasize collaborative principles rather than focusing on terminology or isolated practices" (MORROW; RODVOLD; LEE, 1994, p. 255).

3 Studies of spoken language

The notion that spoken and written forms of language share the same characteristics has long been outdated, but it still orients many English as a Foreign Language (EFL) and English as a Second Language (ESL) coursebooks available in the market (cf. CARTER; MCCARTHY, 2017; RÜHLEMANN, 2008). Studies that emerged from the 1970s (SACKS; SCHEGLOFF; JEFFERSON, 1974; SINCLAIR; COULTHARD, 1975) turned to oral language from a more empirical perspective, with attention given to transcription modes, eventually including their storage in computers. Researchers then realized the need for a better – and as faithful as possible – representation of spoken language. This viewpoint allowed new fields such as Discourse Analysis, Conversation Analysis, and studies of spoken corpora to develop. Nevertheless, the challenges involved in gaining access to or recording natural spontaneous speech and transcribing it hindered advances in compiling large amounts of data, a situation that only began to change through projects such as the Santa Barbara Corpus of American English (DUBOIS, 1991) and the London-Lund Corpus of Spoken English (SVARTVIK, 1990).

The faithfulness of transcriptions is often questioned as transcriptions represent only part of the actual event (ZANETTIN, 2009). However, analysis carried out from empirical evidence yields findings that once were solely based on intuitions (cf. RÜHLEMANN, 2008). Comparisons between spoken and written forms generated materials

aiming at new descriptions of language in a more grammar-usage based style (e.g., CARTER; McCARTHY, 2006), lexically centered (LEWIS, 1993), or even drawing attention to lexico-grammatical patterns (SINCLAIR, 1991). Coursebook writers and material designers then began to employ these findings but still within frameworks built mostly upon generative or universal concepts of linguistics (DAVIES, 2004). Eventually, social theories began voicing the importance of other competences to be included in the EFL/ESL curriculum, including pragmatic, interactional, and strategic competence (CORBETT, 2003; DAVIES, 2004; YOUNG, 2000). Research followed suit, eventually shifting from years of work about grammar, lexicon and pronunciation to a more process-oriented perspective, particularly in studies that used corpora in pragmatics (cf. O'KEEFE; CLANCY; ADOLPHS, 2011), English as a Lingua Franca (ELF) (MAURANEN, 2018), and learner language (GRANGER, 2008). However, this product-oriented focus on lexico-grammatical patterns (and, to a lesser extent, on pronunciation; see JENKINS, 2000) highlighted in the usage-based data was widely perceived as "wrong" data as it did not correspond to the norms prescribed by grammarians and native speaker standards (MAURANEN, 2018). In Mauranen's words, "linguistic structures reflect the demands of communication, not the other way round, with communication shaped by available linguistic structures" (MAURANEN, 2018, p. 13).

The process-oriented approach concerns empirical observations of certain phenomena co-constructed within the interaction. It considers communication as a social, conventional enterprise that evidences transparent elements such as lexical choice, level of politeness, register, and less transparent items such as power relations and cultural factors (including indirect speech acts), among others. Some of these elements are described in literature on communicative strategies (KAUR, 2019), turn initiators (TAO, 2003), fluency enhancement strategies (GÖTZ, 2013), speech acts (ADOLPHS, 2008), mitigation (CAFFI, 1999) and communication breakdowns (GARDINER; DETERDING, 2018), to name a few. Such investigations also review concepts in the teaching of English, especially grammar, which, according to Rühlemann (2008), should focus on the structures of spoken grammar, found in strings of language that contain a "functional profile" (ADOLPHS, 2008), or pragmatic speech act. This can range from a speech act to a false start and is constrained by the context of language production.

4 Method

This study's chosen methodology derives largely from Corpus Linguistics (CL). CL's starting point is the compilation of a corpus, a computer-stored bank of texts collected mostly with research purposes in mind (TAGNIN, 2013) – although more and more uses of corpora are now seen in areas such as glossary making or teaching (CHENG, 2015). To be included in a corpus, texts must meet certain conditions such as emerging from naturally occurring environments, whether in written or spoken form or belonging to any specific genres, among others.

Two key principles underlie CL research: the open-choice principle, and the idiom principle (SINCLAIR, 1991). The first corresponds to the creative use of language, whereas the latter regards the storage of semi-structured language available to the user. The idiom principle is the interest of the present research as it conceptualizes language as socially produced, through entrenchments cognitively stored and conventionalized through common use by a given community; these strings of language, or clusters, spare the speaker the burden of producing new language on every occasion (O'KEEFE *et al.*, 2011). Because the interest of CL is conventionalized patterns, analysis usually starts from generating lists based on the frequency of occurrence in the corpus, which in turn highlight the most frequent words. Researchers then look at them more deeply, using tools such as keyword lists (by comparing two corpora, the researcher can extract those words that are exclusive to or more commonly used in the corpus), but also cluster lists (frequent two-, three-, four- or more strings of words), and concordance lines (the lines of text excerpts in which a node word appears centrally so that it may be observed in its surroundings), among others. The choice of tools depends on the research question.

I now turn to the methodology used in this study. In the investigation of spoken phraseology, that is, patterns commonly used in oral speech, Altenberg (1998) generated two-, three-, four-, five- and six-word clusters and compared their frequency with single word lists in the London-Lund Corpus of Spoken English (http://www.helsinki.fi/varieng/CoRD/corpora/LLC). Through this comparison, the researcher identified clusters corresponding to up to 80% of the corpus. Apart from functional or grammar words such as *in, the,* or *of*, the most frequent single words were not as frequent as many of the two-, three- and four-word clusters. The researcher then grouped these clusters under grammatical categories

such as dependent clauses, independent clauses, and incomplete clauses (ALTENBERG, 1998). Following a similar methodology, McCarthy and Carter (2002) extracted two-, three-, four-, five- and six-word clusters from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE: cf. https://www.nottingham.ac.uk/research/groups/cral/projects/cancode.aspx), but rather than using grammar as an overarching element, they observed pragmatic integrity in the clusters. That is, when analyzed in the concordance lines and in the source texts, each cluster demonstrated common pragmatic categories such as discourse marking, facework, politeness, and purposive vagueness. These categories broadly correlated with the pragmatic routines in Bardovi-Harlig (2012, p. 208) in that in order to be identified as a pragmatic routine, an expression must: (1) contain at least two morphemes; (2) be articulated without any interruption; (3) be repeated in the same way; (4) be dependent on the context; and (5) be community-wide.

To identify the spoken phraseology of Aeronautical English in abnormal situations, I investigated the RadioTelephony Plain English Corpus (RTPEC – PRADO; TOSQUI-LUCKS, 2019). RTPEC consists of 130 audio files transcribed into 110,737 words. All audio files feature communications between pilots and controllers in which abnormal situations occur and presumably contain Plain Aviation English (BIESWANGER, 2016). Guiding the abnormal situations represented in the corpus is another document published by the ICAO, namely Taxonomy of Occurrences, a list that standardizes accident and incident reports (ICAO, 2006). The occurrences presented in this taxonomy refer to operational problems that might occur during a flight such as engine failure, loss of flight controls, bird strikes, weather-related phenomena such as windshear or icing, even human-related scenarios such as problems with passengers or violations such as runway incursions (i.e., inadvertent entry onto the runway). For each of the 33 categories listed in the Taxonomy of Occurrences, there are four to six audio files, of which at least one must have been held in international traffic, that is, an aircraft foreign to that airspace or airport, as a way of ensuring ELF interactions in the corpus (cf. PRADO, 2019).

The transcriptions partially followed the model of Language Into Act Theory (CRESTI, 2000). However, the linguistic or metalinguistic information suggested was not included because the corpus is also intended for pedagogical material design. Still, the principle that oral

language is prosodically centered (see example below) rather than sentence or verb centered conducted the transcriptions. The fact that meaning is constructed through islands defined within prosodic frontiers allows us to observe each island as containing units of meaning, which in turn correspond to a speech act (CRESTI, 2014). Each prosodic unit is represented as an utterance between single slashes, and a full utterance, identified by the fall in intonation, is closed by two slashes. The following extract illustrates this point:

> Uh **you know what** / I'd like to turn back and maybe go to republic if that's okay / uh seven nine November //

The islands are the portions of language between the slashes. Because of the slashes, it is possible to identify "uh you know what" as a string and "I'd like to turn back and maybe go to republic if that's okay" as another string. If the slash had not been used, the researcher might consider "you know" as a string and "what I'd like to" as another. The prosodic fall, identified by the slash, may separate the strings, or the islands. These strings match the idiom principle (SINCLAIR, 1991 – see Section 2) in the sense that conventionalized entrenched language, including collocates, colligates, and clusters, is easily spotted in the concordance lines that exhibit the slashes. Table 1 illustrates this point:

TABLE 1 – Sample of concordance lines with "you know what"

N	Concordande		
1	orty-five // Alright /	**you know what**	/ in that case just pu
2	said Juliet // Okay /	**you know what**	/ for now just hold sh
3	trying to imply // Uh	**you know what**	/ I'd like to turn bac
4	ted fifteen? // Well /	**you know what**	/ they're gone / but q
5	the runway // Okay //	**you know what**	/ Tower / Can you have
6	ed somebody else // Uh	**you know what**	/ most of us sir are l
7	d have known by now //	**you know what**	/ they told us it was

In Table 1, the seven occurrences of the expression "you know what" mostly follow fillers (*alright, okay, uh, well*) and occur at the beginning of utterances. These occurrences were previously selected out of 17 because of their common feature, namely that they function as discourse organizers, a fact we can only observe when investigating

the cluster in question in the source text. It is therefore more practical to identify a cluster as being within an utterance. If an investigated cluster is separated by double slashes, it is disregarded as it does not form a unit of meaning. This transcription model benefits the search for units of meaning that are not semantically transparent, that is, clusters that contain only functional or grammar words. However, the high frequency of such elements in spoken corpora may signal certain uses in the community that could have gone unnoticed if other transcription models had been adopted. Although concordance lines assist the researcher in looking at clusters within utterances, they still do not reveal the context or even the exchange in which the cluster was used (WEISSER, 2018). Therefore, a careful examination of each cluster within the context of production is indispensable.

To run the analysis of the corpus, in line with McCarthy and Carter (2002), I generated two-, three- and four-word cluster lists through Wordsmith Tools (SCOTT, 2016), intending to extract the conventional elements – or the spoken phraseology (ALTENBERG, 1998) – present in the corpus and compare them to the two linguistic areas, fluency and interaction, targeted at in this paper. The most frequent clusters were selected for individual analysis in concordance lines to identify whether or not they belonged to the same cluster, specifically in the same island (CRESTI, 2014), and then in the source text, so as to investigate their pragmatic function by observing the features listed in Bardovi-Harlig (2012). The next section presents the analysis.

5 Analysis and Discussion

To observe whether clusters in the corpus are more frequent than single words, as in McCarthy and Carter (2002), I ran a broader investigation of two lists: a wordlist (TABLE 2) and a two- to four-word cluster list (TABLE 3) by means of Wordsmith Tools (SCOTT, 2016).

TABLE 2 – RTPEC Wordlist with 60 most frequent words

N	Word	Freq.	N	Word	Freq.	N	Word	Freq.
1	THE	2,800	21	IT	615	41	SO	260
2	YOU	2,795	22	S	589	42	GOOD	258
3	TO	2,536	23	HAVE	577	43	WITH	257
4	UH	2,124	24	TURN	526	44	BY	254
5	WE	1,830	25	JUST	521	45	WILL	245
6	AND	1,648	26	IN	508	46	OFF	244
7	ON	1,211	27	ARE	476	47	DOWN	242
8	RIGHT	1,056	28	CAN	474	48	THIS	238
9	FOR	998	29	NOW	449	49	GET	232
10	I	967	30	YOUR	417	50	LIKE	229
11	A	964	31	LL	413	51	BACK	224
12	RUNWAY	931	32	BE	406	52	UP	224
13	THAT	853	33	THANK	392	53	WHAT	223
14	LEFT	837	34	DO	384	54	FROM	218
15	IS	822	35	THERE	366	55	OUT	213
16	RE	756	36	GO	348	56	AIRCRAFT	207
17	AT	683	37	IF	347	57	HERE	201
18	OF	683	38	NEED	295	58	AN	199
19	OKAY	673	39	SIR	285	59	ME	199
20	HEAVY	625	40	GONNA	277	60	WHEN	199

TABLE 3 – RTPEC 2-, 3- and 4-word cluster lists

2-word clusters			3-word clusters		4-word clusters	
N	*Word*	*Freq.*	*Word*	*Freq*	*Word*	*Freq.*
1	WE RE	436	WE RE GONNA	111	CLEARED FOR TAKE OFF	72
2	THANK YOU	392	HOLD SHORT OF	74	THANK YOU VERY MUCH	45
3	ON THE	330	ON THE RUNWAY	62	LINE UP AND WAIT	41
4	YOU RE	273	D LIKE TO	51	WE D LIKE TO	33
5	TO THE	271	I DON T	50	CLEARED TO LAND RUNWAY	32
6	WE LL	233	WE D LIKE	49	ESTABLISHED ON THE LOCALIZER	24
7	AND UH	212	LET ME KNOW	46	HOLD SHORT OF RUNWAY	20
8	UH WE	211	THANK YOU VERY	45	ARE YOU ABLE TO	19
9	YOU CAN	202	YOU VERY MUCH	45	DID YOU COPY THAT	19

10	IF YOU	200	UH WE RE	44	I DON T KNOW	19
11	IT S	186	DO YOU HAVE	42	WOULD YOU LIKE TO	16
12	THAT S	185	YOU RE CLEARED	41	DO YOU WANT TO	15
13	I M	177	AND UH WE	40	JUST LET ME KNOW	15
14	DO YOU	175	SOULS ON BOARD	40	LET ME KNOW WHEN	15
15	WE HAVE	172	WE NEED TO	38	WE RE GONNA HAVE	15
16	WE ARE	164	YOU NEED TO	38	ME KNOW WHEN YOU	14
17	RE GONNA	149	DO YOU WANT	37	RE GONNA HAVE TO	14
18	ARE YOU	137	AT THIS TIME	36	SOULS ON BOARD AND	14
19	YOU HAVE	126	OKAY THANK YOU	36	THE AIRPORT IN SIGHT	14
20	RIGHT NOW	124	I M GONNA	35	WE RE GOING TO	14
21	NEED TO	119	I M SORRY	35	WHEN YOU RE READY	14
22	AND WE	118	TO THE GATE	35	YOU WANT US TO	14
23	DON T	112	AND WE LL	33	AND UH WE RE	13
24	I LL	111	IF YOU CAN	32	I NEED YOU TO	13
25	YOU NEED	106	DO YOU NEED	31	I LL GIVE YOU	12
26	LIKE TO	99	YOU RE GONNA	30	IN FRONT OF YOU	12
27	WHEN YOU	92	WE LL BE	29	PAN PAN PAN PAN	12
28	WILL BE	91	DECLARING AN EMERGENCY	28	WE RE GONNA NEED	12
29	CAN YOU	85	TO THE RAMP	28	ARE YOU READY TO	11
30	YOU WANT	85	UH WE ARE	27	BACK TO THE GATE	11

A comparison of Tables 2 and 3 shows that only 29 single words (TABLE 2) are more frequent than the most recurrent cluster, which is "we're" (TABLE 3). Other than revealing the importance of clusters as inherent in radio communications, the high presence of clusters in this register also supports the view that radio communication in problem-solving situations resembles oral speech, as predicted in Lopez (2013).

As the objective of this study research is not to investigate Standard Phraseology but the Plain English used in radio communications, a stoplist was needed to remove Standard Phraseology words such as numbers, items from the ICAO phonetic alphabet (Alpha, Bravo, Charlie), and proper nouns (airports, airlines), among others. The search for five- and six-word clusters only brought up strings such as "thank you very much sir." Therefore, these were excluded from the list of items to be investigated.

This initial investigation also highlighted how most two-word clusters were in fact segments of three-word clusters such as “do you” (“do you have” or “do you need”). Below is a list of the 100 top three-word clusters; the highlighted expressions are aviation-related and confirm the nature of the communications (TABLE 4).

TABLE 4 – RTPEC 100 most frequent 3-word clusters

N	Word	Freq.	N	Word	Freq.
1	WE RE GONNA	111	51	THANK YOU SIR	22
2	HOLD SHORT OF	74	52	TO THE RIGHT	22
3	**ON THE RUNWAY**	62	53	WHEN YOU GET	22
4	D LIKE TO	51	54	WHEN YOU RE	22
5	I DON T	50	55	DON T HAVE	21
6	WE D LIKE	49	56	GONNA HAVE TO	21
7	LET ME KNOW	46	57	RE GOING TO	21
8	THANK YOU VERY	45	58	RE GONNA HAVE	21
9	YOU VERY MUCH	45	59	ROGER THANK YOU	21
10	UH WE RE	44	60	UH WE HAVE	21
11	DO YOU HAVE	42	61	YOU RE READY	21
12	YOU RE CLEARED	41	62	**AIRPORT IN SIGHT**	20
13	AND UH WE	40	63	IN FRONT OF	20
14	**SOULS ON BOARD**	40	64	IT S A	20
15	WE NEED TO	38	65	LL GIVE YOU	20
16	YOU NEED TO	38	66	**OF THE AIRCRAFT**	20
17	DO YOU WANT	37	67	**TAXI TO THE**	20
18	AT THIS TIME	36	68	UH DO YOU	20
19	OKAY THANK YOU	36	69	WE HAVE A	20
20	I M GONNA	35	70	YOU HAVE A	20
21	I M SORRY	35	71	AND WE RE	19
22	**TO THE GATE**	35	72	AS SOON AS	19
23	AND WE LL	33	73	FOR YOUR HELP	19
24	IF YOU CAN	32	74	**IN THE COCKPIT**	19
25	DO YOU NEED	31	75	**PAN PAN PAN**	19
26	YOU RE GONNA	30	76	WE HAVE UH	19
27	WE LL BE	29	77	WE LL GET	19
28	**DECLARING AN EMERGENCY**	28	78	YOU ABLE TO	19

29	**TO THE RAMP**	28	79	YOU COPY THAT	19
30	UH WE ARE	27	80	AT THE MOMENT	18
31	WOULD LIKE TO	27	81	CALL YOU BACK	18
32	WOULD YOU LIKE	27	82	GIVE YOU A	18
33	YOU WANT TO	27	83	IF YOU WANT	18
34	DON T KNOW	26	84	THAT S WHAT	18
35	IF YOU NEED	26	85	TO THE LEFT	18
36	A LITTLE BIT	25	86	WE LL CALL	18
37	BE ABLE TO	25	87	AND I LL	17
38	LET YOU KNOW	25	88	NEED YOU TO	17
39	THAT S FINE	25	89	OKAY WE LL	17
40	UH WE LL	25	90	SO WE RE	17
41	BACK TO THE	24	91	THANK YOU AND	17
42	**OFF THE RUNWAY**	24	92	WANT US TO	17
43	OKAY WE RE	24	93	YOU KNOW WHAT	17
44	WE VE GOT	24	94	APPEARS TO BE	16
45	WE DON T	23	95	I NEED TO	16
46	WE RE GOING	23	96	KNOW IF YOU	16
47	YOU HAVE THE	23	97	ON THE GROUND	16
48	ARE YOU ABLE	22	98	RE GONNA BE	16
49	**HOLD YOUR POSITION**	22	99	RE READY TO	16
50	**OF THE RUNWAY**	22	100	SIR WE RE	16

Some interesting findings emerge from this list, the first to call attention being the high presence of modal verbs and personal pronouns, two of the language items that according to ICAO (2007) must not be employed in radio communications, but are also common in general English spoken corpora (MCCARTHY; CARTER, 2002). An analysis of each of these clusters first in concordance lines and then in the text they are taken from show that the modal verbs function as mitigators (CAFFI, 1999, p. 882), that is, features related to the management of the interaction that weaken risks such as "self-contradiction, refusal, losing face, conflict, and so forth". Given the problem-solving purpose that oriented this corpus compilation, pilots and controllers seem to attenuate their speech acts, which also change in this scenario as, for example, controllers start to offer alternatives rather than stating commands. The following extract illustrates the mitigation identified in the expression "would/'d like to".

Extract 1:

ATCO	Aircraft seven thirty-six / roger the pan pan / are you ready for the turn here for me? //
Pilot	Uh **we'd uh we'd like to** solve up the problem and **we'd like uh to** return into Sydney / it's better //
ATCO	Aircraft seven thirty-six / **would you like to** return now? //
Pilot	Uh affirm //
ATCO	Aircraft seven thirty-six / turn left heading two one zero / maintain five thousand feet //
Pilot	Left turn heading two one zero / maintaining five thousand / Aircraft seven seven thirty-six //
ATCO	**Would you like to** hold somewhere or are you ready to land now? //
Pilot	We'll keep you advised and tell you later / okay? //
ATCO	Aircraft seven thirty-six / roger / **if you'd like** to hold / what place **would you like to** hold at? //
Pilot	Uuh / we'll advise to you later / we are trying to solve up the problem and we are now <break> engine number one is on idle power / and we are uh <pause> and uuh determing [sic] whether to dump some fuel or uh just check the performance / okay? //

This extract is from a communication about an aircraft that suffered an engine failure after take-off, with the pilots deciding to return to the airport of origin. The extract starts from the pilots saying that they need to work on the problem by means of checklists *and* return to the airport at the same time. The controller rechecks this last information by using "would you like (to return now)." Following the confirmation, the controller gives instructions to enable the pilots to return to the airport, followed by the pilot's readback. However, the controller is still unsure as to whether the pilots need to fly over an area (hold) to prepare the aircraft for landing or if they are ready to land, and thus uses the cluster "would you like to" once again. The pilots use two pieces of information that show they are still not clear as to what their next step should be and when they should take it ("We'll keep you advised and tell you later / okay?"). The controller then asks a question in order to prepare for the next possible action: "if you'd like to hold / what place would you like to hold at?" The pilots finally state their current condition: they are checking their weight and limitations to decide whether or not they will need to dump fuel to reduce weight for landing.

The recurrent use of the expression "would like to" exemplified in this last extract suggests that when pilots and controllers are dealing with an emergency such as engine failure, they tend to mitigate their

language as a means of sharing responsibility over the problem. This can also be seen in the following extract.

Extract 2:

Pilot	Mayday <unreadable> zero five <unreadable> fire on board / **request** immediate turn back to Budapest //
ATCO	Roger / two stations / say again your call sign //
Pilot	Aircraft one six nine five / mayday / **request** turn back and descend to Budapest //
ATCO	Aircraft one six nine five uh roger uh / right is approved / descend to flight level two zero zero //
Pilot	Descend flight level two zero zero / Aircraft one six nine five // Aircraft one six nine five / **can we** turn back to Budapest? //
ATCO	Aircraft one six nine five / affirm / cleared to turn back to Budapest / right turn and uh descend to flight level two zero zero //
Pilot	Right turn flight level two zero zero / Aircraft one six nine five //

The pilot declares an emergency ("mayday") due to fire on board, one of the most critical problems a flight crew can experience. Adhering to Standard Phraseology, the pilot uses the word "request." The controller replies with "roger," a word that means "acknowledged" (but not an affirmative response), states that two radios were in use at the same time ("two stations"), thus blocking the radio frequency, and requests repetition of the call sign (flight number). However, the controller does not refer to which aircraft his request was addressed to. The pilot of the aircraft in the emergency repeats the call sign, the emergency status ("mayday"), and the request ("request turn back and descend to Budapest"). The controller replies once again with "roger," this time also acknowledging the call sign, and gives instructions. However, as the controller does not give any indication that he is complying with the pilot's request, the pilot switches to the use of the mitigation device "can we turn back to Budapest?" The controller finally uses the proper Standard Phraseology to signal to the pilot that they are working together ("affirm / cleared to turn back to Budapest").

The use of "can we" to emphasize the request in an emergency situation reinforces the idea exposed earlier that when involved with a problem, the participants in the interaction under study here migrate to more spontaneous – and mitigated – language. It is worth noting that although "can we" is not a three-word cluster, it was investigated along with the cluster "if you can" (with 24 occurrences), also commonly used for requests.

The second element to be addressed is the high frequency of personal pronouns and referential words such as *here* and *there*, or deixis. Deixis are "aspects of language whose interpretation is relative to the occasion of utterance" (FILLMORE, 1966, p. 220), that is, items that anchor the elements expressed by the participants to the context of production. These can be demonstrative pronouns, personal pronouns, adverbs of place and time, verb tenses, or even verbs such as *come* and *go* (LEVINSON, 2004, p. 74). Such elements, which are highly dependent on the context of production and particularly on the location of the participants in the interaction, should not be used according to Standard Phraseology as precision is a key element in radio communications. However, a further analysis of deictics within their clusters corroborate an investigation by Garcia (2016) of the communication held in the accident of the Airbus 320 that landed on the Hudson River in New York. Through Conversational Analysis, Garcia showed that the portion of language used to describe this unusual event was signaled by linguistic items such as hesitation markers, deixis, and *okay* as a turn opener. Let us observe the transcript of another communication in the following extract.

Extract 3:

ATCO	<interrupted> one two thousand / maintain two five zero knots //
Pilot	Descend to one two thousand and maintain two five zero knots / Aircraft one ninety-two heavy **// and Chicago / just confirm Aircraft one ninety-two heavy / we are cleared down to one two thousand feet / two five zero knots? //**
ATCO	Aircraft one ninety-two heavy / are you declaring an emergency **or just** need to return back as a precaution? //
Pilot	**Uh** just a precautionary return at the moment / we're gonna have the aircraft inspected as it was uh a fairly large flock of birds that made a mess on the front windshield and we're worried about the radome //
ATCO	**Okay** / roger //
Pilot	We are down to one two thousand / **was that last clearance for one ninety-two heavy?** //
ATCO	Aircraft one ninety-two heavy / **uh** affirmative / descend and maintain one two thousand and uh maintain two five zero knots //
Pilot	Down to one two thousand / two five zero knots / Aircraft one ninety-two heavy //

The parts in bold correspond to the elements listed by Garcia (2016). They signal the transition from the rehearsed language of Standard Phraseology (starting from the second turn) to the Plain Aviation English used when the interactions are built around the problem (turns 2-5). These elements, which signal a transition, are usually repetitive, which

can be seen through the investigation of clusters of elements such as "uh we're," "and uh we," "uh we're," and "okay we're" but also functional words such as personal pronouns, hesitation markers, conjunctions (*but, so, and*) and prepositions, confirming Garcia's finding.

Another feature of this discourse is that these elements are related to the organization of the discourse, as can be seen in Table 5.

TABLE 5 – Sample of concordance lines with "and uh we"

N	Concordande
1	tihad four five one // understood // **and uh we'** ll give you five minutes' notice
2	of course all over the windscreen / **and uh we'** caught one of them on one of the
3	tors / fly heading zero niner zero / **and uh we'** ll expect runway two eight center
4	're at the process of slowing down / **and uh we'** ll call the base circuit at one e
5	op on the runway for an inspection / **and uh we'** re gonna evaluate the situation t
6	tially was fire / there is no fire / **and uh we'** are waiting for your notification
7	s // They've got the longer runway / **and uh we'** re gonna get you uh the most uh a
8	/ we're having landing gear issues / **and uh we'** need to sort it out / we're gonna
9	the fuel remaining in pounds? // Uh **and uh we'** have it in kilos // Alright / wha
10	re just starting the checklist now / **and uh we'** try just to uh if that's the whol

Table 5 shows a sample (10 out of 40 occurrences) of the concordance lines with "and uh we" in the center, starting prosodic islands (lines 2-8) or utterances (lines 1 and 9). This implies that the cluster "and uh we" may function as a turn opener or as a strategy for holding the turn as silence may indicate that the other participant can press the radio button to speak. Again, let us turn to the text of production of one of the lines.

Extract 4:

ATCO	Aircraft four nineteen heavy / when you have a chance **just uh** call me back on this frequency please / it's from the Tower //
Pilot	Yup / okay **uh so we** have a chance to call you back **so** / what do you need? //
ATCO	I just need **uh you** you're coming out of Dulles so what's your destination / Aircraft four nineteen heavy? / and your registration number please //
Pilot	Okay / the registration is delta alpha bravo yankee tango / **and the uh** destination was Frankfurt echo delta delta foxtrot //
ATCO	Aircraft four nineteen heavy / thank you very much / have a good night //
Pilot	Thank you **and uh we** uh how do we get onto the position? / we have no signals / nothing is there / so how do we get in? // is there a follow me? //

In this extract, the hesitation marker is also bolded in other clusters. The cluster "and uh we" seen in the last turn is followed by another hesitation marker and then a correction, showing that it was used as a false starter. The other clusters – "uh so we" and "and the uh" – also confirm their function as discourse organizers (starting and holding the turn, respectively). As in Tao (2003), which used CL to support Conversation Analysis studies of repeated language functioning as turn openers, in radio communications, certain clusters seem to function as discourse organizers, supporting Mell's finding (2004) (see Section 2). Other clusters employed here as discourse organizers are: *uh we're* (44 occurrences), *and uh we* (40 occurrences), *I'm gonna* (35 occurrences), *and we'll* (33 occurrences), *you're gonna* (30 occurrences), *uh we are* (27 occurrences), *uh we'll* (25 occurrences), and *okay we're* (24 occurrences), among others.

Specific speech acts also compose this specific genre (cf. BHATIA, 1993). Doc 9838 (ICAO, 2010) offers a list of communicative functions drawn from Mell's 1991 (see Section 2). As Mell's study was based on radio communications in routine situations, his roll of functions is more extensive than those examined in this study (see TABLE 7). Moreover, many of the functions described by Mell are manifested in RTPEC. To illustrate this point, controllers are in charge of giving information about weather, traffic, and airport status but mostly of giving directions, which are transmitted with verbs in the imperative form (such as *climb*, *descend*, etc). Sometimes, as predicted by Standard Phraseology, controllers request pilots to provide some information, such as "report when ready to copy" or "report when reaching flight level 230". However, RTPEC shows that, in abnormal situations, clusters such as "let me know" and "do you have" are more commonly used. This could be related to the mitigation described previously, as can be observed in the concordance lines with "let me know" in the center (TABLE 6).

TABLE 6 – Sample of concordance lines with "let me know"

N	Concordande		
1	wo two left approach and	**let me know**	if you need anything di
2	ou // Romeo Oscar Mike /	**let me know**	if you want lower than
3	lot one five one heavy /	**let me know**	if you get the age of t
4	orth of alpha? // Okay /	**let me know**	if that changes i'll ke
5	two two left localizer /	**let me know**	if you need me to put y
6	y // it's no big deal //	**let me know**	if you need any more as
7	a minute // Okay / just	**let me know**	if you can maneuver all
8	ck to you here // Okay /	**let me know**	if you need anything //
9	n as we are // could you	**let me know**	if there's any change t
10	can sixteen forty / just	**let me know**	if you need any lower t

These concordance lines exhibit some common collocates such as "okay," "if," and "you." When analyzing the concordance lines in the transcript, we find the following extract.

Extract 5:

ATCO	Aircraft one eighty / Kennedy Ground / continue via fox bravo / hold short of runway two two right / remain this frequency //
Pilot	Okay / can we hold on just a second / we need to run a few checklists for one eighty //
ATCO	Aircraft one eighty / roger / **let me know when you're ready to taxi** //

The cluster "let me know" is used when pilot and controller migrate to a more spontaneous discourse implying an undeclared problem (Turn 2). Even though the pilot does not say what the problem is, by stating that he needs to run checklists, he suggests that he has a technical situation to handle. This switch to a more spontaneous code appears to confirm the need for mitigation in radio communications when the abnormal happens.

The linguistic areas of fluency and interaction specified in the ICAO Language Proficiency Rating Scale seem to lack elements such as mitigation or even the signals that call for the collaboration of the participants. It is then possible to confirm that the perspective adopted needs to be updated with more current research from applied linguistics, in this case with recent studies in Pragmatics. Being concerned with language in use, Pragmatics – or pragmatic awareness – should be a useful path for guiding teachers preparing activities or designing materials.

Aeronautical English teachers should also consider the spoken grammar that organizes the conversation between pilots and controllers.

To inform curriculum or pedagogical activities, all 100 three-word clusters in the corpus were analyzed within their context of production and grouped according to their functional profile (ADOLPHS, 2008). This can relate to pragmatic awareness, assisting teachers or material designers with the development of activities. The result is as follows (TABLE 7).

TABLE 7 – Clusters distributed according to their functions

Functions / Speech acts	**Clusters**	**Functions / Speech acts**	**Clusters**
Request	*we'd like (something* or *someone)* *can you* *we need to* *you need to* *if you can*	**Mitigators**	*you know what* *a little bit* (usually before mentioning the problem) *we / I need you to* (more assertive – for instructions) *I don't know if* (for offers or requests)
Request and provide information	*(just) let sb know…* *…when you get a chance* *…when you get to / on* (place) *…when you're ready* *…if / when you have a moment / second / chance* *as soon as * can / possible / practicable* *we'll call you back*	**State abilities or ask about abilities**	*(modal verb) be able to* *If you can* *are you able to*
Offer	*would you like (to)* *do you need (any)* *do you want* *if you need* *if you want* *if you'd like* *if you can* *we / I'll give you* *do you need* *we'll get* *give you a* *do you want us to* *can you*	**Open or hold the turn**	*and uh we* *uh we have* *uh do you* *and we'll / and I'll* *sir we're* *okay we're* *so we're* *uh we'll* *uh we are / we're* *that's what I / we*

State decisions	*we're gonna* *we'd like to* *I'm gonna* *gonna have to* *re going to* *we'll be* *would like to*	**Report other's instructions / decisions made previously**	*that's what*
Agree / allow / thank	*okay thank you* *roger thank you* *okay we're* *that's fine*	**Highlight the current moment**	*right now / now* *at this time* *at the moment* *momentarily* (in the meaning of "soon") *immediately*
Inform of the problem	*don't have* *we have a* *you have a* *we've got* *we don't* *we have uh* *it's a* *appear to be*	**Request information about a problem**	*do you have* *uh do you* *you have the* *can you tell me / give me*

Table 7 presents possible language that may assist teachers in developing syllabi, pedagogical materials, or activities and is not intended for memorization. Instead, the material designer or the teacher may consult this list when preparing activities targeted at Plain Aviation English, not at Standard Phraseology. This list can also inform ICAO Language Proficiency Rating Scale users when assessing a candidate's Plain Aviation English, for example, or when designing proficiency assessment tasks.

Understanding how users organize their discourse can assist learners as well as professional users such as aviators and controllers in signaling problems or abnormal situations or even in comprehending when to remain silent so that other crew with problems can manage their problem with the controller without interference. It also helps pilots recognize when it is time to press the button to take the turn on congested radio frequencies. Students, who are pilots in service, often report that it takes them long minutes before they finally take the turn at airports such as JFK. Another benefit regards the frontiers in the transition between Standard Phraseology and Plain Aviation English. Such an example can also be found in the corpus, as presented in the following extract.

Extract 6:

Pilot	Kennedy Tower / Aircraft eight zero eight zero / reporting balloon / final four right //
ATCO	Say again? //
Pilot	Eight zero eight two heavy / **reporting balloon / four right** //
ATCO	Aircraft eight zero eight two heavy / I'm having trouble understanding you / you are cleared to land four right / can you say again / please speak up //
Pilot	Okay / no problem // cleared to land four right / Aircraft eight zero eight two / **reporting hot balloon** uh final runway four right about five hundred feet //
ATCO	Reporting a bird? / Is that what you're saying? // tell me when you get on the ground //
Pilot	Okay //
ATCO	The wind is three two zero at one zero // eight zero eight two heavy / turn left on foxtrot bravo // Did you have windshear / is that what you are saying? //
Pilot	No / leaving on fox bravo / Aircraft eight zero eight two / **reporting hot air balloon** on final four right about five hundred feet //
ATCO	Balloon / you said? //
Pilot	Balloon //

Here, the Brazilian pilot reports a balloon (an uncommon hazard) in the surroundings of the airport. However, the pilot does not signal the transition from Standard Phraseology to Plain Aviation English to describe the uncommon hazard. By using the word "reporting," the pilot not only uses a word normally considered inappropriate to his participation in this interaction, but he also makes up a collocation that is unfamiliar to the interlocutor. The controller repeatedly tries to understand the pilot, but senses that this situation is not urgent and instructs the pilot to proceed with the landing.

When searching for the word "balloon" in the concordance lines, we find the following instances (TABLE 8).

TABLE 8 – Concordance lines with "balloon"

N	Concordande		
1	void a balloon / we have a	**balloon**	right now on our right ha
2	ow turning left to avoid a	**balloon**	/ we have a balloon right
3	e seven / disconnecting /	**balloon**	now on the right / uh sir
4	e seven / there is another	**balloon**	at uh UTBUR at uh flight
5	way just to go around the	**balloon**	// Roger / report back on
6	blished // We uh we got a	**balloon**	right in the way / we'd l

7	sir // we have flown by a	**balloon**	right at this time // Am
8	tion uh there is a hot air	**balloon**	final on final approach /
9	we are we have a hot air	**balloon**	on flight level one hundr
10	e need information about	**balloon**	// Aircraft zero five sev
11	on uh hot to avoid hot air	**balloon**	on approach / uh many bal
12	above us / there´s a free	**balloon**	flying with a photo // Sw
13	scending to avoid a a free	**balloon**	uh flying around here / d

The concordance lines reveal a common surrounding of the word balloon consisting of "there is", "I can see," and "we are deviating from". These strings reiterate that Plain Aviation English resembles simpler colloquial structures. However, a closer look at the context of production shows that despite the fact that all these lines come from the same radio communication, they are all enunciated by different international aircraft, one of them from a country (the United States) where English is the official language. By exposing students to this fact, teachers may address attitudes toward transfer (in Brazil, reports such as the one described here are made through the Portuguese word *reportar* (report) and the use of Plain Aviation English.

6 Conclusion

The objective of this paper was to identify what linguistic elements constitute fluency and interaction in the Plain Aviation English of air-ground radio communications in abnormal situations. These two linguistic areas were taken from the ICAO Language Proficiency Rating Scale and compared against a corpus built with a two-fold purpose: researching Plain Aviation English and informing pedagogical materials. Corpus Linguistics was shown to be useful in identifying patterns in the Plain Aviation English used in radio communications. Generating cluster lists enabled the analysis of the pragmatic functions of the clusters, identified as pragmatic routines and as items that assist in organizing the conversation. However, these conclusions were not drawn from mere frequencies. Instead, each cluster was examined one by one in concordance lines and in the transcripts, where information about the source was also displayed. These clusters were then grouped into a total of 12 functions, verifying that fluency and interaction can be interconnected into the broader perspective of Pragmatics.

The paper showed that the transition between Standard Phraseology and Plain Aviation English can be signaled by certain linguistic elements such as deixis and hesitation, which are also present in the organization of the conversation between pilots and controllers when sharing responsibility for a problem. This corresponds to the spoken grammar suggested by Rühlemann (2008) and can be pedagogically presented through activities that ask students to reflect on the co-construction of the conversation while also considering key components that act on the pilot-controller relationship.

The problem-solving objective that was inherent in the corpus design allowed for the observation of how the participants in the interaction share responsibility for making decisions and solving problems. Participants mitigate their language toward the same goal by engaging in a verbal-only communication, even in a hermetic context such as aviation, thus supporting Lopez' (2013) claim that social conventions in radio communications cannot be controlled.

Based on the analysis presented in this paper, what should be taken into account is a discussion of pedagogical concepts regarding the content dealt with in the Aeronautical English classroom. In response, a more appropriate syllabus should emphasize the development of the students' capacity of interacting while considering pragmatic awareness and cultural tolerance (DAVIES, 2004). Such a syllabus should consider fluency and interaction as a co-construction by at least two participants in the interaction. It should also contemplate activities that allow the student to enhance pragmatic strategies and pragmatic routines (ISHIHARA; COHEN, 2010). The analysis presented here points to teaching oriented by language use, the co-construction of the interaction among the participants, the context that regulates the community, and pragmatic awareness enhanced in such a way that it allows students to choose how to position themselves in their own community so that they can better perform their functions.

References

ADOLPHS, S. *Corpus and Context:* Investigating Pragmatic Functions in Spoken Discourse. Amsterdam: John Benjamins, 2008. DOI: https://doi.org/10.1075/scl.30.

ALTENBERG, B. On the Phraseology of Spoken English: The Evidence of Recurrent Word Combinations. *In:* COWIE, A. (org.) *Phraseology:* Theory, Analysis, and Applications. Oxford: Oxford University Press, 1998. p. 101-122.

BARDOVI-HARLIG, K. Formulas, Routines, and Conventional Expressions in Pragmatics Research. *Annual Review of Applied Linguistics,* Cambridge, v. 32, p. 206-227, 2012. DOI: https://doi.org/10.1017/S0267190512000086.

BHATIA, V. K. *Analyzing Genre:* Language Use in Professional Settings. London: Longman, 1993.

BIESWANGER, M. Applied Linguistics and Air Traffic Control: Focus on Language Awareness and Intercultural Communication. *In:* HANSEN-SCHIRRA, S.; MAKSYMSKI, K. (org.). *Aviation Communication:* Between Theory and Practice. Frankfurt-am-Main: Peter Lang, 2013. p. 15-31.

BIESWANGER, M. Aviation English: Two Distinct Specialized Registers? *In:* SCHUBERT, C.; SÁNCHEZ-STOCKHAMMER, C. (org.). *Variational Text Linguistics:* Revisiting Register in English. Berlin: de Gruyter, 2016. p. 67-85.

BOROWSKA, A. *Avialinguistics:* The Study of Language for Aviation Purposes. Frankfurt-am-Main: Peter Lang, 2017. DOI: https://doi.org/10.3726/b11037.

CAFFI, C. On Mitigation. *Journal of Pragmatics*, [*S.l.*], v. 31, n. 7, p. 881-909, 1999. DOI: https://doi.org/10.1016/S0378-2166(98)00098-8.

CARTER, R; MCCARTHY, M. *Cambridge Grammar of English*: A Comprehensive Guide. Cambridge: Cambridge University Press, 2006.

CARTER, R.; MCCARTHY, M. Spoken Grammar: Where Are We and Where Are We Going? *Applied Linguistics*, Oxford, v. 38, n. 1, p. 1-20, 2017. DOI: https://doi.org/10.1093/applin/amu080.

CHENG, W. What Can a Corpus Tell Us about Language Teaching? *In:* O'KEEFFE, A.; MCCARTHY, M. (org.). *The Routledge Handbook of Corpus Linguistics*. London; New York: Taylor & Francis Group, 2015. p. 319-332. DOI: https://doi.org/10.4324/9780203856949-23.

CORBETT, J. *An Intercultural Approach to English Language Teaching*. Clevedon: Multilingual Matters, 2003. DOI: https://doi.org/10.21832/9781853596858.

CRESTI, E. *Corpus di italiano parlato*. Florence: Accademia della Crusca, 2000.

CRESTI, E. Syntactic Properties of Spontaneous Speech in the Language into Act Theory. *In:* RASO, T.; MELLO, H. (org.). *Spoken Corpora and Linguistic Studies*. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2014. p. 365-410. DOI: https://doi.org/10.1075/scl.61.13cre.

DAVIES, C. E. Developing Awareness of Crosscultural Pragmatics: The Case of American/German Sociable Interaction. *Multilingua,* [*S.l.*], v. 23, n. 3, p. 207-231, 2004. DOI: https://doi.org/10.1515/mult.2004.010.

DOUGLAS, D. *Assessing Language for Specific Purposes*. Cambridge: Cambridge University Press, 1999. DOI: https://doi.org/10.1017/CBO9780511732911

DUBOIS, J. W. Transcription Design Principles for Spoken Discourse Research. *Pragmatics,* [*S.l.*], v. 1, n. 1, p. 71-106, 1991. DOI: https://doi.org/10.1075/prag.1.1.04boi.

EMERY, H. Developments in LSP Testing 30 Years on: The Case of Aviation English. *Language Assessment Quarterly,* [*S.l.*], v. 11, n. 2, p. 198-215, 2014. DOI: https://doi.org/10.1080/15434303.2014.894516.

ESTIVAL, D.; FARRIS, C; MOLESWORTH, B. *Aviation English:* A lingua franca for pilots and air traffic controllers. London: Routledge, 2016. DOI: https://doi.org/10.4324/9781315661179.

FIELD, J. *Listening in the Language Classroom*. Cambridge: Cambridge University Press, 2009. DOI: https://doi.org/10.1017/CBO9780511575945.

FILLMORE, C. Deictic Categories in the Semantics of "Come." *Foundations of Language,* [*S.l.*], v. 2, n. 3, p. 219-227, 1966.

FRIGINAL, E.; MATHEWS, E.; ROBERTS, J. *English in Global Aviation:* Context, Research, and Pedagogy. London: Bloomsbury, 2020.

GARCIA, A. Air Traffic Communications in Routine and Emergency Contexts: A Case Study of Flight 1549 "Miracle on the Hudson." *Journal of Pragmatics,* [*S.l.*], v. 106, p. 57-71, 2016. DOI: https://doi.org/10.1016/j.pragma.2016.10.005.

GARCIA, A. C. *What do ICAO Language Proficiency Test Developers and Raters Have to Say about the ICAO Language Proficiency Requirements 12 Years after their Publication?* 2015. 115f. Thesis (Masters in Language Testing) – Department of Linguistics and English Language, Lancaster University, Lancaster, 2015

GARCIA, A. C.; FOX, J. Contexts and Constructs: Implications for the Testing of Listening in Pilots' Communications with Air Traffic Controllers. *The Especialist*, São Paulo, v. 41, n. 4, p. 1-33, 2020. DOI: 2318-7115.2020v41i4a4. Available from: https://revistas.pucsp.br/index.php/esp/article/view/49775. Access on: Nov. 15, 2020.

GARDNER, I. A.; DETERDING, D. Pronunciation and Miscommunication in ELF Interactions: An Analysis of Initial Clusters. *In:* JENKINS, J.; BAKER, W.; DEWEY, M. (org.). *The Routledge Handbook of English as a Lingua Franca.* London: Routledge, 2018. p. 224-232. DOI: https://doi.org/10.4324/9781315717173-19.

GÖTZ, S. *Fluency in Native and Nonnative English Speech.* Amsterdam: John Benjamins, 2013.

GRANGER, S. Learner Corpora in Foreign Language Education. *In:* VAN DEUSENSCHOLL, N.; HORNBERGER, N. H. (org.). *Encyclopedia of Language and Education.* Philadelphia: Springer, 2008. v. 4, p. 337-351. DOI: 10.1007/978-3-319-02328-1_33-1.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Manual of implementation of the language proficiency requirements*. Montreal: ICAO, 2004.

INTERNATIONAL CIVIL AVIATION ORGANIZATION. *Aviation Occurrence Categories:* Definitions and Usage Notes. Montreal: ICAO, 2006.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Manual of radiotelephony*. Montreal: ICAO, 2007.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Manual of implementation of the language proficiency requirements* (2nd. ed.). Montreal: ICAO, 2010.

ISHIHARA, N.; COHEN, A. D. *Teaching and Learning Pragmatics:* Where Language and Culture Meet. Edinburgh: Pearson Education Limited, 2010.

ISHIHARA, N.; PRADO, M. The Negotiation of Meaning in Aviation English as a Lingua Franca: A Corpus-Informed Discursive Approach. *Modern Language Journal*, [*S.l.*]. in press.

JENKINS, J. *The Phonology of English as an International Language.* Oxford: Oxford University Press, 2000.

KAUR, J. Communication Strategies in English as a Lingua Franca Interaction. *In:* PETERS, M. A.; HERAUD, R. (org.). *Encyclopedia of Educational Innovations*. Singapore: Springer, 2019. p. 1-5. DOI: https://doi.org/10.1007/978-981-13-2262-4_86-1.

KIM, H. What Constitutes Professional Communication in Aviation: Is Language Proficiency Enough for Testing Purposes? *Language Testing,* [*S.l.*], v. 35, n. 3, p. 403-426, 2018. DOI: https://doi.org/10.1177/0265532218758127.

KNOCH, U. Using Subject Specialists to Validate an ESP Rating Scale: The Case of the International Civil Aviation Organization (ICAO) Rating Scale. *English for Specific Purposes,* [*S.l.*], v. 33, p. 77-86, 2014. DOI: https://doi.org/10.1016/j.esp.2013.08.002.

LEVINSON, S. Deixis. *In:* HORN, L. R.; WARD, G. (org.). *The Blackwell Handbook of Pragmatics*. Malden: Blackwell, 2004. p. 97-121. DOI: https://doi.org/10.1002/9780470756959.ch5.

LEWIS, M. *The Lexical Approach:* The State of ELT and a Way Forward. London: Language Teaching Publications, 1993.

LÓPEZ, S. *Norme(s) et usage(s) langagiers*: Le cas des communications pilote-contrôleur en anglaise. 2013. 420f. Tese (Doutorado em Linguística) – Université de Toulouse Le Mirail, Toulouse, 2013.

MATHEWS, E. *Language Gap*. Alexandria: Flight Safety Foundation, 2012. Available from: https://flightsafety.org/asw-article/language-gap. Access on: Sep. 30, 2020.

MATHEWS, E. How Incomplete Language Standards Threaten Aviation. *Aviation Week*, [*S.l.*], [s.p.], 2020. Available from: https://aviationweek.com/air-transport/opinion-how-incomplete-language-standards-threaten-aviation?fbclid=IwAR262pq8-6McH6YkH7_1PvpsXOAHI0cogGfANrsd7TbU0N4pufWxqeDPSZI. Access on Sep. 30, 2020.

MAURANEN, A. Conceptualising ELF. *In*: JENKINS, J.; BAKER, W.; DEWEY, M. (org.). *The Routledge handbook of English as a lingua franca*. London: Routledge, 2018. P. 7-24. DOI: https://doi.org/10.4324/9781315717173-2.

MCCARTHY, M.; CARTER, R. This, That, and the Other: Multi-Word Clusters in Spoken English as Visible Patterns of Interaction. *Teanga: Irish Yearbook of Applied Linguistics*, Dublin, v. 21, p. 30-52, 2002.. DOI: https://doi.org/10.35903/teanga.v21i0.173.

MCCARTHY, M.; CLANCY, B. From Language as System to Language as Discourse. *In:* WALSH, S.; MANN, S. (org.). *The Routledge Handbook of English Language Teacher Education.* London: Routledge, 2018. p. 199-215. DOI: https://doi.org/10.4324/9781315659824-15.

MCNAMARA, T. Managing Learning: Authority and Language Assessment. *Language Teaching,* Cambridge, v. 44, n. 4, p. 500-515, 2011. DOI: https://doi.org/10.1017/S0261444811000073.

MELL, J. Language Training and Testing in Aviation Needs to Focus on Job-Specific Competencies. *ICAO Journal,* [*S.l.*], v. 59, n. 1, p. 12-14, 2004.

MODER, C.; HALLECK, G. Planes, Politics and Oral Proficiency: Testing International Air Traffic Controllers. *Australian Review of Applied Linguistics*, v. 32, n. 3, p. 25.1-25.16, 2009. DOI: https://doi.org/10.1075/aral.32.3.05mod.

MONTEIRO, A. L. *Reconsidering the Measurement of Proficiency in Pilot and Air Traffic Controller Radiotelephony Communication:* From Construct Definition to Task Design. 2019. 475f. Tese (Doutorado em Linguística Aplicada) – Faculty of Graduate and Postdoctoral Affairs, Carleton University, Ottawa, Ontario, 2019. Available from: https://

curve.carleton.ca/0b65cc09-37d2-449f-804d-a6f804917927. Accessed on Sep. 5, 2020.

MORROW, D.; RODVOLD, M.; LEE, A. Nonroutine Transactions in Controller-Pilot Communication. *Discourse Processes,* [*S.l.*], v. 17, p. 235-258, 1994. DOI: https://doi.org/10.1080/01638539409544868.

O'KEEFE, A.; CLANCY, B.; ADOLPHS, S. *Introducing Pragmatics in Use*. London: Routledge, 2011.

PFEIFFER, A. Inter-Rater Reliability in an Aviation Speaking Test. 2009. 64f. Dissertation (Masters in Linguistics) – Faculty of Linguistics and English Language, Lancaster University, Lancaster, 2009.

PRADO, M. *A relevância da Pragmática no ensino do inglês aeronáutico*: um estudo baseado em corpora. 2019. 336p. Tese (Doutorado em Letras) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, 2019.

PRADO, M.; TOSQUI-LUCKS, P. Designing the Radiotelephony Plain English Corpus (RTPEC): A Specialized Spoken English Language Corpus Towards a Description of Aeronautical Communications in Non-Routine Situations. *Research in Corpus Linguistics*, [*S.l*], v. 7, p. 113-128, 2019. DOI: https://doi.org/10.32714/ricl.07.06.

RÜHLEMANN, C. A Register Approach to Teaching Conversation: Farewell to Standard English? *Applied Linguistics,* Oxford, v. 29, n. 4, p. 672-693, 2008. DOI: https://doi.org/10.1093/applin/amn023.

SACKS, H.; SCHEGLOFF, E.; JEFFERSON, G. A simplest systematics for the organization of turn-taking for conversation. *Language,* Washington, DC, v. 50, n. 4, p. 696-735, 1974. DOI: https://doi.org/10.1353/lan.1974.0010.

SCOTT, M. *Wordsmith Tools* (Version 7). Stroud: Lexical Analysis Software, 2016.

SINCLAIR, J. *Corpus, Concordance, Collocation:* Describing English Language. Oxford: Oxford University Press, 1991.

SINCLAIR, J.; M. COULTHARD. *Towards an Analysis of Discourse.* Oxford: Oxford University Press, 1975.

SVARTVIK, J. (org.). *The London-Lund Corpus of Spoken English:* Description and Research. Lund: Lund University Press, 1990.

TAGNIN, S. *O jeito que a gente diz:* expressões convencionais e idiomáticas. São Paulo: Disal, 2013.

TAO, H. Turn Initiators in Spoken English: A Corpus-Based Approach to Interaction and Grammar. *In:* LEISTYNA, P.; MEYER, C. (org.). *Corpus Analysis:* Language Structure and Language Use. Amsterdam: Rodopi, 2003. p. 187-207. DOI: https://doi.org/10.1163/9789004334410_011.

TOSQUI-LUCKS, P.; SILVA, A. L. Aeronautical English: Investigating the Nature of this Specific Language in Search of New Heights. *The Especialist*, São Paulo, v. 41, n. 3, p. 1-27, 2020. DOI: https://doi.org/10.23925/2318-7115.2020v41i3a2. Available from: https://revistas.pucsp.br/index.php/esp/article/view/47826. Access on: Nov. 15, 2020.

TRIPPE, J.; BAESE-BERK, M. A prosodic profile of American aviation English. *English for Specific Purposes,* [*S.l.*], v. 53, p. 30-46, 2019. DOI: https://doi.org/10.1016/j.esp.2018.08.006.

WEIR, A. *The Tombstone Imperative.* London: Simon; Schuster, 1999.

WEISSER, M. *How to Do Corpus Pragmatics on Pragmatically Annotated Data.* Amsterdam: John Benjamins, 2018. DOI: https://doi.org/10.1075/scl.84.

WIDDOWSON, H. G. Context, Community, and Authentic Language. *TESOL Quarterly,* [*S.l.*], v. 32, n. 4, p. 705-716, 1998. DOI: https://doi.org/10.2307/3588001.

YOUNG, R. Interactional Competence: Challenges for Validity. *In*: ANNUAL MEETING OF THE AMERICAN ASSOCIATION FOR APPLIED LINGUISTICS*,* 2000. Vancouver. *Proceedings* [...]. Vancouver: ERIC Clearinghouse, 2000. p. 1-15.

ZANETTIN, F. Corpora multimediali e analisi dell'interazione: Osservazioni su strumenti e metodologie. *In:* GAVIOLI, L. (org.). *La mediazione linguistico culturale:* Una prospettiva interazionista. Perugia: Guerra Edizioni, 2009. p. 325-255.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1415-1442, 2021

Analyzing the use of personal pronouns in aeronautical communications through CORPAC (Corpus of Pilot and Air Traffic Controller Communications)

O uso de pronomes pessoais em comunicações aeronáuticas: uma análise através do CORPAC (Corpus of Pilot and Air Traffic Controller Communications)

Aline Pacheco
Pontifícia Universidade Católica do Rio Grande do Sul (PUCRS), Porto Alegre, Rio Grande do Sul / Brasil
aline.pacheco@pucrs.br
http://orcid.org/0000-0003-1638-0215

Abstract: This article aims to analyze the use of personal pronouns in aeronautical communications based on CORPAC, a specialized corpus. Pronouns can play an important role in multitasking communicative scenarios such as the one featured in aviation and therefore it is of paramount importance that identities be clearly set in operations. In light of Neville's (2004) study about cockpit's identities, this investigation addresses the frequency and patterns of usage of personal pronouns – especially I, we and you, using corpus linguistic tools. The corpus exploration provides evidence that such pronouns are indeed very frequently used, despite official orientations that do not recommend their use in order to avoid problems such as ambiguity. The examination reveals consistent and interpretable patterns associated to Neville's (2004) assumptions and has significant implications for training and testing purposes in the field of Aeronautical English.

Keywords: aeronautical communications; personal pronouns; corpus linguistics.

Resumo: Este artigo tem como objetivo analisar o uso de pronomes pessoais na comunicação aeronáutica a partir do CORPAC, um corpus especializado. Pronomes podem desempenhar um papel de destaque em cenários comunicativos multitarefa, tais como observados na aviação. Nesse sentido, faz-se importante que as identidades

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1415-1442

sejam claramente definidas nas operações. À luz do estudo de Neville (2004) sobre identidades no cockpit, esta investigação aborda a frequência e os padrões de uso de pronomes pessoais – especialmente "I", "we" e "you", por meio do uso de ferramentas linguísticas de corpus. A exploração do corpus fornece evidências de que tais pronomes são de fato usados com muita frequência, apesar de orientações oficiais que não recomendam seu uso, a fim de evitar problemas como a ambiguidade. A análise revela padrões consistentes e interpretáveis associados às suposições de Neville (2004) e tem implicações significativas para fins de treinamento e teste na área de Inglês Aeronáutico.

Palavras-chave: comunicação aeronáutica; pronomes pessoais; linguística de corpus.

Submitted on September 9th, 2020

Accepted on November 9th, 2020

1 Introduction

Communication is a critical human factor in aviation operations and the effects of poor communications are acknowledged to have highly impacted aviation safety (CUSHING, 1997; DIETRICH; MELTZER, 2002; MATHEWS, 2019; NEVILLE, 2004). Sexton and Helmreich (2000, p. 63) say that "The role of language has been neglected and researchers have recognized the need for a deeper understanding of its roles, characteristics and how it impacts in aviation." More recent research has shown that language specialists have been trying to widen the scope of studies in the field and have successfully managed to shed light on topics which need to be tackled. (SILVA; TOSQUI-LUCKS, 2020; PACHECO, 2019).

Corpus-based research on Aviation English (AE) has become of increasing interest as it enables the researcher to analyze real language occurrences from a variety of tools (BOCORNY, 2011; PRADO, 2019; SARMENTO, 2008; TOSQUI-LUCKS, 2018). It is known that the dialogues between pilots and air traffic controllers (ATCOs) are recorded and available from the Cockpit Voice Recorder (CVR) whenever there is the need for that and especially when there is an event with negative outcomes. Nevertheless, this material is not easily made available for research by airline companies or governmental institutions, and informal or non-authorized recordings can be a problem or can compromise data reliability.

Albeit the challenges posed by this methodology are particularly hard when it comes to corpus compilation in such a high-stakes domain as aviation, the results are very positive and the prospects quite promising. As an example, the International Civil Aviation English Association (ICAEA) has included Corpus Linguistics (CL) as one of the areas of study by its research group,[1] an initiative that will certainly contribute to spread the interest for studies that join together the language of aviation and the wide array of research possibilities offered by CL. We have known about the compilation of some corpora for aviation purposes, such as Corpus da Aviação (CAVI) (BOCORNY, 2008; SARMENTO, 2008), Radiotelephony Plain English Corpus (RTPEC) (PRADO, 2019), OSU Aviation Corpus (MODER, 2013) as well as other corpora mentioned in Lopez (2013), Swinehart (2013), Hinrich (2008).

In order to be able to analyze the language of aviation through a corpus, CORPAC (Corpus of Pilot and ATCO Communications) has been created. It is still in its preliminary stages of compilation and is totally based on open-access information from VASAVIATION,[2] which features emergency situations extracted from Live ATC.[3] The purpose is to explore the tools of analysis offered by CL and to be able to do research from real, spontaneous language use in aviation, covering a range of linguistic features.

It is known that pilots have to work cooperatively in highly coordinated activities throughout the stages of operations, both in the air and on the ground. To perform these tasks successfully, it is of paramount importance that identities be perceived as clearly as possible. In "Beyond the Black Box", Maurice Neville (2004) analyzes how pilots accomplish identities using pronominal forms. According to him, "Pronominal choices are an important aspect of pilots' habitual communicative practice contributing to their awareness of who is doing what and what is going on." (NEVILLE, 2004, p. 33). To coordinate their work, each pilot must be familiar with the duties and responsibilities associated with his/her own identities as well as with the tasks assigned to the other pilot. This will be linguistically performed, mostly, through the use of personal pronouns.

---

[1] https://www.icaea.aero/about/icaea-research-group/.

[2] https://www.youtube.com/channel/UCuedf_fJVrOppky5gl3U6QQ, a You Tube channel with available information.

[3] https://www.liveatc.net/, a paid service which offers access to aeronautical communications in aviation, live or recorded.

Based on Conversation Analysis and data collected in real flights, he explores the use of personal pronouns (subject and object) and possessive adjectives (for him, analyzed as a same category), mainly first person singular (I/me/my), second person (you/your) and first person plural (we/us/our). In his study, he describes "prescribed" and "non-prescribed" pronouns –the former referring to pronouns that are part of wordings spelled out for pilots in official operations manuals and the latter to those not part of the official wording expected to be used. The outcomes of this study show that pronominal choices are related to the creation and presentation of identities and relevant selves in order to comply with the tasks demands through coordinated teamwork.

In this line, this article proposes the investigation of personal pronouns by using CORPAC – a specialized corpus. It aims to analyze the use of some personal pronouns explored by Neville (2004) in aeronautical communications – namely, I, you and we, through tools used in CL research to check aspects regarding their frequency and their clusters. It starts by a brief review on the topics that most closely associate to the ideas approached in our study, such as language as a human factor in aviation, and some previous studies with the use of pronouns in aviation. Next, a section about the method precedes the description of the results obtained by our corpus research. The results are expected to add to the information presented by Neville (2004) and to offer relevant contribution to aeronautical English training, curriculum and test design regarding language in aviation.

2 Language as a factor in aviation communications

The International Civil Aviation Association (ICAO), the United Nations specialized agency for aviation, mandates, as of 2011, that all pilots flying in international airspace have a minimum operational English language proficiency. This is done by tests enforced by the Aviation Authority of each of its member states. Naturally, from this demand, there was an increase in interest about the language of aviation, being referred to as Aviation English, English for Aviation, Aeronautical English, Airpeak, AeroEnglish, Aeroese, Plane English, among others (BIESWANGER, 2016; BOROWSKA, 2017; ESTIVAL *et al.,* 2016; MODER, 2013; SILVA; TOSQUI-LUCKS, 2020).

Moder (2013, p. 227) seems to prefer the use of a more general concept: "Aviation English describes the language used by pilots, air traffic controllers, and other personnel associated with the aviation industry", using "radiotelephony" to comprehend the more specialized communication which occurs between pilots and ATCOs. According to her, AE is made of Phraseology – the prescribed vocabulary and syntax which are part of these highly specialized exchanges, and Plain English, the non-prescribed uses of more common English vocabulary and syntax.

Borowska (2017) understands that AE is an overarching term and that Aeronautical English is more suitable to designate communications solely between pilots and ATCOs (not with mechanic, flight attendants, dispatchers or aviation personnel in general), a distinctive concept which has also been adopted by Silva and Tosqui-Lucks (2020).

The effects of poor communication in aviation can be tragic. The accident of Tenerife, in 1977 is accounted for a miscommunication problem. The phrase "at take-off" ultimately triggered the crash. The KLM pilot uttered it meaning "taking off", using the structure from Dutch, his mother language, to designate continuous activity. The controller understood it as he was supposed to according to standard phraseology – referring to a specific place, waiting for an authorization. It is the deadliest crash in aviation history, killing 583 people. Cushing (1997), in Fatal Words – Communication Clashes and Aircraft Crashes, examines several aeronautical events that had communication issues as a factor, establishing categories of analysis such as problems of reference, inference, compliance; problems with numbers and radios. This manual is taken as a reference for studies of how language is involved and can impact aviation.

In line with this urge for more information about how language can impact safety in aviation, there is the LHUFT (Language as a Human Factor in Aviation) Center, at Embry-Riddle Aeronautical University. The center fosters research that aims to pinpoint linguistic factors involved in aviation communications, in order to have a better perspective of language in communication along with other human factors, so that it can be addressed more properly by the industry and other parts involved. "When accident investigators miss the more subtle effects of language use or language proficiency, the industry underestimates the possible impact and contributory effects of language problems in the accidents being investigated" (MATHEWS, 2019, p. 53). The LHUFT perspective

seeks to "a broad and more accurate understanding role of language in aviation safety, how language, language use, language proficiency, and culture affect aviation safety".[4]

Mathews, Pacheco and Albrighton (2019) discuss the factors that can account for miscommunication in aviation, based on a Taxonomy originally proposed by Mathews in 2013. The taxonomy considers technical, procedural, cultural and language factors in communication and has been very a valuable tool not only for proposing the analysis of specific language problems (as regards phonetics, syntax, semantics or pragmatics), but also for establishing an interface with other aspects that cannot be dissociated from language analyses. Additionally, Pacheco and Souza (2018) conduct a study in an attempt to illustrate the use of this taxonomy as a successful method to investigate aeronautical events that have language as a causal or contributing factor.

Sexton and Helmreich (2000, p. 66) also advocate for a more thorough examination of language use in aviation, stating that "Understanding variations in the language use is important to the extent that language use is related to flight safety". Their study approaches the need for more specific language research in aeronautical communications over the analysis of the occurrences of categories such as pronouns, which we will discuss further in the next section.

3 Pronouns in Aeronautical Communications

Communication in aviation is guided by a series of documents, such as the Manual of Language Proficiency Requirements (ICAO DOC 9835, 2010), the Manual of Radiotelephony (ICAO Doc 9432, 2007), ICAO DOC 4444 Air traffic Management (2016) and Annex 10 to the Convention on International Civil Aviation: Aeronautical Telecommunications (ICAO, 2001).

ICAO Doc 9835 (2010) makes the following reference to pronouns:

> 3.3.10 The principal linguistic characteristics of standardized phraseology (Philps, 1991) are a reduced vocabulary (around 400 words) in which each word has a precise meaning, often exclusive to the aviation domain, and short sentences resulting from the

[4] https://commons.erau.edu/db-lhuft/

> deletion of "function words" such as determiners (the, your, etc.), auxiliary and link verbs (is/are), subject pronouns (I, you, we) and many prepositions. (ICAO, 2010, p. 3-4)

From this perspective, it is understood that the elimination of certain words that are not considered relevant in terms of meaning is supposed to reduce the occurrence of miscommunication through ambiguity (ESTIVAL *et al.,* 2016; MODER, 2013; PHILPS, 1991). The use of call signs (information of the flight, with letters and number pronounced according to what is prescribed) would establish the identification for operations.

ICAO Doc Annex 10 about Aeronautical Telecommunications (2001) does not make specific mention to the use of pronouns, neither does DOC 9432, the Manual of Radiotelephony or ICAO Doc 4444 about Phraseology, even though featuring several instances of "we" and "you" in the example sentences.

Estival *et al.* (2016) present full lexical and functional grammatical categories in a linguistic description of AE. They say that only first and second personal pronouns "I, we, you" are used, and that there is rarely, if ever, a third person pronoun (because a noun phrase is reported in full, assumedly to avoid ambiguity).

According to Borowska (2017), personal, reflexive and possessive pronouns are not generally used in standard phraseology, being "I" an exception in "I say again", "you" in "How do you read?; Are you ready for pushback?; Do you want vectors?; Say your position".

Despite the orientations that discourage the use of pronouns and the conclusions from Estival *et al.* (2016) and Borowska (2017) regarding the non-outstanding use of pronouns in AE and phraseology, other studies show that the use of personal pronouns such as "we", "you", or "it" reveal interesting information for analysis.

Moder and Halleck (2012) claim that AE is a specialized language register through a corpus-based study. In the aviation corpus from Ohio State University, the 20 most frequent words were, in order, "the, to, one, two zero, you, three, five, and, of, four, seven, is, on, six, we, at, it, eight, and right" (p. 144), which differs from the ones in a corpus of general English (Corpus of Contemporary American English was the one she used). In COCA,[5] the most frequent 20 words are" the, is, and, of, a, in,

[5] https://www.english-corpora.org/coca/

to, have, it, I, that, for, you, he, with, on, do, say, this, and they". Among the differences, they highlight the appearance of numbers in the aviation corpus (a peculiar trait of aeronautical communications, to inform flight level, heading, runways and taxiways, call signs, procedures etc.), only three prepositions – "of", "on", and "at", the appearance of only three pronouns – "you", "it" and "we", not "he", "they", and "I", as in the general Corpus.

Prado (2010) also presents a list of the ten most frequent words in a corpus based on aeronautical communications, which are "you", "the", "to", "I", "and", "we", "a", "on", "it", "that". Her list displays three personal pronouns at the top, two articles and a preposition.

Sexton and Helmreich (2000) discuss the relationship of language use and flight outcome measures through the application of a "new" computer-based linguistic method for text analysis, a program called LIWC (Linguistic Inquiry and Word Count). Eighty-five language dimensions were analyzed, including personal pronouns, we, our, us, I, among others. One of their research questions was "how does language use vary across position and or level of workload?". The data were from a NASA study involving a three-person crew: a captain (C), a first officer (FO) and a flight engineer (FE), flying a simulated aircraft for a period of three days.

The conclusions point to the fact that individuals tend to communicate more along periods of high workload, much probably due to the multi-tasking involved in flight deck management. Specifically on the use of pronouns, some of their conclusions were that captains tend to use "we" (the first person plural) more often than FO's and FE's, especially in stressful situations, which could be due to the status and role of the captain. "This role requires more than active team building, and the status affords the right to use the first-person plural ('we need to…, our problem…, let's get out …') when briefing, planning or addressing the crew in conversation" (SEXTON; HELMREICH, 2000, p. 66). Additionally, there was an increase in the use of this pronoun by the three crew members as the familiarity increased along the three days. The use of the first-person plural was highly correlated with performance and could be a marker of familiarity or a more collective orientation towards the crew. Language use of pilots varies as a function of who is talking (C, FO or FE) and as a function of workload (SEXTON; HELMREICH, 2000, p. 66).

In a study dating back to 1994, an NTSB investigation covered flight-crew involved major incidents in U.S. carriers and acknowledged that "we" could be an important marker in that crew familiarity has been implicated as a moderating variable of aviation accidents (NTSB, 1994).

Cushing (1997), in a chapter entitled "Problems of Reference", offers an example of a specific case of confusion that the use of pronoun "we" led to. Two fighters were flying on instrument route and one developed mechanical problem and stated, "We need clearance back to base" (CUSHING, 1997, p. 18). The controller issued an IFR clearance and the aircraft replied, "We are in a left turn and we are climbing to 17,000ft" (CUSHING, 1997, p. 18). From this, the controller interpreted "we" as meaning that both aircraft were returning to home station. However, only the leading aircraft – the one that made the contact, was. The other continued on the original route. The pilot used "we" meaning the crew in that aircraft and the controller understood "we" as the two fighter aircraft flying together and this could have had negative consequences.

4 Pronominal choice in accomplishing cockpit identities

In "Beyond the Black Box", Maurice Neville (2004) analyzes how pilots accomplish identities using prescribed and non-prescribed pronominal forms. There are two formal identities which are assumed for pilots. The first one is automatically given according to their status as professionals, either as a Captain or as a First Officer. The other is related to the functions that they perform in operations, as the Pilot Flying (PF) – the pilot who is actually in charge of the maneuvers to make the aircraft fly, or as the Pilot Not-Flying (PNF or Pilot Monitoring (PM), – the who is in charge of tasks to assist the pilot flying.[6] Pilots have to be clearly aware of who is in charge of what and, in order to share this identity, they make use of pronouns.

[6] "Pilot Monitoring" has been used more recently because it seems to be more appropriate in describing the actual function of the pilot when not performing the actual tasks to fly the plane – as described by the Federal Aviation Administration (FAA), the US Aviation Agency. (https://www.faa.gov/other_visit/aviation_industry/airline_operators/airline_safety/safo/all_safos/media/2015/SAFO15011.pdf) In this article, PNF (pilot not-flying) will be used in accordance with what is used by Neville (2004).

Pronominal choices allow participants to establish how they are related to each other within the interaction and are important for pilots' communicative practice contributing to their awareness of who is doing what and what is going on. So that duties and responsibilities are clearly assigned to identities, pilots have to coordinate their work together.

> Through their pronominal choices pilots develop and demonstrate to one another their evolving understandings of these cockpit identities, from the engine start up and takeoff through to the landing and engine shut down. Pronominal choices indicate which identity pilots are occupying, at any given moment, in a setting where more than one identity may be available and legitimate. Pronominal choices help to allow pilots to make visible and be accountable for their moment-to-moment understanding of the identities they each occupy (NEVILLE, 2004, p. 34).

Crystal (1995, p. 201), in a traditional perspective, defines pronouns as "words that stand for a noun, a whole noun phrase or several noun phrases and a personal pronoun as the main means of identifying speakers, addressees and others". Neville (2004) adopts a different perspective, in accordance with Sacks (1992), who, in Chapters 011 and 11 of his Lectures, respectively, addresses Pronouns and tying Techniques: "The need to tie one's talk to another's preceding talk is a motivation to listen: tying properly shows that one has understood" (SACKS, 1992, p. 716).

In relation to the pronoun "we", Sacks says that it can be used to represent organizational status or capacity – when the speaker talks as an agent, and that a speaker may use "we" for category bound activities, as an indicator of the speaker's category membership". (SACKS, 1992, p. 333).

Neville (2004, p. 36) refers to a research by Malone (1997 *apud* NEVILLE, 2004) who puts that "the first person plural provides 'a powerful resource for calling up involvement obligations that require hearers to interpret who 'we' are at any moment and hence how and where the interaction is proceeding'."

Field (2020) states that the interactive listener retains certain aspects of form in his short-term memory in order to use them in the upcoming responses and is not just concerned with the speakers' meaning. From that, one could assume that personal pronouns seem to be relevant

in assigning references in exchanges and should be given attention in the communication dynamics in order to avoid problems such as ambiguity.

Additionally, the choices of pronouns can be affected by the characteristics of the setting or occasion of an interaction, which can be associated to specific patterns that may be developed over time related to organizational identities.

In order to check pronominal choices that enable pilots to establish who they are talking and listening to one another, Neville (2004) looks at numerous examples of personal pronouns which occur as part of the officially prescribed wording for pilots (in manuals of operating procedures and company policies) and at those which are not in a non-prescribed context.

The example below focuses on "your go" and "my go", part of prescribed words pilots are required to produce.

1 (0.9)
2 C/PNF: I have three three five(.) course bar three five five heading bug,
3 (0.7) A:SEL ADF, (0.2) it's your go.
4 (0.8)
5 FO/PF: my go.
6 (0.5)
7 FO/PF: go-around(.) flight level one eight zero (0.4) with ASEL (0.5)
8 right (of the) the pilot in command info: briefing as discussed.
9 (0.3) (NEVILLE, 2004, p. 40).[7]

This dialogue is said to have happened in the briefing moment before the flight, as the pilots prepare for takeoff. Through their pronominal choices, the pilots explicitly assign their identities as PF and PNF. Other similar examples given by the author are "your departure", "your power levers" and "my yoke", to determine who is in charge of a specific operational task.

The next example features a non-prescribed form:

[7] The reader can refer to the original source for further understanding of the symbols used to transcribe the conversations.

(13.4)
FO/PF:okay we need to plan hh- so the plan shall be:::, (3.4) go downhill
at (0.2) f::orty: (0.3) eight (0.4) mi::les:: er::: (0.4) south of
Destination (0.3) on DME on the GPS, (1.6) we'll expect to be
visual within twentyfive miles make a visual approach:, (1.7) to
join left downwind for left circuit landing runway one ei::ght::.
(0.3) the airfield elevation is eighteen (.) circuit height a thousand
feet is bugged on the altimeter. (0.9) visual procedures left circuit:.
(1.9) we'll be landing flap twentyfi::ve with a:: ah
(2.2) Vref of ninety:ni:ne and (0.2) seventeen point seven (ton),
(1.2) carry ten for a hundred and ni::ne (0.9) and Vfr Vel's a
hundred and nine and fourtee:n. (1.3) <and they're all se:t:.>
(0.8)
C/PNF:0 Set" ecrosschecked).
(0.8)
FO/PF:the fuel on board'll be: six forty, (1.2) it's about an hour and a
quarter's holding, (1.3) not really enough to go anywhere but er
we shouldn't have a problem getting on the ground in an hour.
(3.4)
FO/PF:and ah radio aids we got both the NAYs on Destination no::w we
might as well stick both the AD er ADFs up to Destination too.
(0.7)
((repeating alert tone))
FO/PF:number one ADF identified on Destination now as well.0
(4.3)
C/PNF:that's all understood (NEVILLE, 2004, p. 53).

Although the FO is the one performing the operations – the PF, he makes use of "we" instead of "I" in sentences such as 'we need to plan' (line 2), 'we'll expect to be visual' (line 4), 'we'll be landing flap twentyfi::ve' (line 9), 'we shouldn't have a problem' (line 18), 'we got both the NAVs' (line 20), and 'we might as well' (line 20) in order to make it evident that the activities – the planning, expecting, landing, etc.,

involve both pilots, in a shared identity as crew. Considering the fact that "we" can be interpreted as inclusive or exclusive, the meaning of the pronoun can be often vague and highly context dependent. (BIBER *et al.,* 1999; VAUGHAN; CLANCY, 2013).

Further examples featuring the use of "we" are provided: "we're clear to start", "we're estimating", "we're still traffic to you", "we've got traffic in sight", "we have that", "we don't require it", "we need to plan", "what we'll do", "let's get out of here", among others. They seem to set clear that the activities being performed are being taken as a jointly controlled task. That is, the use of the pronoun 'we" seems to be inclusive of the operational crew members.

The choices for "I", "my", "me", "you" and "your" were analyzed and interpreted as markers to invoke or make salient an individual identity for the other pilot, as shown in the below example:

1 (3.5)
2 FO/PF: take vertical speed and I'll just slow it down a bit more.
3 C/PNF: okay.
4 (4.0) (NEVILLE, 2004, p. 60).

By saying "I'll slow it down", the pilot wants to inform action taken and control of the plane. Other examples – "I'm going to let it run", "I'll take the autopilot's in", "I'll take runway two two", "I've had enough", "I'll have the heading" show that, by choosing one possible pronominal form, pilots are able to adopt for themselves, or assign to another, one of the possible cockpit identities.

The author also mentions that in naturally occurring cockpit talk-in-interaction pronominal choices can be a flexible interactional resource which allow pilots to move in and out of relevant cockpit identities, as shown in the next example:

1 (0.2)
2 C/PF: okay and I'll ah wait until we get the lineup(.) before I take the
3 locks off.
4 FO/PNF:yeah (.)transponders on(.) check's to flight controls.
5 C/PF: and you can tell him we're ready (yeah).
6 (0.2)

7 FO/PNF: yep.
8 (1.4)
9 *FO/PNF:>bravojul<iet:: ()tango ready.*
10 (1.6)
11 FO/PNF:[((coughs))
12 Tower: *[bravo juliet tango.*
13 (1.2) (NEVILLE, 2004, p. 73).

Here, the choices for "I", "we" and "you" seem to portray each one's identities and tasks in the procedure.

Neville (2004) also analyzes what he calls impromptu pronouns, a category that refers to forms that are also non-prescribed, but which occur as "embellishments of prescribed wordings. That is, pilots' talk may include personal pronouns where there are none in the officially prescribed wordings. The personal pronouns are not in the script but are impromptu" (NEVILLE, 2004, p. 76).

For instance, when pilots are running checklists, the prescribed wording would be only "set", or "selected" or received". Instead, in his data, pilots responded like "we've got that", or "you've got flaps ten". To the author, these pronouns do important interactional work as they emerge as part of pilots' accomplishment of their work and "help pilots to make explicit distribution of duties and responsibilities, and the control of various cockpit technologies" (NEVILLE, 2004, p. 77).

As we can see, the investigation proposed by Neville (2004) is significantly contributing insofar it explores a more social aspect comprehended by the use of certain personal pronouns in aeronautical communications. Nevertheless, it does not bring information about the frequency of those structures or a more in-depth exploration of other elements such as lexical items that accompany specific pronoun choices. Our study, then, intends to bridge this gap.

5 Method

CL is an empirical research approach to language use from the exploration of a corpus (a collection of texts as database) through computer-based tools. Our study aims to investigate the use of pronominal forms in a specialized corpus, CORPAC, presented below.

5.1 CORPAC

CORPAC (Corpus of Pilot and Air Traffic Controller Communication) is a corpus that I started to compile in 2017 with the help of two monitor students (not simultaneously) in the Aeronautical Science Program of the Pontifical Catholic University of Rio Grande do Sul. The project originally intended to be a joint work with monitor students in the Letters Program, so that we could have the collaboration of different perspectives in the compilation and analysis of the material – a more technical view on behalf of student pilots and a specialized linguistic contribution from the Letters Program students.[8] A minimum of 100000 words is the target.

This paper is based on a preliminary version of the corpus, from its first stages of compilation – with around 35000 words.

The corpus has been entirely built from emergency situations in aviation extracted from the videos freely made available by VASAviation, which is a Youtube channel that features selected situations from live ATC Emergency Situations/LiveATC). The videos are animations and contain the transcription of the audio. The criteria for the selection are basically about the emergency degree of the event and the availability of the transcription. That is, the video is watched by a student pilot, who then verifies if it actually portrays an emergency situation in aviation and if the transcription corresponds to what is being said.

Student-monitors were briefed about corpus research – its assumptions, entailments and impact and were instructed to:

1. Choose an episode featured on the channel, watch it and check if it actually presented an emergency situation.
2. Fill out a short form in the file "CORPAC INFO" with information about the episode, such as URL, title/ nature of the problem, date, flight/company/ aircraft, where (from/ to), English as a firs/ foreign language, phase of flight, duration of transcripts, and summary of the event.
3. This information can be essential to account for a number of variables in the analysis, such as the nature of the problem, the phase of the flight or if English is being used as a first or foreign language.[9]

---

[8] Currently, the project is on hold due to a number of reasons, but I expect to restart it as soon as possible.

[9] As information about the professionals is not disclosed in the source, it is not possible to accurately claim if the subject is a native speaker of English or not. Student-monitors

4. Write the transcription of the exchange in the same file, between "ATC-Pilot Transcripts and "End of transcript –", as shown above.
5. Transfer ONLY the transcripts to another Word File, "CORPAC", adding just the corresponding number of the event in the INFO file so that we can have access to background information about the event.

Accordingly, CORPAC has (so far) forty-three transcripts of emergency situations (123 pages in a Word file), based on videos which range from three to fifteen minutes and have been produced since 2008. The transcription procedures usually last long, and they also require two computers in order to facilitate the process– one to watch the animation and see the transcript and the other two write the transcription. As for the time spent in the process, student-pilots usually estimate one hour of writing for each minute of recording, altogether.

Despite the limitations of the database so far, the corpus already offers material for us to conduct preliminary linguistic analysis, such as the one presented in this study.

5.2 Software data analysis

With the view to obtain specific data from CORPAC, a popular and freely available software for corpus analysis was used: WordSmith tools.[10] The tools used were Concord, WordList and KeyWords.

The corpus was uploaded and required to generate:

1. A wordlist with the most frequent words
2. Concordance associations, with the most frequent collocate elements of a given pronoun.
3. Keywords showing their keyness value.

The proposed analysis ranges from a general picture of the use of personal pronouns in CORPAC to a more detailed look at the pronouns "we", "you" and "I", given their role in aeronautical communications as shown in Neville (2004).

---

were asked to fill out the form based on the company and other factors such as language proficiency and accent. As I evolve with the project and counting on the help of Student-monitors from the Letters program, I intend to conduct a more detailed categorization of this feature considering other factors and sources.

[10] https://lexically.net/wordsmith/

6 Results and Discussion

CORPAC totaled 36846 tokens and 1794 types. Table 1 below shows the twenty most frequent words:

TABLE 1 – The twenty most frequent words in CORPAC

Rank	Word	Frequency
1	THE	873
2	YOU	784
3	TO	764
4	AND	638
5	WE	611
6	TWR	537
7	ONE	496
8	TWO	456
9	UH	435
10	RUNWAY	419
11	ON	408
12	FOR	368
13	A	347
14	APP	306
15	AT	283
16	I	273
17	THREE	272
18	RIGHT	271
19	OF	255
20	IS	245

Source: Produced by the author.

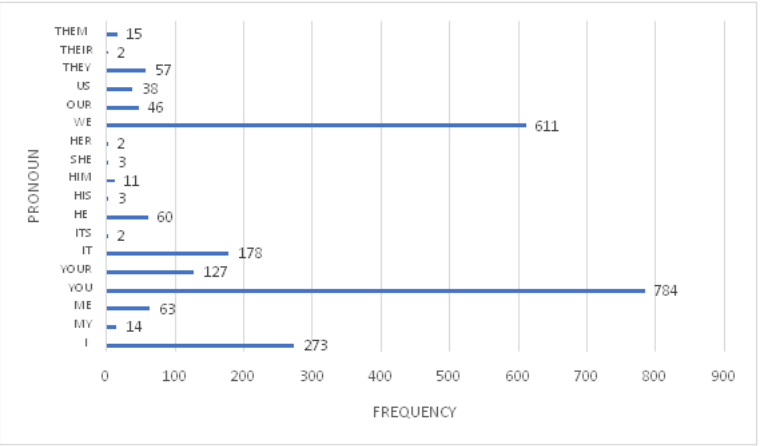
Results show two pronouns topping the list: “You”, in second, with 784 occurrences and “we”, coming in fifth, with 611. “I” occupied the 16th position in the rank, occurring 273 times. This is similar to what was found in Moder and Halleck (2012): “you ” and “we” in the top positions, as well as other word categories: numbers are indeed frequent,

the article "the" tops the rank in both corpora and preposition "to" is also commonly frequent. Our results also resemble Prado's (2010) – her top ten list features the same pronouns in CORPAC, and article "the".

Most are closed class words – determiners, prepositions, pronouns, conjunctions. Open class words are represented by items such as "runway", "twr" (tower), "right", and "app".[11]

The following graph presents an overall picture of pronominal occurrences in our corpus, taking into account first, second and third person pronouns.

GRAPH 1 – Pronominal Occurrence in CORPAC

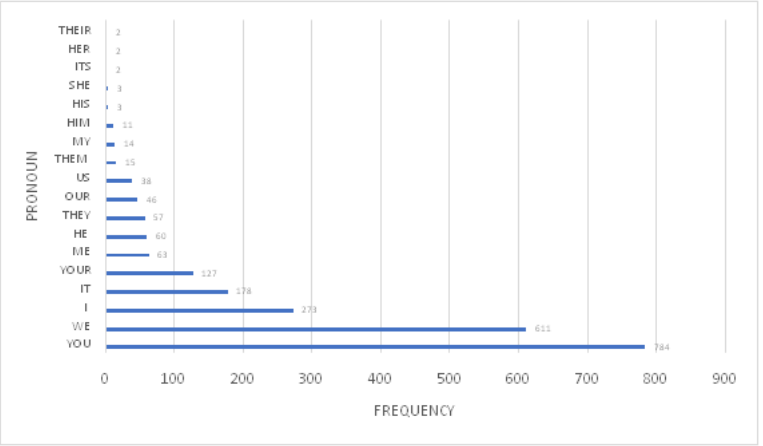


Source: Produced by the author.

We present the same graph featuring all the pronouns differently organized – ranked by their frequency as follows.

[11] Additional analyses combining the most frequent open and closed class words would be interesting insofar it could determine more precisely the association between the most frequent pronouns and nouns in the corpus. Although this proposal goes beyond the scope of this article, it should be considered as forthcoming research following this investigation.

GRAPH 2 – Personal Pronouns organized by Frequency



Source: Produced by the author.

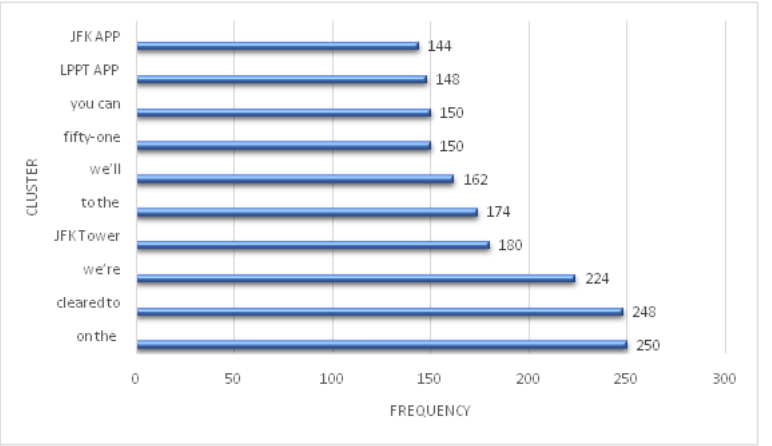
"You", "we" and "I" top the rank of frequency. They correspond, respectively, to 2,13%, 1,66% and 0,74% of the words in CORPAC. "You" had a +365.69 keyness[12] value and "we", a +282.17, coming in second and third position, only behind the article "the", with a value of +399,13, which is significant compared to all the other words in the corpus.

Having in mind the orientations not to use pronouns in aeronautical communications (as seen previously), we could say these numbers can be considered representative. Especially acknowledging Neville's (2004) assumptions that analyze the importance of these personal pronouns in assigning identities. In other words, the use of personal pronouns is not encouraged in aeronautical communications to avoid ambiguity and still "we", which is more likely to cause ambiguity than "I", is used almost three times more. Pilots seem to need to resort to it to optimize communication.

A further analysis of the two-word clusters in CORPAC show pronouns and prepositions topping the occurrences.

[12] The "keyness" value of a word can be obtained through the tool Keyword, uploading the target corpus and another reference corpus for comparison. In this study, the reference corpus used was BNC Spoken corpus, retrieved from http://www.natcorp.ox.ac.uk/using/index.xml?ID=freq

GRAPH 3 – 10 most frequent two-word Clusters in CORPAC



Source: Produced by the author.

"We're" is in third, "we'll" in sixth and "you can" in eighth. This appears to reveal significant information about aeronautical communications: despite having their use discouraged, pronouns seem to have a degree of importance in aviation exchanges given their high frequency, analyzed as a one-word or as a cluster.

A more detailed examination of the three-word clusters with the personal pronouns "I", "you" and "we" demonstrate that these elements seem to play a relevant role in Pilot-ATCO or Pilot – Pilot communication.

The most frequent clusters that feature "I' are "I don't" (10); And" I'll" (9); " I mean I" (7); "I'll get" (6) and "I'm just" (6). The most frequent three-word groups of words with "you" are "If you can" (15); "Do you have" (14); "Do you want" (13); "You have the" (13) and "Thank you very…" (11). The clusters that display "we" the most are "We're gonna" (15); "We'll" (14); "We're" (13); "We're going" (12) and "We'll be" (12).

The data show that three pronouns are associated with auxiliaries or modal verbs. Revising the instances in numbers, we verify that clusters with "we" happen more frequently than those with "I". The top five occurrences of "We" in clusters show the use of "will" or "be going to", which can be interpreted as a manifestation of plans, intentions or future actions. As pointed out by Neville (2004), "we" is a linguistic element that pilots resort to in order to establish shared identity. Sexton and Helmreich

(2000) have also mentioned that the use of "we" seems to reinforce the idea of "team-building" and that this use may be increased along the time shared in the cockpit by a sense of familiarity of the crew members.

As for the pronoun "you", data form CORPAC show that they are significantly frequent and used to assign clear identity in terms of pilots performance in operations, as in "If you can", "Do you want" and "you have the" – the last one associated with examples provided by Neville (2004) mentioned previously. The form "you" can be a singular or a plural pronoun and this flexibility probably accounts for its high occurrence and for possible ambiguities as well. An ATCO can use "you" addressing a pilot of a specific flight or and, depending on the content of the utterance, such as weather warning, other pilots can interpret it as a general remark. This is why other indications, such as the call sign (the identification of the flight) have to be used in order to mitigate possible ambiguities.

The following examples are extracted from CORPAC and illustrate the use of "we" in real emergency situations. Example (1) below features communicative strategies used in order to clarify the identity of "we":

(1) "We are not clear of the runway, we are on the runway. Cathay Zero-Seven-One is on the runway, crossing."

The pilot uses "we" twice, and the call sign right after it to make sure that the ATCO understands "we" as the crew in that specific flight, not another aircraft. A similar strategy can be observed in the next example.

(2) "(JFK APP) \x96 Delta 1888, it seems like the rate of turn is a little bit slower. Am I right to assume it\x92s gonna take you longer to turn?

(DAL 1888) \x96 We\x92re working on it, Delta 1888, we can tighten it up.

(JFK APP) \x96 Endeavour 3323, turn right heading 130, vectors for an emergency aircraft inbound."

The repetition of the callsign, that is, the code that identifies the flight – in this case, "Delta 1888", appears to confirm information about who "we" is referring to.

Example (3), on the other hand, brings an instance of "we" being used in a context where it could cause ambiguity.

(3) "GYI TWR – Seneca three-seven-Tango, right closed traffic; report midfield downwind runway one-seven left.

N5337T – Left closed traffic and report midfield right downwind. Three-seven-Tango.

PR-ITB – (...) Runway one-seven left. India-Tango-Bravo.

GYI TWR – Three-seven-Tango, I need you right traffic and report midfield right downwind.

N5337T – Right traffic and report right downwind. Three-seven-Tango.

GYI TWR – Papa-Romeo-India-Tango-Bravo, that was stepped on. Say your position from the airport.

PR-ITB – Radial one-three-zero. Now four miles.

GYI TWR – Papa-Romeo-India-Tango-Bravo, thank you. And make left traffic runway one-seven left. Report midfield left downwind.

PR-ITB – OK\x85 I understand we are cleared to land runway one-seven left.

N478BK – North Tex Tower- North Texas Tower, Cessna eight-Bravo-Kilo; three miles final."

When the pilot of the flight PR-ITB utters the sentence "I understand we are cleared to land runway one-seven left", he is not following communication rules stated by aeronautical phraseology which require the repetition of the call sign when reading back an instruction in order to avoid miscommunication. He seems to be unaware of the possibility of ambiguity, reinforced by the fact that, previously in the conversation, it is clear that the controller had to call his attention by saying "Papa-Romeo-India-Tango-Bravo, that was stepped on" when he readback an instruction meant to be directed to flight Seneca three-seven-Tango. The controller successfully detected that the pilot read back an instruction which was not assigned for him and was probably monitoring the phraseology deviations from this pilot in a way that, when he used "we" without clearly saying who "we" was referring to, he managed to

understand. Still, it is a potential example of how pronouns can cause ambiguity if identities are not clearly assigned.

Therefore, considering that the use of pronouns is not encouraged in aeronautical communications balanced against the fact that the results found in CORPAC suggest that they are more used than what would be expected, one would think about more risks for ambiguity. Even bearing in mind that pronominal choices seem to be justified by a reason such as identity assignment (NEVILLE, 2004; SEXTON; HELMREICH, 2000), in an ideal training context, learners should be made aware of the orientations from the official documents that regulate communications in a prescribed way, and should also be informed of the actual language occurrences in order to be better prepared to interact.

In other words, learners can benefit a lot from this non-prescriptive research perspective offered by CL. Real examples can be used, discussed and explored in class. Ambiguity is a problem in aviation exchanges which should definitely be mitigated, and corpus-based investigation seems to be a helpful tool.

7 Final Considerations

The aim of the study presented in this article was to analyze the use of personal pronouns in aeronautical communication based on CORPAC, a specialized corpus which is under compilation. To accomplish this goal, some concepts involved in the discussion were reviewed as were some studies that address the use of pronouns in exchanges in aviation, which appear to be significant despite orientations to avoid their use due to possible ambiguity.

Results from CORPAC about information regarding frequency and clusters associated with "I", "you", and "we" demonstrate that personal pronouns are frequent and seem to appear in constructions that are relevant for identities to be clearly assigned in such a high-stakes domain as aviation operations. After our preliminary analysis, the actual use of pronouns appears to mirror this communicative necessity.

It should be noted that, in accordance with the non-prescriptive approach of CL, this study is not meant to investigate the use of pronouns to assign rules in which they have to be employed in aviation. It is intended to describe the occurrence of some pronouns in real, spontaneous source of aviation language use. On that matter, CL showed to be a

fundamental tool to raise quantitative and qualitative data from such specific language domain. Likewise, the research also contributed to CL in that it reinforced the importance of empirical investigation to be confronted against formal orientations.

I understand that such information should be taken into account in aviation language training, not only in the level of the teaching practice of pronominal structures, but also in the metalinguistic level – pilots and ATCOs would profit from learning about how much their behavior in operations can be associated with the use of a pronoun. If there is orientation to avoid their use and if the findings from CORPAC show that they are frequently employed, learners should be made aware of the entailments and implications of their pronominal choices. Furthermore, testing practices could also benefit from this information, inasmuch task management can be reflected by the proper and clear use of pronouns.

This study is a preliminary examination on the use of pronouns in aeronautical communications. The limitations do not allow for a further analysis on a more detailed look at the actual occurrences of "I", "you" and "we" in longer sentences and in their context of utterances so to check possible deviations from Standard Phraseology. It would be interesting to go beyond frequency and cluster analysis and extend the information provided so far to compare it with the prescribed standard language to be used in aviation and to envision possible problems of ambiguity.

Additionally, further analyses of the occurrences of "I", "you" and "we" regarding a more thorough examination of the linguistic context, as well as a more particular investigation of the other pronouns, are necessary if we want to understand better and better the issues of aeronautical communications in order to promote aviation safety.

References

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E.; HIRST, G. *The Longman Grammar of Spoken and Written English.* Harlow: Pearson Education, 1999.

BIESWANGER, M. Aviation English: Two Distinct Specialized Registers? *In:* SCHUBERT, C.; SANCHEZ-STOCKHAMMER, C. (ed.). *Variational Text Linguistics*: Revisiting Register in English. Berlin: Mouton de Gruyter, 2016. p. 67-85.

BOCORNY, A. E. P. *Descrição das unidades especializadas poliléxicas nominais no âmbito da aviação*: subsídios para o ensino de inglês para fins específicos (ESP). 2008. 230f. Tese (Doutorado em Estudos Linguísticos – Teorias do Texto e do Discurso, Lexicografia e Terminologia: Relações Textuais) – Faculdade de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

BOCORNY, A. E. Panorama dos estudos sobre a linguagem da aviação. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 11, n. 4, p. 963-986, 2011.

BOROWSKA, A. *Avialinguistics*: The Study of Language for Aviation Purposes. Bern: Peter Lang, 2017.

CRYSTAL, D. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press, 1995.

CUSHING, S. *Fatal Words*: Communication Clashes and Aircraft Crashes. Chicago: The University of Chicago Press, 1997.

DIETRICH, R.; MELTZER, T. *Communication in High Risk Environment*. Hamburg: Linguistiche Berichte, 2002.

ESTIVAL, D.; FARRIS, C.; MOLESWORTH, B. *Aviation English*: A Lingua Franca for Pilots and Air Traffic Controllers. London: Routledge, 2016.

FIELD, J. Idle Chatter: What Really Goes on in Tests of Interactive Communication? *In*: *CRELLA Symposium*, [S.l.], 2020. Available at: https://www.youtube.com/watch?v=ONM7xcJTPD0&feature=youtu.be. Access on: July 11th, 2020.

HINRICH, S. W. *The Use of Questions in International Pilot and Air Traffic Controller Communication*. 2008. 294 p. Thesis (Doctorate in Philosophy) – Oklahoma State University, Stillwater, OK, 2008.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Annex 10 to the Convention on International Civil Aviation*: aeronautical telecommunications. Montreal: International Civil Aviation Organization, 2001.

INTERNATIONAL CIVIL AVIATION ORGANIZARION (ICAO). *Manual of Radiotelephony DOC 9432-AN/925.* Montreal: International Civil Aviation Organization, 2007.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Manual of Implementation of the Language Proficiency Requirements (DOC9835-AN/453)*. 2. ed. Montreal: International Civil Aviation Organization, 2010.

INTERNATIONAL CIVIL AVIATION ORGANIZATION (ICAO). *Air traffic management*: DOC 4444. Montreal: International Civil Aviation Organization, 2016.

LOPEZ, S. *Norme(s) et usage(s) langagiers*: le cas des communications pilote-contrôleur en anglais. 2013. 435f. Thèse (Doctorat en Linguistique Anglaise) - Université Toulouse le Mirail, Toulouse, 2013.

MALONE, M. *Words of talk*: the presentation of self in everyday conversation. Cambridge: Polity Press, 1997.

MATHEWS, E. English in Global Aviation: Historical Perspectives. *In*: FRIGINAL, E.; MATHEWS, E; ROBERTS, J. (ed.). *English in Global Aviation*: Context, Research and Perspectives. New York: Bloomsbury Publishing, 2019. p. 3-25.

MATHEWS, E.; PACHECO, A.; ALBRITTON, A. Language as a Human Factor in Aviation. *In*: FRIGINAL, E.; MATHEWS, E; ROBERTS, J. (ed.). *English in Global Aviation*: Context, Research and Perspectives. New York: Bloomsbury Publishing, 2019. p. 55-78

MODER, C. L. Aviation English. *In*: PALTRIDGE, B.; STARFIELDE, S. (ed.). *The Handbook of English for Specific Purposes*. West Sussex: Wiley-Blackwell, 2013. p. 227-242.

MODER, C.; HALLECK, G. Designing Language Tests for Specific Social Uses. *In*: FULCHER, G.; DAVIDSON, F. (ed.). *The Handbook of Language Testing*. New York: Routledge, 2012. p. 137-149.

NEVILLE, M. *Beyond the Black Box*: Talk-in-Interaction in the Airline Cockpit. London: Ashgate Publishing, 2004.

NTSB, 1994. Available at http://libraryonline.erau.edu/online-full-text/ntsb/safety-studies/SS94-01.pdf. Access on: Aug. 27, 2020.

PACHECO, A. *English for Aviation*: Guidelines for Teaching and Introductory Research. Porto Alegre: EdiPUCRS, 2019.

PACHECO, A.; SOUZA, G. Classificação e análise de acidentes aeronáuticos baseada em taxonomia considerando a língua como fator humano na aviação. *In*: SCARAMUCCI, M.; TOSQUI-LUCKS, P.; DAMIÃO, S. (org.). *Pesquisas sobre inglês aeronáutico no Brasil*. Campinas: Pontes, 2018. p. 23-47.

PHILPS, D. Linguistic Security in the Syntactic Structures of Air Traffic Control English. *English World-Wide*, [*S.l.*], v. 12, n. 1, p. 103-124, 1991. DOI: https://doi.org/10.1075/eww.12.1.07phi

PRADO, M. C. A. Corpus de inglês oral na aviação em situações anormais. *Aviation in Focus*, Porto Alegre, v. 1, n. 1, p. 48-57, 2010.

PRADO, M. C. A. *A relevância da pragmática no ensino do inglês aeronáutico: Um estudo baseado em corpora* [The relevance of pragmatics in the teaching of aviation English: A corpus-based study]. Doctoral dissertation, Universidade de São Paulo, 2019. Available at: www.teses.usp.br. Access: 1 Jun. 2020.

SACKS, H. Lectures on Conversation. Oxford: Basil Blackwell, 1992. 2 v.

SARMENTO, S. *O uso dos verbos modais em manuais de aviação em inglês*: um estudo baseado em corpus. 2008. 262f. Tese (Doutorado em Estudos Linguísticos – Teorias do Texto e do Discurso, Lexicografia e Terminologia: Relações Textuais) – Faculdade de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2008.

SEXTON, B.; HELMREICH, R. Analyzing Cockpit Communications: The Link between Language, Performance, Error and Workload. *Journal of Human Performance in Extreme Environments*, Cambridge, v. 5, n. 1, p. 63-68, 2000. DOI: https://doi.org.10.7771/2327-2937.1007.

SILVA, A. L. B.; TOSQUI-LUCKS, P. Around the World in Aeronautical and Aviation English Courses. *Revista CBtecLE*, São Paulo, v. 2, n. 1, p. 418-440, 2020.

SWINEHART, N. Aviation English Corpus Linguistics: Using the Right Phraseology? *Aviation English Corpus Linguistics*, Ohio University, Athens, Ohia. , p. 2-5, 2013.

TOSQUI-LUCKS, P. Aplicações de corpora no ensino e na avaliação de inglês aeronáutico: estado da arte, reflexões, direcionamentos. [Applications of corpora in the teaching and testing of aeronautical English: state-of-the-art, reflections, guidelines.] *In*: SCARAMUCCI, M.; TOSQUI-LUCKS, P.; DAMIÃO, S. M. (ed.). *Pesquisas sobre inglês aeronáutico no Brasil*. Campinas: Pontes, 2018. p. 89-114.

VAUGHAN, E.; CLANCY, B. Small Corpora and Pragmatics. *In:* ROMERO-TRILLO, J. (ed.). *Yearbook of Corpus Linguistics and Pragmatics 2013:* New Domains and Methodologies. Dordrecht: Springer, 2013. p. 53-73. DOI: 10.1007/978-94-007-6250-3_4.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1443-1484, 2021

Weather events in air traffic control standards and communication: discourse patterns and implications for language teaching and assessment

Eventos meteorológicos em normas e comunicações de controle de tráfego aéreo: padrões discursivos e implicações para o ensino e a avaliação de línguas

Rafaela Araújo Jordão Rigaud Peixoto
Department of Airspace Control (DECEA), Rio de Janeiro, Rio de Janeiro / Brazil
rafaela.peixoto@gmail.com
https://orcid.org/0000-0002-3504-8405

Patrícia Tosqui-Lucks
University of São Paulo (USP), São Paulo, São Paulo / Brazil
Airspace Control Institute (ICEA), São José dos Campos, São Paulo / Brazil
patricialucks@gmail.com
https://orcid.org/0000-0001-9104-2123

Abstract: Weather events affect air traffic control (ATC) in many ways, for there are many situations that need to be reported in pilot-controller communication. This paper attempts to analyze the language used to express the impact of meteorological phenomena to air traffic operations, particularly in regard to aeronautical English, that is, the communication used during radiotelephony by air traffic controllers in training situations. For that, two types of analyses will be carried out: one regarding the formulaic structure of lexical units using 11 Aeronautical Meteorology terms within the ATC context (phase 1); and another one concerning the use of these terms by students in three ATC courses (for TWR, ACC and APP facilities) and how it affects their performance during communication activities in a learning environment (phase 2). These analyses will be based on rationales of lexical semantics for terminology; corpus linguistics (CL), comprising English for Specific Purposes (ESP) and learner corpora; and considerations about vocabulary assessment on aeronautical English exams. Results suggest that terminological patterns discussed in this paper show how meaning is dependent on

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1443-1484

context, and how lexical semantic analysis of terms may contribute to reveal nuances of language used in a specialized context. In this way, it indicates courses have been efficient in teaching and practicing the use of the main meteorological terms related to aeronautical English and that, despite some mistakes students make, evidence points out that they are able to report weather conditions to pilots and to understand pilots' requests in a proficient level concerning vocabulary.

Keywords: meteorology; aeronautical English; terminology; learner corpus; language assessment.

Resumo: Eventos meteorológicos afetam o controle de tráfego aéreo (ATC) de diversas formas, dado que muitas situações precisam ser reportadas na comunicação entre piloto e controlador. Este artigo pretende analisar a linguagem utilizada para expressar o impacto de fenômenos meteorológicos para operações ATC, particularmente quanto ao uso de inglês aeronáutico, ou seja, a comunicação utilizada durante a radiotelefonia, por controladores em situações de aprendizagem. Para isso, duas análises foram realizadas: em relação à estrutura formulaica de unidades lexicais contendo 11 termos de Meteorologia Aeronáutica no contexto ATC (fase 1); e quanto ao uso desses termos por alunos de três cursos ATC (para os órgãos operacionais TWR, ACC e APP) e como isso afeta seu desempenho durante as atividades de comunicação em um ambiente de aprendizagem (fase 2). Essas análises serão fundamentadas nas teorias de semântica lexical para terminologia; linguística de corpus (LC), compreendendo Inglês para Fins Específicos (ESP) e corpora de aprendizes; e considerações sobre avaliação de vocabulário em exames de proficiência de inglês aeronáutico. Os resultados sugerem que os padrões terminológicos discutidos mostram como os significados dependem do contexto, e como a análise léxico-semântica de termos pode contribuir para revelar nuances da linguagem utilizada em contexto especializado. Desta forma, demonstrou-se que os cursos foram eficientes no ensino e na prática do uso dos principais termos meteorológicos e que, apesar de alguns erros cometidos, as evidências apontam que os estudantes foram capazes de reportar condições meteorológicas e compreender as solicitações dos pilotos com nível de proficiência adequado em relação a vocabulário.

Palavras-chave: meteorologia; inglês aeronáutico; terminologia; corpus de aprendizes; avaliação de línguas.

Submitted on October 12th, 2020

Accepted on December 7th, 2020

1 Introduction

The extent of weather events affecting air traffic control (ATC) is generally taken for granted, but it varies greatly, from the amount of

water film on the runway on a rainy day to volcanic ashes coming from another country as situations that need to be reported in pilot-controller communication. In this way, this paper attempts to analyze the language used to express the impact of meteorological phenomena to air traffic, particularly when it occurs in international traffic, and these professionals need to use English to communicate.

After a few fatal accidents which had communication problems as contributing factors, the International Civil Aviation Organization (ICAO) issued, in 2004 (with a reviewed second edition in 2010), the Manual of Language Proficiency Requirements, known as Doc 9835, in order to establish some parameters for English language proficiency, involving listening and speaking skills, for international pilots and air traffic controllers (hereafter, we will use the term 'controllers') who work in multilingual environments. According to this document, these professionals should be able to communicate through a highly specific code for aviation purposes, i.e. *aeronautical standard phraseology*,[1] and *plain language* whenever phraseology does not suffice to communicate in non-routine situations. The concepts of standard phraseology and plain language, which constitute the essence of the aeronautical English, are explained in Table 1, as follows:

TABLE 1 – Definitions of phraseology and plain English.

Term	Definition/Conceptualization
Phraseology (standard phraseology)	It is a code used by pilots and air traffic controllers, in a limited number of restrict and predictable communicative events characterized by short phrases and reduced vocabulary which allows a concise, precise and efficient transmission of information related to a flight.
Plain English, plain language	It is the use of the English language in radiotelephony communication that exceeds the use of standard phraseology, when it is not sufficient, but that should mirror phraseology, keeping its characteristics and specificities, as well as the same critical safety requirements such as intelligibility, non-ambiguity and concision.

Source: Adapted and translated from Scaramucci; Tosqui-Lucks; Damião (2018, p. 300).

[1] ICAO recommendations for the use of standard phraseology can be found in Doc 9432, Manual of Radiotelephony (ICAO, 2007) and Doc 4444, Air traffic management (ICAO, 2016).

By considering those definitions on phraseology and plain language, Tosqui-Lucks and Silva (2020a, 2020b) discuss the aeronautical English concept, explaining that controllers and pilots have to make crucial decisions, which require high levels of attention, focus and memory. Proficiency in a foreign language may pose major stress in those situations, due to interlinguistic aspects, different intercultural backgrounds, possible code-switching and other pragmatic issues (Cf. TOSQUI-LUCKS; SILVA, 2020a).

Concerning *plain English* as one of the elements of radiotelephony communications, it must be used according to the same parameters of conciseness, precision, objectivity, intelligibility and unambiguity that govern the use of phraseology (ICAO, 2010, p. 3-5), i.e., in no way having the connotation of English for use in common everyday situations (SCARAMUCCI, 2011), nor for use in other aviation contexts, which escape communication by radiotelephony.

In this sense, it is paramount to be aware of phraseological patterns in aeronautical language as well as making proper use of specialized terminology in air traffic control communication. Therefore, in an attempt to follow these recommendations, the Department of Airspace Control (DECEA), a military organization of the Brazilian Air Force, attributed to the Airspace Control Institute (ICEA), responsible for ongoing training of professional controllers, the mission to develop both aeronautical English courses and an aeronautical English test to make sure all controllers involved with international traffic have at least the minimum required proficiency level (PL) to ensure safety in Brazilian skies.

Thus, ICEA has been developing on-site courses and, since 2015, on-line courses too, aimed at professionals who work at the three different ATC facilities: tower (TWR), mainly responsible for landing and take-off; approach control (APP), in charge of operations when the aircraft are flying after take-off or preparing to land; and area control center (ACC), responsible for aircraft on cruising level. Professionals working at these three different facilities have specific characteristics, responsibilities and tasks to perform, addressed accordingly in the three online courses developed for each of them, as will be more detailed in the next sections of this paper.

Concerning development and application of the Aeronautical English exam for Brazilian Air Traffic Controllers (EPLIS),[2] ICEA follows ICAO guidelines, which prescribes 6 PL, from which PL 4 is the minimum required to operate internationally; and assesses six independent descriptors, i.e. structure, vocabulary, pronunciation, listening comprehension, fluency and interaction. In this paper, we will focus on the descriptor vocabulary, based on 11 selected terms related to meteorology as follows: (1) rain, (2) wind, (3) wind shear, (4) turbulence, (5) wake turbulence, (6) conditions, (7) lightning, (8) formation, (9) cloud, (10) fog, and (11) thunderstorm. In this way, discourse patterns in the context of weather events in air traffic control standards and communication, and their implications for language teaching will be analyzed.

To study discourse patterns in air traffic control standards, and language teaching implications in air traffic communication during learning activities, this paper is organized in the following way: presentation of the theoretical panorama comprising rationales of lexical semantics, corpus linguistics (CL), including ESP and learner corpora, and considerations about vocabulary assessment on aeronautical English exams; detailed methodology explaining the compilation process of the reference corpus and the learner corpus, and the methodology design; discussion of discourse patterns regarding weather events in air traffic control phraseology standards (phase 1), specifically addressing formulaic structure of lexical units using 11 Aeronautical Meteorology terms within the ATC context; discussion of weather events in air traffic control communication in the learner corpus, based on the use of these terms by students in three ATC courses (for TWR, ACC and APP facilities) and how it affects their performance during communication activities in a learning environment (phase 2). In the last section, we will consider some implications for Aeronautical English teaching and make suggestions for addressing the weather terms on courses.

[2] In Portuguese, EPLIS stands for *Exame de Proficiência em Inglês Aeronáutico do Sistema de Controle do Espaço Aéreo Brasileiro.*

2 Theoretical Foundation

2.1 Phraseological patterns: a lexical semantic approach to terminology

For the study of terminology, it is paramount to verify the patterns of language, as how they relate to other terms in a language. According to Hunston (2010, p. 158), "observing pattern involves identifying similarity and forming notional categories." In this sense, a word or term with the same meaning may be considered to have a different pattern, as it related differently to other collocates or its cotext.

To exemplify this perspective, Hunston (2010) analyzes verbs in a corpus used in her research to identify objective-subjective nature based on collocates, arguments and cotext, and for the verb react, she lists eight patterns:

> (1) REACT followed by a subordinate clause indicating stimulus; […]
> (2) REACT followed by the preposition to; […]
> (3) REACT followed by an adverb and then by the preposition to; […]
> (4) REACT followed by a to-infinitive clause indicating consequence; […]
> (5) REACT followed by the preposition with answering the question 'how?'; […]
> (6) REACT followed by the preposition with answering the question 'what?'; […]
> (7) REACT followed by a full stop; […]
> (8) Other lines:
> 4 two-thirds of the radical pairs reacting (in a field of typically only [...]
> 13 efforts you may find the magician reacting too early or late. Also bear in. (HUNSTON 2010, p.160.)

Along with this perspective, Sinclair (2008) advocates that phraseological study must stem from the analysis of collocates (coselection), and not lexical and grammatical structures alone. If they are treated independently, without considering a differentiated meaning when combined in a specific way (phrases), studying phraseologies would not be fruitful.

Sinclair expands this perspective by classifying analysis of meaning in three levels: (1) contextual settings, as studied by Firth, using cotext analysis; (2) phraseological, by analyzing collocational frameworks (Cf. RENOUF; SINCLAIR, 1991); and (3) lexical and grammatical, where the grammatical stance presupposes a pool of possible choices, in which abstract patterns underlie meaning, and the lexical stance detail lexical items according to the meaning they create.

The interdependence of elements, as cotext, is one of the main contributions of CL, since it enables analysis of how terms actually behave in a language, not considering them as "closed" structures. In fact, defining a term is always a very complex task, as there is no set standard that works for all situations. In the case of specialized fields, this issue is even more sensitive (Cf. FINATTO, 2001; PEIXOTO, 2020), as there is a traditional perspective based on an Aristotelian point of view that word senses could be devoid of subjectivity by attributing general content (genus) + specification (differentia). The problem is that such a clear-cut perspective does not work so smoothly in most contexts, as meanings are more related to the word environment, i.e., how words/terms relate to other lexical items around this main given term.

In this way, the main contribution of lexical semantics is that it relates the semantic content of words to other words and associations, named combinatorics. In the air traffic environment, for example, the term 'conditions', analyzed in this paper, may bear the same general content of "the possibility of a situation to happen", but the way it relates to other collocates actually specializes this meaning. For example, 'air traffic conditions' is different from 'meteorological conditions': while the first one may refer to the general context from departure to take-off, including weather conditions and aircraft conditions, the second one is more related to weather phenomena such as clouds, snow or thunderstorm.

Since it relies on contextual variables, lexical semantics take into consideration concepts and relations ideally extracted from running text. In this way, relations between concepts may range greatly, and polysemy becomes an issue as it is more sensitive to precisely define the whole scope of a specific word definition.

When it comes to teaching specialized language, the need of standardization tends to be stressed, but it must consider language in context. In this sense, semantic labels intend to delineate the conceptual

structure from a relational perspective, not only a definitional one, so as to enable understanding variation as part of language concepts, not as deviation.

In the case of multiword expressions, they may also be considered terms, and, as a matter of fact, most entries in specialized dictionaries are multiword terms. In the analysis carried out in this paper, 'wind shear escape', for example, clearly has a more specific definition than 'wind shear' itself.

Considering polysemy as typical in language, as it entails variation, leads to the elimination of the useless concern of trying to have many clear-cut definitions to situations where only nuances apply. In addition to that, it addresses cases where interferences with general language may occur, i.e. "a lexical item can denote a concept in a specialized field and convey a different meaning in everyday situations" (L'HOMME, 2020, p. 81). This is the case with the term 'fog', which has specificities regarding the range of visibility, something that is not taken into account in everyday situations but is very relevant for the specialized context, as explained by Peixoto (in press) in the following excerpt:

> Regarding 'fog' (FG), 'haze' (HZ) and 'mist' (BR), the classification depends on humidity and visibility issues. 'Fog' is reported when the air is at about 100 per cent humidity and the visibility is less than 1000 m [Cf. ICAO 2005], while 'mist' presents visibility ranging from 1000 m and 5000 m, and relative humidity above 90 per cent. [Cf. ICAO 2005]. On the other side, 'haze' are "extremely small particles invisible to the naked eye and sufficiently numerous to give the air an opalescent appearance [...], usually only a few thousand feet thick, but may extend upwards to 15,000 feet (4,600 meters) [...]" and visibility may vary "greatly, depending on whether the pilot is facing into or away from the sun" (FAA, 2016: 16-5). Concerning those categories, although "mist may be considered an intermediate between fog and haze" (ibidem), identifying those phenomena may be critical, as "there is no distinct line between any of these categories" (ibidem). In Portuguese, fog (FG), haze (HZ) and mist (BR) are translated as 'nevoeiro', 'névoa seca' e 'névoa úmida'. (PEIXOTO, in press).

A relational perspective also allows for an enduring definition, as concepts may change in time, mainly due to our understanding of knowledge, and structural definitions may need updates. Within this context, the focus of the lexical semantic approach is comprehending where the terms are located in a language system, considering interrelations.

Errors are commonly the focus of linguistic analyses within a learning environment but finding regularities in the use of language contributes greatly as errors are not necessarily related to a very low level of proficiency. Ebeling and Hasselgård (2015) have shown learners from higher levels of proficiency make an equivalent number of mistakes mostly because they use more complex structures. In that sense, grammatical mistakes are more related to verbal structures which are not commonly used in language, and more varied lexical structures tend to work as a thermometer to measure the actual level of proficiency.

Based on findings by Nesselhauf (2005), Thewissen (2008) and Chen (2013), Ebeling and Hasselgård (2015) clarify that it is more important to analyze the type of error a learner makes, not only the quantity of errors. More sophisticated structures such as phrasal verbs are more error-prone than simple structures, yet phrasal verbs are mostly used by more advanced learners. As a result, Nesselhauf (2005) has noticed in her investigation that free combinations account for 25% of errors while collocations account for 40% of errors. Of course, this must take into account student background as well as personal effort of individuals in the learning process. As Meunier explains,

> Individual differences typically include aptitude, motivation, identity issues, personality traits, type of working memory, socio-educational background, language proficiency in the mother tongue (L1) and other languages learnt, but also numerous aspects related to cognitive restructuring. (MEUNIER, 2015, p. 385.)

Meunier (2015) expands this perspective by resorting to Bartning and Forsberg's (2006) study, indicating that the use of prefabricated language is a more skilled capacity in comparison to the use of simple verbal morphology. In this sense, the students' abilities to actually develop more sophisticated proficiency depends greatly on the communicative style of learners, which we believe could also be nurtured by following certain strategies. In Meunier's words, "whilst verbal morphology

displays what they call a strict development (p.19), prefabricated language does not seem to follow such strict development and is more sensitive to input and to the communicative style of individual learners" (MEUNIER, 2015, p. 392)

To enable a more representative assessment of students' proficiency based on language used by them, the analysis of collocational patterns may be more relevant, since it allows for a more contextual perspective, considering how words relate to each other to constitute meaning. Ebeling and Hasselgård (2015) enlighten us on the relevance of idiomatic phrasal constructs and explain that:

> 'Collocation' is defined as involving some degree of fixedness/ restriction on the combinations of verb with noun. This definition separates collocations from free combinations, in which the verb and the noun combine without arbitrary restriction, and idioms, in which both verb and noun have lost their original meaning, or which can only be used with the idiomatic sense in restricted environments. (EBELING; HASSELGÅRD, 2015, p. 220).

When considering discourse patterns in a given specialized language, adjectives are specifically harder to be captured in a conceptual structure (Cf. L'HOMME, 2020) since their meanings may have subtleties which would only be understood when analyzed in context. In language learning, adjectives are also considered the most complex structure precisely due to their combinatorial nature, also comprising specific order in multiword expressions.

By considering this complex panorama of weather events affecting air traffic operations, discourse patterns were analyzed according to a lexical semantic approach for terminology, to assess semantic relations of Aeronautical Meteorology terms, based on a classification of semantic labels as described in Table 2.

TABLE 2 – Description of semantic labels

#	Label	Description
01	CHARACTERISTIC	It refers to the trait, quality or property of the meteorological condition. E.g. 'cold ~'
02	CHARACTERISTIC / INTENSITY	It is a label which combines the labels characteristic and intensity.
03	DIMENSION	It refers to the size or dimension of the meteorological condition E.g. 'small ~'
04	DURATION	It refers to the time elapsed since the beginning of the meteorological condition or continuously. E.g. '~ during the night'
05	EPISODE	It refers to an occurrence as an episode or instances of the meteorological condition. E.g. '~ registration'
06	EPISODE / INTENSITY	It is a label which combines the labels episode and intensity.
07	FORECAST	It refers to a forecast, observation or notification of a meteorological condition. E.g. 'observed ~'
08	FORM	It refers to the objective form of the meteorological condition, generally of concrete nature. E.g. '~ pellets'
09	INFORMATION FACTOR	It refers to an information or data factor with the purpose of quantifying the meteorological condition in some way. E.g. '~ data'
10	INSTRUMENT	It refers to instruments or devices used to measure or forecast a meteorological condition. E.g. '~ sensors'
11	INTENSITY	It refers to the level of intensity of a meteorological condition, generally associated with another feature (label). E.g. 'strong ~'
12	LAYOUT	It refers to the layout or arrangement of the meteorological condition in the overall scenario. E.g. '~ vertical profile'
13	LOCATION	It refers to the location where the meteorological condition takes place, which can range from a cardinal direction or a geographical position, to a city or an airport. E.g. '~ *no aeroporto'* ['~ at the airport']

14	MANAGEMENT	It refers to procedures derived from decisions taken to manage problems E.g. '~ mitigation techniques'
15	MOVEMENT	It refers to movement or continuous occurrence of a meteorological condition. E.g.'blowing ~'
16	PARAMETER	It refers to a standard used as comparison within a framework of meteorological conditions. E.g. 'minimum ~'
17	PHENOMENON	It refers to an occurrence which precisely characterizes the meteorological condition. E.g. '*precipitação de ~'* ['~ precipitation']
18	REFERENCE	It refers to a standard used as spatial indication of a meteorological condition. E.g. 'minimum height of ~'
19	RELATED TERM	It refers to another term which is semantically related to the term analyzed. E.g. '~ and precipitation'
20	TYPE	It refers to a meteorological condition of a particular kind, class or group. E.g.'surface ~'
21	TYPE / DIMENSION	It is a label which combines the labels type and dimension.
22	TYPE / INTENSITY	It is a label which combines the labels type and intensity.
23	UNIT OF MEASUREMENT	It refers to a unit of measurement used to indicate a physical quantity regarding the meteorological condition. E.g. '~ *em (200) hP'* ['~ in (200) hP']
24	VARIATION	It refers to a variable state of a meteorological condition. E.g. '~ gradient'

Source: Adapted from Peixoto; Pimentel (2020).

There are more labels in the original paper by Peixoto and Pimentel (2020), and some others may be created to address the semantic nature of additional terms, as it was the case of the label 'management', which represents the relational context of the term '~ mitigation techniques', for example.

Such lexical semantic terminological research is best equipped with corpora resources because it enables words to be analyzed in context, identifying different forms of concept or meaning expression. In a specialized approach, corpus containing institutional documents is

a relevant contribution because it contains language and perspectives of experts in the field. The next section will address this theoretical issue more thoroughly.

2.2 Corpus Linguistics: English for Specific Purposes and learner corpora

Many authors attest the benefits of CL to research and teach vocabulary (BERBER-SARDINHA, 2011; SCHMITT, 2000; STEFANOWITSCH, 2020; TAGNIN, 2006; TOSQUI-LUCKS; PRADO, in press). According to Schmitt (2000), corpus evidence has shown two important things: (i) that a very limited number of high-frequency words do the bulk of the work in language, and it is crucial that students master them; and (ii) that words tend to collocate, that is, multiword strings seem to act as a single lexeme. In fact, the author says that a major direction in vocabulary studies today is "researching these multiword units through corpus evidence to establish their frequency and behavior" (SCHMITT, 2000, p. 89).

This is part of a move from lexis as individual words to be considered in isolation toward viewing them as integral parts of a larger discourse, and it is valid to general English and English for Specific Purposes (ESP) discourse too. In this matter, Stefanowitsch (2020, p. 215) complements that all corpora consist of orthographically represented language, and this makes it easy to retrieve word forms. To him, the focus on words is also due to the fact that the results of research using CL have proved that words (individually and in groups) are more interesting and show a more complex behavior than traditional, grammar-focused theories of language. As an example, we can consider the word 'wind', which has different uses and meanings depending on the impact it has for aircraft landing, and can be expressed in multiwords such as 'crosswind', 'tailwind', 'downwind leg', etc.

Still considering CL for teaching vocabulary, Berber-Sardinha (2011) states that most pedagogical tasks focus on concordances, and presents some text-centered and multi-genre alternatives. The author also highlights some areas that may deserve attention in the larger context of Brazilian educational CL. Some of them are represented in this study: more research about it on academic level, more integration with diverse areas, more application on educational contexts, more pedagogical materials and teaching resources based on corpora and more

integration with distance education. For the latter, Berber-Sardinha, in the above-mentioned work, says that both distance learning and CL are technological areas that can profit a lot if instructional designers learn more about CL tools.

Gavioli (2005) states that corpus work in ESP appears to match teachers' and learners' requirements particularly well, for corpus analysis highlights recurrent features of language. The possibility of having instruments to describe the routine aspects of ESP language is a key teaching issue in ESP courses, where the teacher is often split between the need to be both an expert in the foreign language and an expert in the specialized discipline. Corpora of specialized texts seem to be a very useful instrument in isolating and providing indications about key lexical, grammatical or textual issues to deal with in ESP classes. Creating corpora from specialized texts is relatively easy and inexpensive for most teachers who are familiar with computers, and analyzing such pools of texts with concordancing software may suggest relevant lexico-grammatical items and the way they are used to deal with in the ESP class and the way they are used (GAVIOLI, 2005, p. 5). The author also highlights the advantages of "home-made" corpora created *ad hoc* for some particular teaching or learning purpose, which is our case. Even though there is some criticism about using corpus for pedagogical reasons because of a possible "confusion between what is scientifically interesting and what is pedagogically useful" (GAVIOLI, 2005, p. 27), she supports data-based corpus analysis for English as a Foreign Language (EFL) teaching because it can help researchers and material designers in producing more authentic descriptions of language usage which, in their turn, may improve teaching and reference materials.

Tosqui-Lucks and Prado (in press) state that, for many years, vocabulary selection for course content was made intuitively by material designers. With the advent of CL, computational tools started to be used as a source of information for textbooks. In ESP areas, this is even less common, and only recently CL findings started to be used in aviation. The authors present a list of corpora of aviation and aeronautical English compiled internationally and results from studies with four different corpora compiled with international and Brazilian pilots and controllers, considering ESP and learner corpora.

Gilquin (2015) states that, like any corpus, the learner corpus is a collection of machine-readable authentic texts (which can be written or

be transcripts of spoken data) sampled to be representative of a particular language or language variety. What makes the learner corpus special is that it represents language as produced by foreign language learners; and what makes it different from the data used in earlier second language acquisition studies is that it seeks to be representative of this language variety. To tackle the issue of degree of naturalness when defining learner corpora, the author cites Granger's (2008, p.338) definition of learner corpora as "electronic collections of (near-) natural foreign or second language learner texts assembled according to explicit design criteria" suggesting that they may be comprised of texts that are not, strictly speaking, naturally occurring texts. This is because, especially for foreign language learners, the target language only fulfils a limited number of functions, most of which are restricted to the classroom context. To this matter, Römer (2004) adds that the problem of authenticity in English language teaching has been discussed for many years. To her, "what authenticity really means in a language teaching context, which different types of authenticity play a role and whether or not we want to teach authentic English to our pupils are highly controversial questions among linguists and didacticians" (RÖMER, 2004, p. 153).

Tagnin (2006) states that a learner corpus can provide useful data to detect specific difficulties of language learners and consequently inform the production of pedagogic material to address these problem areas. To the author, a learner corpus can provide useful data to detect such specific difficulties and consequently inform the production of pedagogic material to address these problematic areas, but one of the problems with textbooks used in Brazil for teaching a foreign language is that most are written by foreign authors unacquainted with Brazilian students' difficulties. Then, in an attempt to overcome possible limitations and fulfill specific needs of the Brazilian context, we have compiled a learner corpus, with productions from controllers during in-service distance learning training. The discussion of language patterns in this learner corpus will be based on guidelines for the vocabulary descriptor of language assessment, published by the International Civil Aviation Organization (ICAO), as discussed in the next section.

2.3 Considerations about vocabulary assessment on aeronautical English exams

The documents that guide aeronautical English teaching and assessment, according to ICAO regulations, are Doc 9835 (ICAO, 2010) and Circular 323 (ICAO, 2009).[3] The first one defines aeronautical radiotelephony communications (Chapter 3) and provides guidance on language proficiency teaching and assessment for pilots and controllers (Chapter 7; Chapter 6), while the other complements it by presenting specific recommendations for course designs, both classroom-based and through distance learning. The mentioned Circular details the design and development of language courses emphasizes that language teachers should be trained to teach this very particular type of ESP and enumerates a few characteristics of aeronautical communication: it is essentially oral, with no visual cues, and employs a very specific vocabulary, as clear and unambiguous as possible, "because it involves risk management not only for pilots and ATCOs but for society at large" (TOSQUI-LUCKS; SILVA, 2020a, p. 3).

These documents reinforce that both teaching and assessment should be guided by ICAO Language Proficiency Rating Scale, Annex 1, Doc 9835 (ICAO, 2010), for speaking and listening proficiency only, according to six differentiating PL (being 1 the lowest, 6 the highest and 4 the minimum to be considered operational). There are recommendations for assessing the candidates holistically and analytically. Doc 9835 presents the following holistic descriptors:

> Proficient speakers shall:
> a. communicate effectively in voice-only (telephone/ radiotelephone) and in face-to-face situations;
> b. communicate on common, concrete and work-related topics with accuracy and clarity;
> c. use appropriate communicative strategies to exchange messages and to recognize and resolve misunderstandings (e.g. to check, confirm, or clarify information) in a general or work-related context;

[3] In this paper, we are referring to the second edition of Doc 9835 (2010), which was revised and included a great part of Cir 318 (2009) about Aviation English assessment – but the first edition of Doc 9835 was published in 2004, thus, earlier than Cir 323 (2009).

> d. handle successfully and with relative ease the linguistic challenges presented by a complication or unexpected turn of events that occurs within the context of a routine work situation or communicative task with which they are otherwise familiar; and
> e. use a dialect or accent which is intelligible to the aeronautical community. (ICAO, 2010, Appendix I.)

As for the analytical assessment, there are band descriptors for pronunciation, structure, vocabulary, fluency, comprehension and interaction. Specifically about the category vocabulary, the analytical scale for PL4 states that:

TABLE 3 – ICAO rating scale vocabulary PL4

Vocabulary range and accuracy are **usually** sufficient to communicate effectively on common, concrete, and work related topics. **Can often** paraphrase successfully when lacking vocabulary in unusual or unexpected circumstances.

Source: ICAO (2010) Attachment A (our emphasis)

The same rating scale presents the following description for vocabulary PL3 (that is, not suitable for international traffic):

TABLE 4 – ICAO rating scale vocabulary PL3

Vocabulary range and accuracy are **often** sufficient to communicate on common, concrete, or work related topics but range is limited and the word choice often inappropriate. Is **often unable** to paraphrase successfully when lacking vocabulary..

Source: ICAO (2010) Attachment A (our emphasis)

If we look at the description for PL2, i.e. "limited vocabulary range consisting only of isolated words and memorized phrases", it is clear that this level of proficiency is far behind NP3.

So, comparing PL3 and PL4, it is possible to conclude that, concerning the vocabulary descriptors of the rating scale, what differentiates a controller PL3 and a PL4 is the ability to use the vocabulary to communicate effectively on common, concrete, and work-related topics in a *usually sufficient way, with appropriate lexical range and accuracy*. Another important aspect is the ability to *often paraphrase*

successfully when lacking vocabulary in unusual or unexpected circumstances. This distinction is not always so clear, for 'usually' and 'often' are sometimes difficult to measure during an interview, but it is crucial, considering that PL4 is allowed to operate with international traffic and PL3 is not – a high-stake decision with many important consequences – people's lives, ultimately.

Römer (2017) questions the traditional separation between lexis and grammar on rating scales. The author claims that reasons for this separation come from a structuralist view of language testing researchers' understanding of language proficiency. According to her:

> More recent models of language ability, including the influential model of Bachman and Palmer (1996, 2010), continue this separation of lexis and syntax as distinct aspects of "grammatical knowledge", separating these aspects of language ability from knowledge of language functions, which is subsumed under "pragmatic knowledge". Based on this view of language, many influential rating scales in language testing have traditionally treated lexis and grammar separately. (RÖMER, 2017, p. 478).

On the other hand, Römer (2017) argues that recent integrative and functionally oriented approaches to language learning have a more holistic approach to language proficiency, considering lexico-grammatical knowledge as a single category. According to her, CL offers an important contribution to this view that the phrase, rather than the individual word, is the fundamental unit of language, and that a great deal of communication consists of fixed expressions that defy simple categorization into either vocabulary or grammar. This approach is beginning to be considered in the area of language assessment too. While a major problem with many rating scales is that their descriptors are not based on analyses of empirical linguistic evidence but come from intuitive judgments, "corpus studies of lexico-grammar provide such empirical evidence that may be useful in informing the development, validation, and use of rating scales for speaking assessment" (RÖMER, 2017, p. 478).

We share this view and believe that words have to be analyzed in context, so that different forms of concept or meaning expression can be identified, as we have discussed in this section, and seems to be unanimous in contemporary studies of lexical semantics, terminology, CL, teaching and assessment. We hope that this work can offer a small

contribution to spread this view to high-stakes speaking tests, as is the case of aeronautical English assessment.

Starting from weather situations considered extremely relevant to ATC or that could be problematic or cause confusion, presented in Doc 9835 (ICAO, 2010) and in the Reference Corpus, based on our experience in the teaching of aeronautical English for over 10 years, and also on data presented in research carried out by Tosqui-Lucks and Prado (in press), our procedure was to investigate the use of the 11 selected terms and their variations in the three subcorpora to analyze how these terms are used by students, considering the vocabulary descriptor of ICAO rating scale.

Thus, our approach to CL is predominantly corpus-based and data-informed, since we look for particular linguistic characteristics already pre-established in the corpora and analyze the data found to verify the occurrences, frequencies, concordances and relevant information to understand the phenomena (RAYSON, 2008). At other times, we also follow the data-driven approach, when, for example, we observe the most frequent words generated in the Wordlist tool, or even the words with wrong spelling, which can be an indication of a pronunciation problem (as in 'mantein', 'turbulance'), vocabulary (as in 'dicende', 'buid-up') or even when the choice of words is wrong, as in the case of 'approximation' by 'approach' or 'alternative' by 'alternate', as we will explain later. In order to detect problems in the students' production, we did not correct anything in the corpus, we kept the spellings exactly as in the original. This decision forced us to carefully analyze the occurrences to see whether or not the same word was written in different ways, as in the case of 'condictions', 'wheather', 'confirme', 'intencions'. The details of the methodology are presented in the following section.

3 Methodology

As mentioned before, this paper has two phases: the first one, based on lexical semantics applied to terminology, to analyze formulaic structure of lexical units using Aeronautical Meteorology terms within the ATC context; and the second one, to analyze the use of these terms by students in three ATC courses (for TWR, ACC and APP facilities) and how it affects their performance during communication activities in a learning environment. For that, we selected some key aeronautical meteorology (AER MET) terms particularly used in ATC phraseology,

and studied their lexical semantic relations in a reference corpus of ATC international and Brazilian standards; and in a learner corpus of air traffic control communication in learning situations. The key terms were selected based on their relevance in the corpora, as related to ATC communication, and the following ones were extracted: (1) rain, (2) wind, (3) wind shear, (4) turbulence, (5) wake turbulence, (6) conditions, (7) lightning, (8) formation, (9) cloud, (10) fog, and (11) thunderstorm. It is important to highlight that the occurrence[4] of those words are related to the relevance of some meteorological phenomena in air traffic situations applied to the Brazilian context, in terms of occurrence and frequency of some weather events.

The software used for corpora analysis is AntConc (ANTHONY, 2019), a freeware corpus analysis toolkit for concordancing and text analysis, chosen because its interface is simple, very user-friendly and provide adequate tools for the purposes of this paper. In addition to that, Anthony (2019) provides many downloadable guides and video tutorials on the software website that may guide unexperienced teachers into the basics of CL analysis. The software also enables the use of regular expression (REGEX) commands, as a way to extract terms which have spelling variation or are misspelled in the learner corpus, as in the case of variations 'wind shear' and 'windshear', and misspelled occurrences of 'wind sheer'. In this way, we believe – and hope – that aeronautical English researchers and teachers can be inspired by our ideas and try to use a similar methodology to compile a learner corpus with the production of their own students.

The architecture of our corpora (reference corpus and learner corpus) is described in the table 5.

[4] As our aim in this paper is analyzing discourse patterns concerning aeronautical meteorology terms used within the air traffic control context, 'occurrence' of terms refer to different instances of use of a term, i.e. the exact same instance of use was not counted as another occurrence. For example, in spite of the fact 'heavy rain' appears many times in the learner corpus, this was only considered one occurrence; but 'moderate rain', even though similar in structure, was considered another occurrence of the term 'rain'.

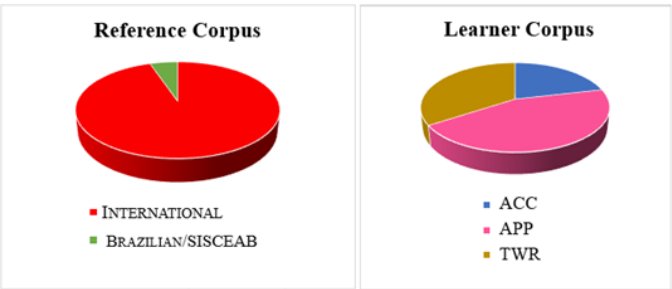
TABLE 5 – Corpora architecture (word types and word tokens)

CORPUS	SUBCORPUS	WORD TYPES	WORD TOKENS
REFERENCE CORPUS (ATC PHRASEOLOGY IN ENGLISH)	INTERNATIONAL	11119	581414
	BRAZILIAN/SISCEAB	3331	32202
LEARNER CORPUS (IN ENGLISH)	ACC	1052	13258
	APP	1763	27559
	TWR	1520	21249

Source: Authors' own elaboration.

The proportion of each subcorpus in the corpora composition is best represented in the figure 1.

FIGURE 1 – Corpora Architecture (proportional comparison)



Source: Authors' own elaboration.

As indicated in Table 5 and in Figure 1, the Reference Corpus is composed of ATC phraseology publications in English, set as standards by the international organizations ICAO, WMO and FAA, and by the Brazilian authority DECEA (SISCEAB system); and the Learner Corpus is composed of learning situations carried out during courses applied to the context of Area Control Center (ACC), Approach Control Center (ACC), and Tower (TWR).

The ATC phraseology publications compiled for the reference corpus are Doc 4444 (ICAO, 2016), Annex 3 (ICAO 2018), Doc 732 (WMO 2003), ORDER JO 7110.65W (UNITED STATES, 2015), MCA

100-16 (BRAZIL, 2018) and ICA 105-12 (BRAZIL, 2014). They were selected because they are guidelines which specifically address the use of phraseology within the ATC context, as published by official institutions dealing with aviation regulations also comprising meteorological instructions: the International Civil Aviation Organization (ICAO), the World Meteorological Organization (WMO), the Federal Aviation Administration (FAA, United States) and the Department of Airspace Control (DECEA, Brazil).

In this sense, Doc 4444 (ICAO, 2016) prescribes rules for Air Traffic Management; Annex 3 (ICAO, 2018) focuses on guidelines for the provision of Meteorological Service for International Air Navigation; Doc 732 (WMO, 2003) is a Guide to Practices for Meteorological Offices serving Aviation; Order JO 7110.65W (UNITED STATES, 2015) is an Air Traffic Organization Policy on phraseology and procedures; MCA 100-16 (BRAZIL, 2018) is the institutional documentation for ATC Phraseology within the Brazilian Airspace Control System (SISCEAB);[5] and ICA 105-12 (BRAZIL, 2014) prescribes VOLMET Phraseology to be used in the SISCEAB system as well. As it can be visualized in Figure 1, the Brazilian/SISCEAB subcorpus is much shorter because it mostly comprises ATC phraseology used within Brazilian specific situations, by following standardized phraseology in English, originally prescribed by ICAO and WMO.

Regarding the learner corpus, it was compiled from evaluated activities that are part of a series of distance learning courses offered to Brazilian Controllers, called "Go4it". There are three different courses: for area control center (ACC); approach control (APP) and tower (TWR). In each activity, the student must record an audio about the topics studied on that module, followed by the respective script. Since the activities were produced by students, it is only natural that they make mistakes. We opted for using the scripts with errors, not the versions corrected by the teachers, because the corrections could affect the results. So, we kept the problems with spelling, grammar or vocabulary. Considering that the courses have emphasis on speaking and not writing, some students do not worry too much about reviewing spelling mistakes on the scripts, because they will be graded mostly for their oral performance.

[5] In Portuguese, SISCEAB stands for "Sistema de Controle do Espaço Aéreo Brasileiro".

The three subcorpora used in the paper correspond to the scripts of the "Weather events" module of the three courses. The learners are in-service controllers, male or female, military or civil employees of Brazilian Air Force, enrolled in the courses offered from 2015 to 2018. Most of them have PL3 according to ICAO rating scale, but some have PL4 and need to revalidate their level, what occurs every 3 years.[6]

Each course lasts 8 weeks and comprises 8 modules, being the first one introductory (*Getting Started*) and seven of specific content: *Air Communication, ATC Jobs, Medical Emergencies, Parts of the Aircraft, Phases of Flight, Operational Events,* and *Weather en route*. Among the specific content modules, only five reproduce pilot-controller communications, and were compiled in the learner corpus: *Operational Events, Air communication, Phases of Flight, Medical Emergencies and Weather Events*. The other modules offer different kinds of oral activities, such as reporting a real situation or telling a story based on pictures. Having explained the compilation process of the three learner subcorpora, we will hereafter refer to it simply as "learner corpus" for the sake of this article, as contrasted to the "reference corpus".

The compiled reference corpus was used in phase 1, and the learner corpus was used in phase 2; and the analysis focused on studying collocates (in the lexical semantics theory, it is called 'combinatorics') of each main term as listed according to a 3L-3R parameter, from which the first 50 ranked were analyzed. Then, we focused on left and right combinatorics of terms, and also associative patterns (relations), to proceed to a lexical semantic analysis (L'HOMME, 2020) by attributing semantic labels (PEIXOTO; PIMENTEL, 2020), and discourse patterns were discussed based on occurrences in the corpora. In addition to that, phase 2 approached language difficulties of learners, according to ICAO descriptors discussed in item 2.3 of this paper.

The methodology design for the work carried out in this paper is summarized in Table 6 as follows.

---

[6] The learner corpus was compiled within an ATC military organization and its use is allowed only for previously authorized research, because of national safety reasons. In order to follow the recommended practices of the Committee on Publication Ethics, students signed a term of consent agreeing on the use of the data collected from their production within the course for research purposes, regarding that their identities are preserved.

TABLE 6 – Methodology Design

#	Activity	Description
01	Corpora compilation (reference corpus and learner corpus)	Compilation of publications on ATC phraseology comprising aeronautical meteorology situations, from official institutions (reference corpus) and from activities in a learning environment (learner corpus).
02	Extraction of key terms	Generation of wordlists and extraction of 11 terms related to weather situations which are critical for air traffic operations.
03	Analysis of discourse patterns in air traffic control phraseology standards (Phase 1)	Analysis of the formulaic structure of lexical units using 11 Aeronautical Meteorology terms within the ATC context, by studying left and right combinatorics of AER MET terms as appearing in the reference corpus.
04	Analysis of air traffic control communication in Aeronautical English courses (Phase 2)	Analysis of language structure as produced by students in classes of air traffic communication for Area Control Centers (ACC), Approach Control Centers (APP) and Towers (TWR), based on ICAO descriptor vocabulary of language assessment.

Source: Authors' own elaboration.

4 Weather events in air traffic control phraseology standards: discussion of discourse patterns

Meteorological conditions affect a varied range of air traffic control situations, related not only to en route events but also to air traffic operations particularly during landing/approach and take-off procedures. Runway conditions partly depend on meteorological conditions, especially when it comes to water effects leading to runway contamination, in addition to specific traits of the runway, which makes it more prone to water accumulation or not.[7]

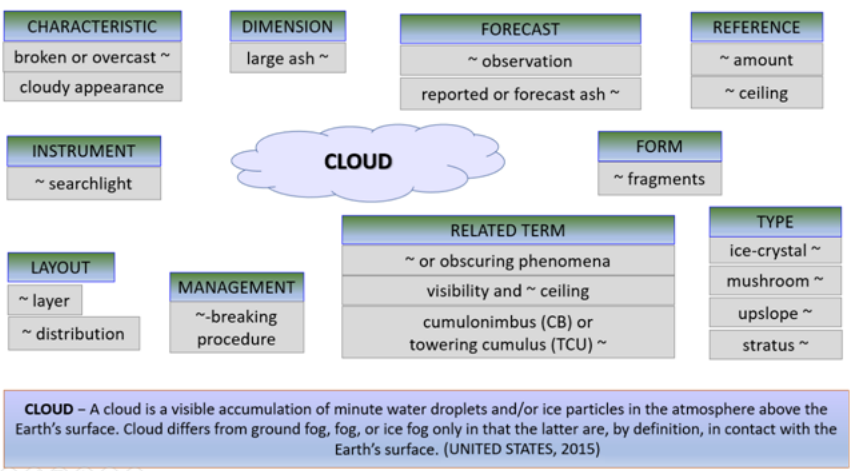
[7] The International Civil Aviation Organization (ICAO) has recently issued some guidelines to address the types of runway contamination: the New Global Reporting Format (GRF) for Runway Surface Conditions (2019), based on the Takeoff and Landing Performance Assessment (TALPA) model issued by the Federal Aviation Administration (FAA) in 2016. ICAO Member States were demanded to implement the GRF grid assessment by November 2020; however, due to the COVID-19 pandemics, the deadline was postponed to November 2021. More information on GRF guidelines can be found at <https://www.icao.int/safety/Pages/GRF.aspx>.

By considering this holistic panorama, the WMO (2003) classifies possible aviation hazards as: **(a) in-flight hazards** such as icing, turbulence, lightning and volcanic ashes; **(b) hazards in the phases of approach and take-off,** including wind shear effects, turbulence, convective activity and freezing precipitation on aircraft; **(c) weather hazards affecting the acceptance capability of hub airports**, considering capacity for de-icing, and runway and apron snow clearance; **(d) weather hazards affecting the capacity of air routes**, such as mesoscale convective systems, volcanic ash and severe turbulence; and **(e) weather hazards affecting ground operations, passenger ground transportation and safety**, resulting from lightning, strong winds or hail, for example. In this sense, the Annex 3 (ICAO, 2018, p. 4-5) mentions that minimum present weather phenomena to be identified at airports, to enable safety of operations, are "rain, drizzle, snow and freezing precipitation (including intensity thereof), haze, mist, fog, freezing fog and thunderstorms (including thunderstorms in the vicinity)."

By considering this complex panorama of weather events affecting air traffic operations, discourse patterns were analyzed by following the lexical semantic approach discussed in item 2.1.

To illustrate how labels were attributed to collocates of a term, Figure 2 presents the semantic profile of the term 'cloud', by showing lexical semantic occurrences with this term and their respective semantic labels.

FIGURE 2 – Semantic profile of the term 'cloud'



Source: Authors' own elaboration.

In the reference corpus, the definition for 'cloud' was found in Order JO 7110.65W (UNITED STATES, 2015), in the glossary listed at the end, as well as definitions of two other selected terms, described in the following way:

> WAKE TURBULENCE − Phenomena resulting from the passage of an aircraft through the atmosphere. The term includes vortices, thrust stream turbulence, jet blast, jet wash, propeller wash, and rotor wash both on the ground and in the air.
>
> WIND SHEAR − A change in wind speed and/or wind direction in a short distance resulting in a tearing or shearing effect. It can exist in a horizontal or vertical direction and occasionally in both
>
> CLOUD − A cloud is a visible accumulation of minute water droplets and/or ice particles in the atmosphere above the Earth's surface. Cloud differs from ground fog, fog, or ice fog only in that the latter are, by definition, in contact with the Earth's surface. (UNITED STATES, 2015).

Concerning 'wind shear', there was another related term ('wind shear escape'), with a more specific meaning:

> WIND SHEAR ESCAPE − an unplanned abortive maneuver initiated by the pilot in command (PIC) as a result of onboard cockpit systems. Wind shear escapes are characterized by maximum thrust climbs in the low altitude terminal environment until wind shear conditions are no longer detected. (UNITED STATES, 2015).

However, occurrences for 'wind shear escape' were not so prolific, as the only collocates found were '~ complete', '~ maneuver', and '~ procedures', so 'wind shear escape' was not classified independently.

Regarding the profile of semantic labels for each term, Table 7 shows the total of lexical semantic occurrences and the total of labels, and also compares the semantic density of the selected terms, by calculating the total of labels per total of occurrences.

TABLE 7 – Profile of semantic labels for each term in the reference corpus

Term	Total of occurrences	Total of labels	Semantic density
rain	11	4	36%
wind	22	9	41%
wind shear	15	5	33%
turbulence	22	7	32%
wake turbulence	18	8	44%
conditions	24	8	33%
lightning	6	3	50%
formation	4	3	75%
cloud	55	10	18%
fog	17	5	29%
thunderstorm	11	5	45%

Source: Authors' own elaboration.

The occurrences of the selected terms often come as 'ADJECTIVE + TERM' or 'NOUN + of + TERM' / 'TERM + of + NOUN' (perception of ~; possible effects of ~; ~ of great vertical extent; ~ of operational significance) or adverbial structures such as '~ around the periphery of an airport'. In addition to that, it is interesting to note there were some few hyphenated constructions such as '~-breaking procedure' and '~-prone areas'; and there were also passive structures such as 'partially covered by ~' and 'algorithmically derived ~ warnings', '~ networks to ATS'.

When it comes to the productivity of semantic labels, RELATED TERMS are the most common, with 82 occurrences, then TYPE and LAYOUT, with 23 and 22 occurrences. The labels DURATION, INTENSITY and MOVEMENT only had one occurrence each. The most diverse terms were 'formation' and 'lightning', accounting for 75% and 50% of semantic density, respectively. And 'cloud' and 'fog' are the most uniform terms, i.e., with less semantic variation, of only 18% and 29% respectively.

However, in the case of 'cloud', there were many occurrences of the semantic label TYPE (9), REFERENCE (7) and LAYOUT (7), of more objective nature. 'Fog' also had a more objective standard, with prevalent occurrences of CHARACTERISTIC (3) and LAYOUT (3) too. Regarding 'lightning' and 'formation', with more density, 'formation' has a more objective profile (TYPE semantic label is prevalent) while 'lightning'

showed a more procedural perspective to terms being used, with two occurrences of the label INSTRUMENT.

'Cloud' also showed major label variation (10 out of 18 classified in this paper were applied), as well as 'wind' (9 labels), indicating higher relevance of those terms to the field of aeronautical meteorology. As a matter of fact, the World Meteorological Organization (WMO, 2003) states that "the primary forecast elements are the surface wind, visibility, weather and cloud." (p. 13). In this line, the International Civil Aviation Organization (ICAO, 2018) states important weather information related to aviation as "information on visibility, runway visual range, present weather and cloud amount, cloud type and height of cloud base" (p. 4-6). This is convergent to the previously discussed perspective of weather influence to runway conditions, as a direct product of weather phenomena (Cf. ICAO, 2018; WMO, 2003).

In addition to these findings, some interesting cases have to be highlighted and discussed, as shown in the reference corpus. Regarding 'conditions', it is interesting to note that this term was used a significant number of times in the text with the meaning of possibility or objective condition of air traffic elements (surface conditions; and conditions, such as workload, traffic volume, the quality/limitations of the radar system) not related to weather phenomena. In this paper, however, occurrences were selected only when 'conditions' referred to general standards of atmospheric phenomena, not conditions as status of equipment, for example. The polysemy shown here, however, stresses the possible nuance of terms, only clarified when considering the contextual reference to collocates. As discussed at the beginning of this paper, defining the whole scope of pattern of a term is always a very sensitive task. Although runway conditions may be indeed related to meteorological phenomena it does not constitute a weather situation in itself since it is not an ongoing process, but the product or result of a previous meteorological condition.

Regarding 'lightning', there were some occurrences of 'blue lightning event', particularly in Order JO 7110.65W (UNITED STATES, 2015), but with a different meaning when compared to primary concept of 'lightning' within the aeronautical context. As a matter of fact, 'blue lightning events' refers to "reports of possible human trafficking". As publicized in the website of the U.S. Department of Transportation, the Department to which FAA belongs, this expression is explained as:

> The Blue Lightning Initiative (BLI), led by the Department of Transportation, the Department of Homeland Security, and U.S. Customs and Border Protection, is an element of the DHS Blue Campaign. The BLI trains aviation industry personnel to identify potential traffickers and human trafficking victims, and to report their suspicions to federal law enforcement. To date, more than 100,000 personnel in the aviation industry have been trained through the BLI, and actionable tips continue to be reported to law enforcement. (UNITED STATES, 2020).

As lightnings may pose major threats to air traffic operations, airports use human observation as well as specific detection equipment to support weather phenomena analysis. Annex 3 (ICAO, 2018) informs that

> At aerodromes with human observers, lightning detection equipment may supplement human observations. For aerodromes with automatic observing systems, guidance on the use of lightning detection equipment intended for thunderstorm reporting is given in the Manual on Automatic Meteorological Observing Systems at Aerodromes (Doc 9837). (ICAO, 2018, p. APP 3-13)

The use of 'formation' is quite often related to aircraft arrangement (join-up and breakaway) during performed flights, generally conducted in VFR weather unless otherwise approved, as indicated in Order JO 7110.65W (UNITED STATES, 2015). The few cases where formation is used in the context of weather phenomena is when referring to 'formation/cell operations' and 'formation/cell envelope'.

A related term is 'build-up', which is shown in broader aviation literature of WMO and ICAO as generally referring to some accumulation of substances as water, snow or ice ('build-up of ice', and 'ice build-up', 'water build-up'); or accumulation of some sort of reaction, such as 'build-up of static electricity'. In other situations not related to weather phenomena, 'build-up' is also used in the sense of evolution of services or operations as in 'build-up of services' and 'volcano build-up to an eruption'. This latter sense is more related to a general sense of "an increase, especially one that is gradual" or "an increase in the amount of something over a period of time", as indicated in the Cambridge Dictionary.

The term 'cloud' is the one showing most interesting lexical semantic associations. 'Cloud' is the general term comprising specific types of cloud, such as CB (Cumulonimbus) or TCU (Towering Cumulus)

clouds, which is often used independently as well (as CB or TCU only, without the word 'cloud'). Occurrences of CB in the corpora mostly refer to meteorological codes to be used in forms and systems; and there are occurrences of 'cumulonimbus CB', and its variation 'cumulunimbus CB', only found in ICA 105-12 (BRAZIL, 2014).

It is important to highlight that the analysis carried out in this paper did not intend to find overall patterns for these terms but phraseological patterns in the ATC language prescribed in the compiled reference corpus, which addresses ATC and weather situations.

This terminological perspective contributes to deepen understanding on how terms work and confirm the perspective that they are very inter-related to adjectival patterns, which require more emphasis on adjective order for example, as well as specific adjectives to be collocated with related nouns, as more broadly discussed in the next section.

5 Weather events in air traffic control communication in learner corpus: discussion and implications for Aeronautical English courses

When it comes to the learner corpus, it is important to emphasize that language patterns may vary a little due to the fact it is a controlled learning environment. For example, related terms are much higher in this learner corpus regarding more common weather phenomena such as rain (17), wind (6) and lightning (10), and how they are associated to other phenomena or situations such as 'runway', 'fog', 'gust', 'hailstones', 'lightning', 'CB', 'tailwind', 'thunderstorm', 'turbulence', 'visibility', 'wind', 'wind shear', 'thunderstorm' and 'instrument conditions', in the case of 'rain'; 'rough chop', 'position', 'rain', 'temperature' and 'visibility', in the case of 'wind'; and 'rain', 'thunderstorm', 'turbulence', 'hailstones', 'electrical failure', 'CB', 'flashflood', 'engine', 'visibility' and 'runway lights', in the case of 'lightning'.

Overall, adjectives played an important role in the usage of expressions containing these terms in the learner corpus. For example, 'heavy' was the most common collocate with terms analyzed, especially with 'rain'. Among those, some adjective occurrences reflect linguistic calque, as in the case of 'strong' used instead of 'heavy': 'strong rain' in place of 'heavy rain'. However, occurrences such as 'weak rain' are not

contained in the learner corpus. In that sense, most of those adjective uses are intensifiers, with other occurrences with 'dense', 'intense', 'light', 'moderate' and 'severe'.

As the learner corpus is also representative of the aeronautical language in use, it also contains more verbs, due to the intent to comprise more situated communication, with higher reference to location as well. In our study, there is a varied range of verbs which were used with 'turbulence' and 'lightning', a pattern which was not specifically explored in the semantic labels in this paper but is relevant to be mentioned. In the case of 'turbulence', verbs such as 'passing through', 'flew through', 'went through', 'passed through', 'suffering', 'facing' and 'experiencing' were used in many instances and also indicate some level of interference from Portuguese. For 'lightning', verbal constructions were mostly based on verbs 'strike' and 'hit', in both active and passive voices, with constructions such as [VERB IN PASSIVE VOICE + DIRECT OBJECT]; [VERB IN PASSIVE VOICE + INDIRECT OBJECT]; [VERB IN ACTIVE VOICE + DIRECT OBJECT]; AND [VERB IN ACTIVE VOICE + INDIRECT OBJECT]. Some examples are 'striked[8] by a ~'; 'a strong ~ struck the engine'; 'a strong ~ struck us''; a ~ has struck us'; 'a ~ has struck my left engine'; '~ stroke our landing equipment'; 'hit by a ~'; 'a ~ hit our left wing'; 'a ~ hit us'; 'we were hit/struck by a ~'; and 'I had my right wing hitted for a ~ strike'. The consequences are sometimes reported and usually related to some kind of technical failure as in "We was hit for a lightning strike and had an electric system failure".

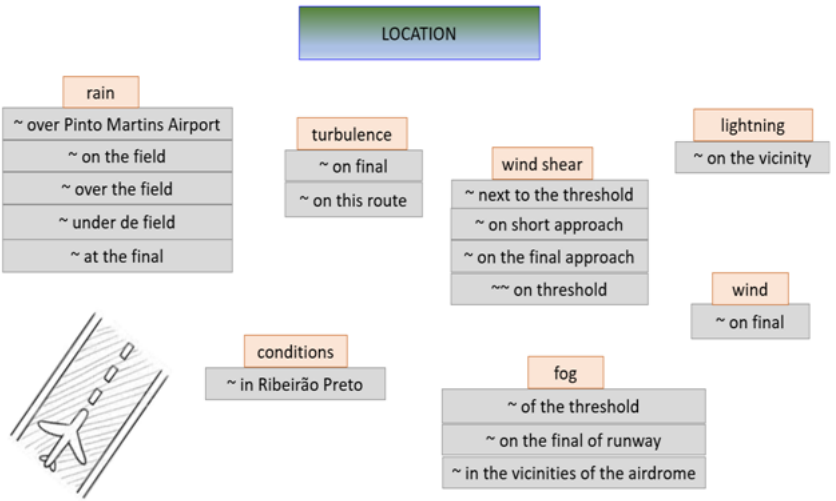
Concerning the term 'conditions', likewise in the reference corpus, there are occurrences which are directly related to meteorological phenomena and some others which comprise a broader scope regarding runway conditions. There is one special example which is in the "crossroads" of this differentiation: instrument meteorological conditions (IMC) and instrument flight rules (IFR) conditions, both found in the learner corpus. While IMC literally mentions the meteorological factor, IFR focuses on the use of instrument rules, applied in cases when the airport has such poor weather conditions that it is necessary to rely more

[8] As mentioned in the methodology, we did not correct students' grammar errors. In these examples, the incorrect forms '*striked'* and '*stroke'* were used by students instead of the correct form '*struck'*. We will not refer to the grammatical correct form of other examples.

heavily on flight instruments to be able to land the aircraft. This example is particularly interesting to emphasize how meteorological phenomena affect many air traffic situations as a whole, as described in WMO (2003) when mentioning weather hazards.

Regarding the semantic label 'location', it was also found in a high number in the learner corpus, as a clear indication of fine-grained weather report during all phases of flight, as it happens in 'rain on final (approach)', 'rain on over the field', 'rain over the field', 'rain over Congonhas', 'over the Guamá river', 'rain under the field', 'rain in the threshold on the runway', 'rain is approaching the aerodrome', 'over the airdrome, and 'over the runway'. Some of these occurrences for each term are illustrated in the following figure.

FIGURE 3 – Occurrences of the semantic label LOCATION



Source: Authors' own elaboration.

A summary of lexical semantic occurrences and labels for each key term analyzed in this paper, and the corresponding semantic density, is shown in the table 8.

TABLE 8 – Profile of semantic labels for each term in the learner corpus

Term	Total of occurrences[9]	Total of labels	Semantic density
rain	29	5	17%
wind	22	10	45%
wind shear	14	5	36%
turbulence	25	8	32%
wake turbulence	1	1	100%
conditions	13	7	54%
lightning	17	5	29%
formation	11	8	73%
cloud	9	6	67%
fog	18	6	33%
thunderstorm	11	6	55%

Source: Authors' own elaboration.

As indicated in the table, terms 'rain', 'lightning', 'turbulence' and 'fog' have the lowest semantic density, with 17%, 29%, 32% and 33%, respectively; and 'formation' and 'cloud' have the highest diversification of semantic labels, accounting for 73% and 67% respectively.[10]

When it comes to didactic applications of aeronautical meteorological terms, there are some interesting aspects to note. Some uncountable nouns as 'rain' are used as countable nouns, with the inclusion of indefinite article, as in "We are undergoing a formation and facing a heavy rain". In the same way, an indefinite article is also used in "We received in the short end a tailwind with 15 knots" and "We have a electrical failure due to a lightning, strike the airplane". Sometimes the article may be used or omitted, as in "We are facing a thunderstorm on FL180." and "There is Thunderstorm over Porto Velho Airdrome, pay attention.".

In aeronautical communication, it is paramount to provide sensitive information on weather (ICAO, 2010). The importance of

[9] Wake turbulence' was not taken into account because there was only one occurrence, then semantic density was 100%.

[10] 'Wake turbulence' was not taken into account because there was only one occurrence, then semantic density was 100%.

warning pilots regarding these meteorological conditions is present in some excerpts in the corpus, as in the listed occurrences:

> Attention, the runway 13 is slippery.
> (2) Fortaleza is below minimum VFR due to bad weather, heavy rain. Caution, for your information the aircraft has just landed before said he went through a chop on final and other one reported a windshear on short final.
> (3) Fortaleza is operating IFR conditions below minima due to heavy rain, the last aircraft that landed, reported thunderstorm with lightning strike, when he was 3nm out.
> (4) Rain and low visibility on final, alternate to Manaus airport.

In some examples, other consequences or damages are also informed as in:

> we are facing severe turbulence and we are loosing oxigen.
> (2) we are in severe turbulence and we have lost the weather radar.
> (3) Right winglet was broken due to severe turbulence/ has some damage, probably caused by the turbulence /turbulence and one part of my cargo broke.
> (4) We got severe turbulence, shaking too much.

Another very important aspect to describe the weather conditions is gradation (from very bad to good), as in "Waiting more than 15 minutes for a better weather condition", "weather conditions become better to runway 24, few clouds / standby" and "keep hold at this position waiting for weather conditions to improve". The communication regarding the criticality level of weather leads to requests by the controllers, such as "descend", "divert", "immediate descent" and "descend immediately" or required actions such as "avoid turbulence". Sometimes modal verbs (must / need / will) are also used for that, as in "turbulence. I need descend now", "turbulence I need divert to my alternative airport" "turbulence. We must land on the nearest AD" and "turbulence. We will need a firefighter, because we…"

When taking into consideration language patterns of the learner corpus, it is important to list all relevant occurrences in order to foresee possible mistakes and try to find regularities to assess those mistakes properly (RAYSON, 2008). Particularly in the learner corpus there are some other occurrences/mistakes which are relevant for a learning environment but which is not a general language pattern in

standard communication (RÖMER, 2004; PEIXOTO, 2020). Verbs and prepositions seem to add to this, especially in terms of crosslinguistic interference/variation (linguistic calque).

If we compare the semantic labels occurring in the reference corpus and in the learner corpus, it is possible to notice there are some peculiarities regarding language patterns (Table 9).

TABLE 9 – Semantic labels in the reference corpus and in the learner corpus

#	Semantic Label	Reference corpus	Learner corpus
01	CHARACTERISTIC	12	1
02	CHARACTERISTIC / INTENSITY	10	16
03	DIMENSION	6	2
04	DURATION	1	2
05	EPISODE	3	4
06	EPISODE / INTENSITY	-	3
07	FORECAST	12	10
08	FORM	1	-
09	INFORMATION FACTOR	2	2
10	INSTRUMENT	7	2
11	INTENSITY	1	2
12	LAYOUT	23	10
13	LOCATION	-	20
14	MANAGEMENT	4	-
15	MOVEMENT	1	-
16	PARAMETER	8	2
17	PHENOMENON	-	4
18	REFERENCE	9	11
19	RELATED TERM	82	59
20	TYPE	21	15
21	TYPE / DIMENSION	-	1
22	TYPE / INTENSITY	-	1
23	UNIT OF MEASUREMENT	-	3
24	VARIATION	2	-

Source: Authors' own elaboration.

In the learner corpus, there was more variation of semantic labels, including EPISODE / INTENSITY, LOCATION, PHENOMENON, TYPE / DIMENSION, TYPE / INTENSITY and UNITS OF MEASUREMENT. This can be explained by the nature of reported communication, giving specific details on where and how weather phenomena are taking place. This finding is mostly suggested, on the one hand, by the higher occurrence of LOCATION labels, in a total of 20 occurrences regarding all terms in the learner corpus; and, on the other hand, by the fact the semantic label MANAGEMENT does not appear in the learner corpus, along with the absence of semantic labels FORM and VARIATION.

6 Final Remarks

Terminological patterns discussed in this paper show how meaning is dependent on context, and how lexical semantic analysis of terms may contribute to reveal nuances of language used in a specialized language. Likewise, this approach also contributes to deepen understanding of language used by students, especially regarding the descriptor vocabulary, prescribed in ICAO rating scale.

However, it is important to stress that analyses carried out in the reference corpus as compared to the learner corpus are illustrative, since occurrences in the learner corpus are controlled and depend on other variables beyond proportional occurrences in natural language expression. Findings suggest learner corpus language focuses on occurrences which are found to be related to more common daily situations, especially within the Brazilian context; and, based on that, semantic density in both corpora is not expected to be the same.

Therefore, results show that the courses have been efficient in teaching and practicing the use of the main meteorological terms related to aeronautical English and that, despite some mistakes students make, evidence indicates that they are able to report weather conditions to pilots and to understand pilots' requests in a proficient level concerning vocabulary. As we've mentioned before, we believe in a more integrated analysis of language production by students, considering the context and the blocks of unit instead of looking at isolated words. In this sense, CL is an efficient tool for analyzing the production of groups of students.

Concerning implications for teaching, there are many analyses that can be conducted by a teacher using the resources of CL. The software used in this paper is free and easy to use with little training – tutorials are widely available. For existing courses, which is the case here, by looking at the concordance lines is possible to compare students' use of the terms, their collocates, the context of use and adjust instruction if necessary. It is possible to monitor a student's development and address him/her individually. It is also possible to apply a data-driven approach and, by showing concordance lines to students, raise their awareness in relation to misuses of a term or the most frequent collocations of it, contrast its use in the learner corpus and the reference corpus, among others. The results presented here can also help researchers and material designers collect authentic descriptions of language usage in a learning environment which, in their turn, may improve teaching and reference materials. This is especially important in the case of aeronautical English, since there are not many courses or material available in the market that deal with specific needs of Brazilian air traffic controllers.

As for implications regarding language testing, we hope this kind of analysis helps teachers benchmark their students' performance in relation to what is expected to NP4 according to the ICAO rating scale. Results also advocate in favor of a more integrated scale and could be used as an argument for ICAO revision of its 16-year-old rating scale. A follow-up suggestion for future research would be to analyze the results of the same students who took the three courses from where the subcorpora were compiled at EPLIS, to check their performance in weather-related tasks and if they achieved PL 4 or above in the descriptor vocabulary, but this is beyond the scope of this paper.

Declaration of contribution

Rafaela Rigaud Peixoto wrote the theoretical foundation section on phraseological patterns and lexical semantic terminological approach, and contributed to the introduction section on aeronautical English. Regarding the methodological planning of the paper, she developed the methodology design, compiled the reference corpus, and articulated the methodological procedures of phase 1 and phase 2. Rafaela performed lexical semantic analysis of weather events in air traffic control phraseology standards, and of weather events in air traffic control communication in learner corpus,

regarding phraseological structures and semantic labels of combinatorics of the 11 selected terms, by using *AntConc* concordancing software. She wrote the abstract in Portuguese and in English. Patrícia Tosqui-Lucks wrote the introductory concepts of aviation and aeronautical English; phraseology and plain English; and ICEA responsibilities concerning teaching and assessment of Brazilian ATCOs. As for the theoretical foundations, she wrote the discussion about learner corpora and ICAO proficiency requirements, including the rating scale descriptors for vocabulary assessment. As for the methodology, Patrícia inserted the data of the learner corpus she compiled into the *Antconc* software and contributed to the analysis of language structure of the 11 selected lexical items as produced by students in classes of air traffic communication for Area Control Centers (ACC), Approach Control Centers (APP) and Towers (TWR).

Acknowledgment

This study was financed in part by the *Coordenação de Aperfeiçoamento de Pessoal de Nível Superior* – Brasil (CAPES) – Finance Code 001.

References

ANTHONY, L. *AntConc (Version 3.5.8)* [Computer Software]. Tokyo: Waseda University, 2019. Available from: https://www.laurenceanthony.net/software. Access on: August, 2020.

BARTNING, I.; FORSBERG, F. Les séquences préfabriquées à travers les stades de développement en français L2. *In*: CONGRÈS DES ROMANISTES SCANDINAVES, 16e., 2006, Roskilde. *Actes* […]. Roskilde: Department of Language and Culture, Roskilde University, 2006. p. 1-22.

BERBER-SARDINHA, T. Como usar a linguística de *corpus* no ensino de língua estrangeira – por uma linguística de *corpus* educacional brasileira. *In:* VIANA, V.; TAGNIN, S. E. O. (org.). *Corpora no ensino de línguas estrangeiras*. São Paulo: HUB Editorial, 2011. p. 301-356.

BRAZIL. Comando da Aeronáutica. Departamento de Controle do Espaço Aéreo. *ICA 105-12:* Fraseologia Volmet. Rio de Janeiro, 2014. Available from: https://publicacoes.decea.gov.br/?i=publicacao&id=4072. Access on: Sep. 20, 2019.

BRAZIL. Comando da Aeronáutica. Departamento de Controle do Espaço Aéreo. *MCA 100-16:* Fraseologia de Tráfego Aéreo. Rio de Janeiro, 2018. Available from: https://publicacoes.decea.gov.br/?i=publicacao&id=4072. Access: Sep. 20, 2019.

CHEN, M. Phrasal Verbs in a Longitudinal Learner Corpus: Quantitative Findings. *In:* GRANGER, S.; GILQUIN, G.; MEUNIER, F. (ed.). *Twenty Years of Learner Corpus Research:* Looking Back, Moving Ahead. Corpora and Language in Use - Proceedings 1. Louvain-la-Neuve: Presses universitaires de Louvain, 2013. p. 89-101.

EBELING, S. O.; HASSELGÅRD, H. Learner Corpora and Phraseology. *In:* GRANGER, S.; GILQUIN, G.; MEUNIER, F. (ed.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 2015. p. 207-230.

FINATTO, M. da G. K. *Definição terminológica*: fundamentos teórico-metodológicos para sua descrição e explicação. 2001. 395f. Tese (Doutorado em Estudos da Linguagem) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2001.

GAVIOLI, L. *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins, 2005.

GILQUIN, G. From Design to Collection of Learner Corpora. *In:* GRANGER, S.; GILQUIN, G.; MEUNIER, F. (ed.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 2015. Digital version. DOI: 10.1017/CBO9781139649414.002

GRANGER, S. Learner Corpora in Foreign Language Education. *In:* Van DEUSEN-SCHOLL, N.; HORNBERGER, N. H. (ed.). *Encyclopedia of Language and Education.* 2. ed. New York: Springer, 2008. v. 4, p. 337-351.

HUNSTON, S. How Can a Corpus Be Used to Explore Patterns? *In:* O'KEEFFE, A.; McCARTHY, M. (ed.). *The Routledge Handbook of Corpus Linguistics*. London; New York: Routledge; Taylor & Francis Group, 2010. p. 152-166.

INTERNACIONAL CIVIL AVIATION ORGANIZATION. *Manual of Radiotelephony:* Doc 9432. Montreal: ICAO, 2007.

INTERNACIONAL CIVIL AVIATION ORGANIZATION. *Guidelines for Aviation English Training Programs:* Circular 323 NA/185. Montreal: ICAO, 2009. Available from: https://www.icao.int/safety/lpr/Documents/323_en.pdf. Access on: Mar. 23, 2020.

INTERNATIONAL CIVIL AVIATION ORGANIZATION. *Manual on the Implementation of ICAO Language Proficiency Requirements:* Doc. 9835 AN/453. 2. ed. Montreal: ICAO, 2010. Available from: https://skybrary.aero/bookshelf/books/2497.pdf. Access on: Mar. 23, 2020.

INTERNATIONAL CIVIL AVIATION ORGANIZATION. *Air Traffic Management. Doc 4444*. Montreal: ICAO, 2016. Available from: https://ops.group/blog/wp-content/uploads/2017/03/ICAO-Doc4444-Pans-Atm-16thEdition-2016-OPSGROUP.pdf. Access on: Mar. 23, 2020.

INTERNATIONAL CIVIL AVIATION ORGANIZATION. *Annex 3 to the Convention on International Civil Aviation.* Meteorological Service for International Air Navigation: parts I and II. 20. ed. Montreal: ICAO, 2018.

L'HOMME, M.-C. *Lexical Semantics for Terminology*: An Introduction. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2020.

MEUNIER, F. Developmental patterns in learner corpora. *In:* GRANGER, S.; GILQUIN, G.; MEUNIER, F. (ed.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 2015. p. 379-400.

NESSELHAUF, N. *Collocations in a Learner Corpus*. Amsterdam: Benjamins, 2005.PEIXOTO, R. A. J. R. P. Nas asas da tradução: elaboração de glossário de Meteorologia Aeronáutica. *Revista CBTecLE*, Santa Ifigênia, v. 2, n. 1, [s.p.], 2020.

PEIXOTO, R. A. J. R. P. Aeronautical Meteorology Glossary: A Discussion on Term Definition in the ANACpedia Termbase. *The ESPecialist*, São Paulo, v. 41, n. 3, p. 1-26, 2020.

PEIXOTO, R. A. J. R. P. Terminology of Aeronautical Meteorology Codes: A Systematization by Using Corpus. *TradTerm*, São Paulo, v. 37, n. 1, in press.

PEIXOTO, R. A. J. R. P.; PIMENTEL, J. M. M. Aeronautical Meteorology in Aeronautical Language and in Aviation Language: a hybrid field? *The ESPecialist*, São Paulo, v. 41, n. 4, p. 1-24, 2020. DOI: 10.23925/2318-7115.2020v41i4a2

RAYSON, P. From Key Words to Key Semantic Domains. *International Journal of Corpus Linguistics,* Birmingham, v. 13, n. 4, p. 519-549, 2008.

RENOUF, A.; SINCLAIR, J. Collocational Frameworks in English. *In:* AIJMER, K; ALTENBERG, B. (ed.). *English Corpus Linguistics*: Studies in Honour of Jan Svartvik. London: Longman, 1991. p. 128-143.

RÖMER, U. Comparing Real and Ideal Language Learner Input. *In:* ASTON, G.; BERNARDINI, S.; STEWART, D. (ed.). *Corpora and Language Learners*. Amsterdam: John Benjamins Publishing Company, 2004. p. 151-168.

RÖMER, U. Language Assessment and the Inseparability of Lexis and Grammar: Focus on the Construct of Speaking. *Language Testing*, [*S.l.*], v. 34, n. 4, p. 477-492, 2017.

SCARAMUCCI, M. V. R. O Exame de proficiência em língua inglesa para controladores de voo do SISICEAB: uma entrevista com Matilde Scaramucci. *Aviation in Focus,* Porto Alegre, v. 2, n. 1, p. 3-12, 2011. Available from: https://geia.icea.gov.br/geia/publicacoes.php. Access on: Jan. 23, 2020.

SCARAMUCCI, M. V. R.; TOSQUI-LUCKS, P.; DAMIÃO, S. M. (ed.). *Pesquisas sobre inglês aeronáutico no Brasil.* Campinas: Pontes Editores, 2018.

SCHMITT, N. *Vocabulary in Language Learning*. Cambridge: Cambridge Language Education, 2000.

SINCLAIR, J. Envoi. The Phrase, the Whole Phrase, and Nothing but the Phrase. *In:* GRANGER, S.; MEUNIER, F. (ed.). *Phraseology:* An Interdisciplinary Perspective. Amsterdam; Philadelphia: John Benjamins Publishing Company, 2008. p. 407-410.

STEFANOWITSCH, A. *CL:* A Guide to the Methodology. Berlin: Language Science Press, 2020. (Textbooks in Language Sciences 7). Available from: http://langsci-press.org/catalog/book/148. Access on: Sep. 7, 2020.

TAGNIN, S. E. O. A Multilingual Learner Corpus in Brazil. *In:* WILSON, A.; ARCHER, D.; RAYSON, P. (ed.). *Corpus Linguistics around the World*. Amsterdam; New York: Rodopi, 2006. p. 195-202.

THEWISSEN, J. The Phraseological Errors of French-, German- and Spanish-Speaking EFL Learners: Evidence from an Error-Tagged Learner Corpus. *In:* TEACHING AND LANGUAGE CORPORA CONFERENCE (TALC 8), 8th., 2008, Lisboa. *Proceedings* […]. Lisboa: Associação de Estudos e de Investigação Científica do ISLA-Lisboa, 2008. p. 300-306.

TOSQUI-LUCKS, P.; PRADO, M. C. de A. *Corpora* de inglês aeronáutico: desafios para o estudo da área e proposta de trabalho conjunto. *Tradterm,* São Paulo, v. 37, n. 1, in press.

TOSQUI-LUCKS, P.; SILVA, A. L. B. C. Aeronautical English: Investigating the Nature of this Specific Language in Search of New Heights. *The Especialist*, São Paulo, v. 41, n. 3, p. 1-27, 2020a. DOI: https://doi.org/10.23925/2318-7115.2020v41i3a2

TOSQUI-LUCKS, P.; SILVA, A. L. B. C. Da elaboração de um glossário colaborativo à discussão sobre os termos "inglês para aviação" e "inglês aeronáutico". *Estudos Linguísticos*, São Paulo, v. 49, n. 1, p. 97-116, 2020b. Available from: https://revistas.gel.org.br/estudos-linguisticos/article/view/2561. Access on: May 2, 2020.

UNITED STATES. Federal Aviation Administration. U.S. Department of Transportation. Air Traffic Organization Policy. *ORDER JO 7110.65W*: Air Traffic Control. Washington, D.C, 2015.

UNITED STATES. U.S. Department of Transportation. *Blue Lightning Initiative*. 2020. Available from: https://www.transportation.gov/administrations/office-policy/blue-lightning-initiative. Access on: Sep. 25, 2020.

WORLD METEOROLOGICAL ORGANIZATION. *Guide to Practices for Meteorological Offices serving Aviation*. (WMO, n.732). Geneva: WMO, 2003.

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1485-1527, 2021

Corpus linguistics and continuous professional development: participants' prior knowledge, motivations and appraisals

Linguística de corpus *e formação profissional contínua: conhecimento prévio, motivações e avaliações dos participantes*

Vander Viana
University of East Anglia, Norwich / United Kingdom
vander.viana@uea.ac.uk
http://orcid.org/0000-0003-3079-4393

Lu Lu
Hong Kong Polytechnic University, Hong Kong / Hong Kong
lu-cbs.lu@polyu.edu.hk
http://orcid.org/0000-0002-6049-6154

Abstract: Previous studies on the application of corpus linguistics (CL) to education have primarily examined language-related contexts where students are pursuing a formal degree (e.g. undergraduate and Master's programs). Little do we know about the informal learning of CL especially by (but not limited to) academics/professionals who are not educated and/or do not work in language-oriented fields. The present study addresses these research gaps by examining the perspective of participants in a non-credit-bearing continuous professional development (CPD) project aimed at academics/professionals in a range of disciplines, who did not need to have any prior knowledge of CL. More specifically, we administered a questionnaire to 28 participants of a UK-based CPD project on CL with a view to researching four main aspects: (i) these participants' CL background; (ii) their motivations to participate in this type of project; (iii) the advantages and barriers of employing CL in their teaching practice; and (iv) their appraisal of corpus analysis integration in their research practice. The results point out to the role of CPD projects in democratizing access to CL education both to language-oriented and non-language oriented academics/professionals and in potentially raising their interest in CL learning. Lack of knowledge is perceived to

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1485-1527

be the main barrier in embedding corpus approaches to teaching and research, thus reinforcing the relevance of developing formal and informal CL learning opportunities for academics/professionals in different fields.

Keywords: corpus linguistics; continuous professional development; educational corpus integration; evaluation of corpus use in professional practices; corpus application to teaching and research; language teacher education; translator education; interdisciplinarity.

Resumo: Estudos sobre a aplicação da linguística de *corpus* (LC) à educação examinaram uma série de contextos diferentes – principalmente aqueles em que os alunos recebem um diploma de colação de grau (por exemplo, cursos de graduação e mestrado). No entanto, pouco se sabe a respeito da aprendizagem informal da LC, especialmente por (mas não se limitando a) acadêmicos/profissionais que não tem uma formação educacional e/ou não trabalham em áreas relacionadas aos estudos da linguagem. A presente pesquisa preenche essas lacunas, examinando a perspectiva dos participantes de um projeto de formação profissional contínua destinado a acadêmicos/profissionais de várias disciplinas, que não precisavam ter conhecimento prévio de LC. Mais especificamente, administramos um questionário a 28 participantes de um projeto de formação profissional contínua na área de LC realizado no Reino Unido com o objetivo de pesquisar quatro aspectos principais: (i) a formação educacional em LC dos participantes; (ii) suas motivações para participar desse tipo de projeto; (iii) as vantagens e barreiras de empregar a LC em suas práticas pedagógicas; e (iv) suas avaliações sobre a integração da análise de *corpus* em suas práticas de pesquisa. Os resultados apontam para o papel dos projetos de formação profissional contínua na democratização do acesso à educação em LC para profissionais tanto da área de estudos da linguagem quanto de outras áreas e no potencial aumento do interesse desses profissionais na aprendizagem de LC. A falta de conhecimento é percebida como a principal barreira para a incorporação de abordagens de corpus para o ensino e a pesquisa, reforçando assim a relevância do desenvolvimento de oportunidades de aprendizagem formal e informal para acadêmicos/profissionais em diferentes áreas.

Palavras-chave: língüística de *corpus*; formação profissional contínua; integração educacional de *corpora*; avaliação do uso de *corpora* em práticas profissionais; aplicação de *corpora* no ensino e na pesquisa; formação de professores de línguas; formação de tradutores; interdisciplinaridade.

Submitted on September 10th, 2020

Accepted on October 21th, 2020

1 Introduction

This special issue of *Revista de Estudos da Linguagem* aims to take stock of the achievements and challenges of corpus linguistics (henceforth CL) over the years. While it would be challenging to precise exactly when CL started (see VIANA; ZYNGIER; BARNBROOK, 2011), Johansson (2008) clarifies that Jan Aarts first proposed the term *corpus linguistics* in the 1980s. In this decade, we also start to observe the academic uptake of corpus studies mainly due to the popularization of personal computers. In all these past years, CL has considerably evolved and has afforded new perspectives to our understandings of language use.

Corpus approaches have been used to examine different languages and their specific uses; however, the educational impact of CL has not been explored to the same extent. Naturally, it would be factually inaccurate to claim that there is little research on this topic: previous studies have investigated the integration of corpus analysis in numerous classroom settings. These settings include different languages being taught/learned (e.g. O'SULLIVAN; CHAMBERS, 2006 on French), educational levels (e.g. FRANKENBERG-GARCIA, 2015 on Master's students), countries (e.g. TODD, 2001 on Thailand), and disciplines (e.g. HAFNER; CANDLIN, 2007 on law students).

A review of the literature, however, reveals that much of the work conducted to date focuses on language-oriented educational contexts (e.g. FARR 2008; GAN; LOW; YAAKUB, 1996; HEATHER; HELT 2012; ZAREVA 2016) and degree-awarding settings where CL is taught in a compulsory or an optional module (e.g. BUENDÍA-CASTRO; LÓPEZ-RODRÍGUEZ, 2013; FRANKENBERG-GARCIA, 2015; GALLEGO-HERNÁNDEZ, 2015b). In other words, disciplines other than language-related ones and educational programs which are not credit-bearing remain underexplored in the research literature on educational applications of CL. To address these two research gaps, the present study innovates by investigating the perspective of participants from a range of disciplines in a non-credit-bearing continuous professional development (CPD) project. More specifically, it focuses on four main aspects here: (i) the CL background of participants who are drawn to CPD opportunities like this one; (ii) their motivations to participate in it; (iii) the advantages and challenges of employing CL in their teaching practice; and (iv) their evaluation of the integration of corpus analysis in their research practice.

To this end, a questionnaire was administered to the participants of a CL CPD project in the UK. The empirical data were analyzed in a bottom-up way which combined quantitative and qualitative analytical methods.

The present paper is divided into seven sections. Following this introduction, Section 2 reviews the literature on the integration of CL in two professional fields – language teaching and translation. In Section 3, we describe the CPD project that was offered to the research participants. Section 4 presents the methodological procedures adopted in this study. In Section 5, we describe our research participants, clearly indicating how they differ from the population samples in most of the studies conducted to date. The results of our analysis are presented and discussed in Section 6 before some final remarks are made in Section 7.

2 Literature review

CL has revealed its potential to contribute to several occupations – from language-oriented ones such as lexicographers and materials developers (FLOWERDEW, 2012; O'KEEFFE; MCCARTHY, 2010) to those which do not necessarily have a direct language component such as healthcare practitioners (CRAWFORD; BROWN, 2010) and lawyers (HAFNER; CANDLIN, 2007). In the present paper, we focus our attention on the embedding CL into teacher education (especially language teacher education) and translators, the two occupations that have received most attention in the research literature. The following subsections review the available research literature on corpus embedding in the education of these two professional groups.

2.1 CL in language teacher education

Many publications have highlighted the contribution that CL can bring to the field of language teaching (for useful summaries and overviews, see BIBER; REPPEN, 2015; O'KEEFFE; MCCARTHY, 2010). However, language teachers' use of corpus approaches in their language classrooms is far from being the mainstream practice (BOULTON 2010; RÖMER, 2010). The proponent of data-driven learning, Johns (1991) writes about two challenges in the integration of CL into language teaching. One challenge relates to teachers' roles, which need to change to 'a director and coordinator of student-initiated research' (p. 3). The other challenge relates to the use of traditional

teaching materials, which may have their accuracy questioned when concordancing tools take central stage in the classroom, and actual language use is analyzed. These two challenges (see also ASTON, 2011; CONRAD, 2011; VIANA, 2011) may help to explain why teacher education programs have not extensively incorporated modules on corpus analysis (CALLIES, 2019; FARR, 2010; GRANATH, 2009; MCCARTHY, 2008). The next two subsections will be dedicated to the integration of CL in the professional development of, respectively, pre-service and in-service teachers.

2.1.1 Pre-service language teacher education

Empirical studies on the effectiveness of using corpora in professional development have underlined the potential integration of CL in teacher education programs. As argued in Breyer (2009), the dual role – as a student and a future teacher – of pre-service teachers enables them to build a strong knowledge base in corpus queries and analyses as a learner before they expand what they have learnt to their workplace.

Several studies have focused on the use of CL in the teaching of vocabulary and grammar in pre-service English language teacher education (e.g. HEATHER; HELT, 2012; FARR, 2008; ZAREVA, 2016). Gan, Low, and Yaakub (1996) contrasted corpus approaches with traditional teacher-centered pedagogy at a Malaysian university regarding the teaching of vocabulary skills. Students in a pre-service teacher education program in Teaching English as a Second Language (TESL) were divided into an experimental group, which had five two-hour sessions on computer-based concordancing exercises, and the control group, which followed the teacher-centered approach with consultation from dictionaries. Pre- and post-tests revealed the experimental group excelled in the use of words in context. The benefits of CL in grammar teaching are well observed in Farr's (2008) sample of postgraduates in English Language Teaching (ELT), Zareva's (2016) survey on trainee teachers in the field of Teaching English to Speakers of Other Languages (TESOL), and Heather and Helt's (2012) grammar course for teachers of English as a Second Language (ESL). Difficulties and problems that may constrain the implementation of corpus approaches were also identified. One of the constraining factors is student teachers' language proficiency levels. In Heather and Helt's (2012) semester-long English grammar

course for pre-service teachers, the researchers collected students' questionnaire responses, their critique of prescriptive grammar rules and their design of supplementary teaching materials with corpus approaches. The findings revealed that pre-service teachers' grammatical knowledge affects the interpretation of corpus results to students. For instance, one student teacher with weak class performance wrongly categorized the auxiliary use of *be* as the main verb.

Apart from the above studies focusing on general English, the educational application of CL has been explored in ESP settings as well (e.g. VIANA; BOCORNY; SARMENTO, 2018). Hüttner, Smit and Mehlmauer-Larcher (2009) implemented corpus tools to ESP teaching: a small and specialized English corpus was built and contrasted with general reference corpora such as the BNC. This comparison offered a handy and reliable approach for students to identify obligatory and optional moves as well as formulaic expressions. As regards the tools for ESP teaching, two of the 32 pre-service teachers in Ebranhimi and Faghih's (2017) research believed that the free corpus software AntConc was useful in ESP education. This is because AntConc enables both learners and teachers to build a specialized corpus and generate a keyword list. However, student teachers also face challenges in CL-informed ESP classes. In Leńko-Szymańska's (2017) semester-long CL course, she collected students' end-of-semester assignments (e.g. self-compiled ESP corpora and corpus-based lesson plans) and argued that pre-service teachers only mastered basic CL technical skills at the lexical and phraseological levels, leaving other language features (e.g. register differences) barely untouched. Her study demonstrates that a semester-long course is not sufficient for pre-service teachers, who may lack the confidence and expertise to design CL-based activities for their own students in the future.

2.1.2 In-service language teacher education

Previous research on the interface between CL and in-service language teacher education has examined current professionals' use of corpora through questionnaires. Mukherjee (2004) investigated the actual use of CL in German secondary-school language teaching practice. His research results revealed that English language teaching had been hardly influenced by language features attested in corpora. Also drawing on the German context, Römer's (2009) survey of 78 secondary-school

English language teachers uncovered some of their desires and problems that could be well addressed by CL such as better teaching materials and reference resources other than non-corpus-based descriptions. More recently, Chen *et al*. (2019) analyzed the questionnaire responses provided by 54 in-service teacher participants of a data-driven learning workshop in Hong Kong. Among other findings, the results revealed correlations between teachers' prior knowledge of CL and their evaluation of the difficulty of corpus tools and between teachers' motivation for professional development and their adoption of data-driven learning. The types of investigation reviewed so far in this section are similar to the one that we have conducted in that we also adopted a questionnaire as our research instrument (cf. Section 4); however, our population sample is distinct as will be discussed in Section 5.

Another area that has been explored in the interface between CL and in-service teacher education is current professionals' use of corpora and corpus resources. One example of these resources is the Teachers of English Education Nexus (TeleNex), a website that is aimed at supporting the work of primary and secondary English language teachers in Hong Kong. Based on the analysis of 1,294 teacher-generated questions over eight years, Tsui (2005) advocates that schoolteachers' use of corpora is an effective way to help them understand the meaning and usage of a linguistic item. Corpus data were preferred over dictionary definitions, especially regarding queries on synonyms (such as *finally* vs. *lastly*) and stylistic patterns that seem to go against traditional prescriptive rules (e.g. whether sentences can start with conjunctions like *because*, *but* and *and*).

CL research has similarly investigated teachers' language use (e.g. CHAMBERS; O'RIORDAN, 2007; FARR, 2006). For instance, Vaughan (2007) examined how teachers use jargons and humor to maintain their membership identity in teacher-teacher talks at staff meetings. Farr (2006) researched trainer-trainee interactions in language teaching to uncover some of the linguistic and communicative features of this type of professional discourse such as self-disclosure strategies (e.g. 'all of us would…') and the high-frequency use of compliments (e.g. 'good') (see also FARR, 2005, 2010). Drawing on a US context, Reppen and Vásquez's (2007) and Vásquez and Reppen's (2007) action research over two semesters explored the spoken interaction between four pairs of teachers and their respective supervisors in post-observation meetings. Results from the first semester highlighted supervisors' higher talking time

in the collected data and prompted a change in supervisory practice. More specifically, supervisors created questions to give teachers the opportunities to ponder and generate more talk. The comparison of data in Semesters 1 and 2 revealed a change in the post-observation meetings: there was an increase in the total number of words produced by teachers whereas supervisors' amount of talk decreased or remained approximately the same. Reppen and Vásquez's (2007) and Vásquez and Reppen's (2007) studies illustrate the key role that corpus research can play in the reconsideration of professional practices. In their case, corpus findings triggered a change in supervisors' and teachers' roles in post-observation meetings, encouraging the teachers to take center stage in these meetings.

2.2 CL in translator education

Translation is another area that CL has influenced over the years potentially due to the facilitating role that corpora can have in translation tasks. For example, parallel corpora can facilitate the search for translation correspondences; and corpora of trainee translators and monolingual target-language corpora can be examined to identify and evaluate trainee professionals' translation solutions. The following sections will review empirical studies on the integration of CL in, respectively, pre- and in-service translator education.

2.2.1 Pre-service translator education

Previous studies on pre-service translator education have examined a few different contexts. From a geographical perspective, these studies have taken place, for example, in Denmark (e.g. LAURSEN; PELLÓN, 2012), Germany (e.g. KRÜGER, 2012), Spain (e.g. GALLEGO-HERNÁNDEZ, 2015b; MONZÓ-NEBOT, 2008; RODRÍGUEZ-INÉS, 2009), and the UK (e.g. FRANKENBERG-GARCIA, 2015). From an educational perspective, these studies have primarily examined degree-awarding courses – either at the undergraduate (e.g. BUENDÍA-CASTRO; LÓPEZ-RODRÍGUEZ, 2013; GALLEGO-HERNÁNDEZ, 2015b; ZANETTIN, 2001) or postgraduate level (e.g. FRANKENBERG-GARCIA, 2015).

The benefits of corpus introduction in pre-service translator education have been well argued and advocated. Working with a specialized monolingual first-language corpus, Bowker (1998), for

example, conducted a pilot study in a French-to-English translation classroom and showed that corpus analysis helped trainee translators understand the translation subject, terminology and idiomatic expressions. Zanettin (2001) drew on undergraduates' Italian-to-English translation of a newspaper text on the Olympic Games to illustrate how corpus exploitation can enable student translators to contrast source and target languages and to facilitate the selection of translation correspondences. Similar benefits have been reported in Rodríguez-Inés's (2009) Spanish-English translation class for 26 final-year Spanish students and Monzó-Nebot's (2008) legal translation courses for third- and fourth-year undergraduates in Spain.

The use of the web as a corpus (e.g. GATTO, 2014) in pre-service translator education has also been examined, and its advantages have been identified. The MA Specialized Translation students in Krüger's (2012) research believe that web concordances such as WebCorp can provide immediate solutions to language-related doubts that do-it-yourself (DIY) corpora (i.e. ad-hoc self-compiled corpora) may fail to solve. In another web-as-corpus study, Buendía-Castro and López-Rodríguez (2013) conducted an experiment with third-year undergraduate students in Translation and Interpreting at a university in Spain. These students were tasked with the translation to English of a research article excerpt on swine flu originally written in Spanish. It showed that the use of automatically built specialized corpora compensate for pre-service translators' lack of discipline-specific knowledge.

In addition to advantages, the literature on corpus integration in pre-service translator education has identified challenges. For example, Rodríguez-Inés (2010) highlighted the amount of time required in learning about corpora and their use. She pointed out that this may result in a time loss for pre-service translators to develop their translation skills and competence in a strict sense. In Gallego-Hernández's (2015b) study, the pre-service translator participants were split in their evaluation of the difficulty (N=14) or easiness (N=11) of corpus methods, and they indicated that time was a factor in their engagement with CL. Mixed feelings towards corpus work are also observed in Frankenberg-Garcia's (2015) study on 13 MA students in Translation at a UK university. While students appreciated exploiting corpora to assist them with the handling of unfamiliar terminologies and idiolects, they had some trouble in choosing appropriate corpora or corpus tools.

2.2.2 In-service translator education

The central conundrum in the integration of CL in in-service translator education is akin to the one observed in language teacher education (cf. Section 2.1). Although translators have begun to become aware that corpus approaches may support their day-to-day professional practice (e.g. VARELA-VILA, 2009), the uptake of these approaches is still relatively reduced (see, for instance, BOWKER, 2004; FRÉROT 2016; GALLEGO-HERNÁNDEZ, 2015a; JÄÄSKELÄINEN; MAURANEN, 2006; MELLANGE, 2006). Similar to pre-service translator education (cf. Section 2.2.1), time appears as one of the major barriers in translators' corpus uptake (ASTON, 2009; WILKINSON, 2006). This barrier is often noticed if translators have to compile their corpora (GALLEGO-HERNÁNDEZ, 2015a). It seems that translators would be more open to corpus approaches if the translation task involves a very large or interdisciplinary text, or if the translators themselves work full time.

A vicious circle can be identified in the integration of CL and professional translation. Perhaps because corpus uptake is not widespread, corpus skills are not mentioned in person specifications for translation posts as observed in Bowker's (2004) investigation of Canadian job adverts. At the same time, the lack of recognition of corpus skills as a sought-after ability in job ads does not encourage professional translators to acquire these skills (see also FRÉROT 2016). For example, Jääskeläinen and Mauranen's (2006) survey into Finnish timber industry found that corpora and concordance tools were not widely used, especially among freelance translators.

There are, however, some positive prospects in the integration of CL in in-service translator education. The survey of 1,015 translators and interpreters around the world in the Multilingual eLearning in LANGuage Engineering Project (MELLANGE, 2006) revealed some promising aspects: 20.2% of the respondents had used concordancers, and 82.0% would be interested to learn more about corpus-based translation skills. Gallego-Hernández's (2015a) survey equally indicates promising results: nearly half of the 526 participants indicated that they engaged in corpus exploration in their translation practice at a frequency that varied from 'sometimes' to 'very often'.

The present literature review indicates that the potential CL has to offer to professional education has not been fulfilled yet. In other words,

corpus approaches do not seem to have entered mainstream education or professional education courses. In the present study, we examine the perspectives of academics/professionals from different backgrounds on the integration of CL in their practices. The following section describes the CPD project that was offered to these participants.

3 CPD project on CL

This research investigated the perspectives of participants in a blended CPD project on CL funded by the British Academy. Merging research, teaching and learning perspectives, the project aimed at showing participants how to develop their CL skills and their students'/ supervisees'. The face-to-face element consisted of three day-long events spread over one year (i.e. June, September and December) with sessions delivered by experts in the field (e.g. Marina Bondi, Paul Thompson, Ute Römer). While Chen *et al*.'s (2019) research is also on a non-credit-bearing CL workshop, their target participants were limited in professional terms (i.e. it was aimed at English language teachers) and the length of their sessions was shorter (i.e. two three-hour long workshops, totaling six contact hours). The online space in our CPD project provided a further means for interaction among participants and for their learning to be consolidated over time with asynchronous input from the same team of experts.

This CPD project did not assume any prior knowledge of CL. The face-to-face and online activities were planned in such a way that participants would be introduced to the main concepts in CL before putting these concepts into practice in hands-on sessions and exploring the application of CL to their teaching and research.

The three face-to-face events had different but complementary foci. Participants were first introduced to the basics of language education and CL before they had two full days examining how this could be applied to language in general and to language for academic purposes. A decision was made to focus on English since this was the only shared language among all attendees, but the transferable nature of corpus skills was stressed.

4 Methods

We decided to use a questionnaire to collect data for the present study. Despite its inherent limitations (e.g. the potentially thin data to be collected), a questionnaire was the most appropriate option for our data collection plans. It required a reduced time commitment from participants, thus potentially increasing the final volunteer sample. This can be seen in our response rate, which will be discussed in Section 5.

In addition to the cover sheet, the questionnaire consisted of 25 questions divided into three parts. The first one contained questions about personal matters (e.g. sex, age, home country), participants' work experience, their educational background, and language knowledge and proficiency. Part 2 contained questions on participants' prior knowledge of CL as well as of related matters such as discourse analysis and statistics. Part 3 was dedicated to participants' reasons for registering for the CPD project, their expectations of it, and their appraisals of CL application to teaching and research.

At the beginning of the first event, participants were invited to complete the questionnaire anonymously. From an ethical perspective, our decision to ask participants to answer the questionnaire in the first face-to-face event could be challenged. While this is not unusual (e.g. FRANKENBERG-GARCIA, 2015; GAN; LOW; YAAKUB, 1996), we thoroughly considered whether the questionnaire should be answered online before the event or in person at the first event. We opted for the latter option because of two main reasons. Firstly, our target participants were primarily academics and/or professionals, who would probably struggle to find the time to answer the questionnaire before the event. Secondly, we felt it was essential for us to get to know the participants and to introduce the project to them in person before making any requests.

We were, however, aware that our request to answer the questionnaire in the first face-to-face session could be seen as a potential imposition by our participants, which would limit their perceived scope for declining to do so. This potential imposition is lower than in previous studies involving students where the researcher is also the teacher in charge of assessing the student participants (e.g. FRANKENBERG-GARCIA, 2015). Our relationship with the participants did not take place in any formal educational context where they would be evaluated for a credit-bearing module, for example. This was an optional CPD project

for which the participants had decided to register and to which they had already been accepted.

We followed four main steps in order to reassure participants of their freedom to decide whether or not to answer the questionnaire.

1. We explained to them the voluntary nature of their participation, and they had an opportunity to ask any questions before the questionnaire was distributed. They could naturally ask further questions at any point in time as well.
2. Participants completed the questionnaires anonymously. While we asked for some background information, it does not allow us to identify them – nor is it important to our research either.
3. Both researchers kept a physical distance from the participants during questionnaire completion so that they did not feel coerced to complete it. We would only approach specific participants if they called us to clarify any questions that they had.
4. The questionnaires were returned anonymously: we asked the participants to put their questionnaires in a manila envelope, which was placed at the back of the room. The envelope was only opened at the end of the first day after the participants had already left the venue. As no other writing sample was collected from participants throughout the CPD project, they were reassured that their identities were never disclosed to us. This procedure means that, once the participants returned their questionnaire, they could not withdraw from the research anymore. However, we felt that this was a fair compromise to ensure their anonymity, which we believed to be of higher importance in this research.

Following the completion of the paper questionnaires, participants' responses were digitized verbatim so that we could investigate the data electronically independently. The data collected through closed questions were analyzed quantitatively while the participants' answers to open-ended questions were analyzed qualitatively and quantitatively. Our approach to open-ended answers meant that they were initially studied in a bottom-up manner to identify themes, which were then quantified based on the number of occurrences.

Both of us were involved in the analysis. The first author analyzed all the data independently initially. He then shared the results with the

second author, who compared the results with her original analyses. She checked the quantitative results for accuracy and the qualitative results for thoroughness. There were only minor discrepancies in the qualitative analyses, which were resolved by discussing each of the relevant cases. Before the results are presented in Section 6, the following section will detail the participant sample in the present study.

5 Participants

A total of 36 registered participants were expected to attend the face-to-face events. Out of this total, three had expressed their impossibility in attending the first event, three were speakers who had to either arrive late or leave early, and two were the CPD project organizers, who are also the authors of this paper. This resulted in a pool of 28 potential participants, all of whom agreed to contribute to the study and answer the questionnaire. While the sample may be considered small, we worked within a non-interventionist research paradigm with the participants of a specific, real-life educational CPD project. As reviewed in Section 2, other pedagogical studies have researched a similar or even smaller number of participants. For example, Farr (2008) examined a sample of 25 MA student teachers in her questionnaire-based evaluation on participants' perception of corpus-assisted courses; Frankenberg-Garcia's (2015) study drew on the data provided by 13 Master's students in Translation at a UK university; Zareva (2016) analyzed 21 trainee teachers' responses to a questionnaire aimed at evaluating a corpus-based course design.

Our study had a 100% return rate, which is high for non-course/degree-based questionnaire studies. In Römer's (2009) research with in-service teachers, for instance, 78 out of 120 questionnaires were completed and returned. However, the difference in the overall population sample must be acknowledged. Our decision to request participants to complete the questionnaire in the first face-to-face event after we had initially established rapport with the participants (cf. Section 4) may have contributed to this high return rate.

5.1 Personal characteristics

Our participant sample is varied. We had a total of 16 female participants and 12 male ones, which is a somewhat even split. This differs

from many previous studies in which females considerably outnumber males (cf. 16 female vs. 5 male student teachers in ZAREVA, 2016) or the distribution of sex is not disclosed (cf. CHEN *et al.*, 2019; HEATHER; HELT, 2012; LEŃKO-SZYMAŃSKA, 2014).

Participants' ages vary from 23 to 55 with the mean being 41 years old. The age in our sample is older than in previous studies: participants' mean age in Zareva (2016) is 35.4 years old, and the participants' ages in Vásquez and Reppen (2007) range from the mid-20s to mid-30s. This difference is not surprising given the CPD project (cf. Section 3) and the primary occupation of the target participants (see Section 5.3).

5.2 Countries of residence/origin and language knowledge

Participants were asked about their countries of residence and origin. The answer to the former question indicates that most of them (N=26) lived in the UK (i.e. two participants decided not to answer this question) at the time of data collection. This is understandable and unsurprising given that the project entailed three face-to-face events in this country in one calendar year (i.e. June, September and December). This set-up would make it difficult for overseas participants to join the on-site events. In relation to participants' home countries, while we observe that most respondents are from the UK (N=20), there is more diversity with two participants from China and one participant from each of the following countries: Canada, France, Germany, India, Malaysia and Russia.

The range of nationalities described above helps to explain participants' language knowledge. English is the first language of most participants (N=15). Participants also reported having German (N=2), Latin (N=1), and Mandarin (N=2) as their first language. Altogether, nine participants decided not to answer this question.[1] English language command was not an issue: most participants (N=22) self-assessed themselves as proficient in the Common European Framework of Reference for Languages (CEFR), that is, either at C1 or C2 level. Only four participants declared to be at independent user levels (B1 and B2), and two decided not to answer this question. In terms of other languages, participants indicated that they knew 16 other languages to varying

[1] The total of 29 first languages is due to one participant's reporting to speak both German and Latin as first languages.

degrees of proficiency. The most recurrent additional languages were French (N=11), German (N=7), Japanese (N=3) and Russian (N=3). There were also mentions to Danish, Dutch, Hindi, Italian, Malay, Mandarin, Spanish, Telugu, Thai and Turkish to cite some examples.

5.3 Educational and professional background

Our project targeted a specific group of academics/professionals. All of the participants held at least a first degree and either had or were working towards higher degrees. One participant had a Diploma, 17 were educated to Master's level or were reading for one, and 10 were doctoral degree holders or were taking such a course. Most of these participants were affiliated with a higher education institution (N=27), encompassing both universities (in most cases) and colleges. Considerably smaller numbers worked at other educational institutions – e.g. schools and local authorities (N=4), and publishers (N=2).[2]

With regard to their occupation, most participants were based in an educational environment: language teachers (N=13), university lecturers (N=10), students (N=9). There were two other professions represented in the sample (i.e. a publisher and a corpus developer), and one participant decided not to answer this question.

The educational and professional profile outlined above coheres with our participants' ages. As their average age is 41 years old and as most participants are in their 40s and 50s (N=17), they have higher educational degrees (i.e. at postgraduate level) and have considerable work experience in their respective fields. The profile is also aligned with the target participant group for the CPD project, namely, academics and/or professionals.

5.4 Distinctive features of our population sample

Our participant sample stands out from the samples in previous studies due to our focus on a CPD project. Our participants held or were pursuing higher degrees – generally Master's or a doctorate. While there are studies with participants at postgraduate levels, they are generally with

[2] The total in this case is higher than the overall number of participants (N=28) because some of them declared more than one affiliation. The same is the case for the participants' reported occupations.

student cohorts (e.g. FARR, 2008; FRANKENBERG-GARCIA, 2015; KRÜGER, 2012), who learn about CL as part of their degrees – either on a compulsory or on a voluntary basis. Our CPD project differed from these courses in that it did not lead nor contribute to any educational degree. Participants' decision to register for this project was probably not because they may have felt it was compulsory to do so nor because they would be awarded a certificate at the end of it (see also CHEN *et al.,* 2019). Instead, as this was a voluntary CPD project, they were potentially intrinsically motivated to do so. After all, they had to make several commitments in relation to, for example, time (e.g. travelling to the venue and attending the three full-day events) and money (e.g. paying for their travel expenses).

In relation to their occupation, most participants worked in educational environments – be they language teachers or university lecturers. While nearly one-third of our participants were students (N=9), only five of them were exclusively students. The other four were either students who worked as teachers (N=3) or a student who was a lecturer (N=1). There have been studies conducted with language teachers, especially schoolteachers (cf. MUKHERJEE, 2004; RÖMER, 2009; TSUI, 2005). However, there seems to be a research gap concerning university lecturers – a notable exception is Chen *et al.*'s (2019) study.

Another stark difference between our study and the available literature has to do with our participants' most recent teaching experience. Understandably, most of them either teach English language (N=15) or work with English language teacher education (N=5). While participants similar to both groups have already been investigated in other educational and national contexts (e.g. CHEN *et al.*, 2019; FARR, 2008; LEŃKO-SZYMAŃSKA, 2014; RÖMER, 2009; TSUI, 2005), we have a more comprehensive range of disciplines being represented in our sample, which includes Applied Linguistics, Dementia, Education, History, Japanese, Russian and Sociology. One participant represented each of these disciplines except for Sociology, which was taught by two participants. The thinly spread disciplinary representation in this study does not allow us to make any specific points about them individually. However, the participant sample as a whole helps us to advance our current knowledge and understanding of the appeal of CL beyond the exclusively language-related disciplines, and it opens up an exciting new area of exploration in order to help us deepen the impact of CL across disciplines.

6 Results

The results of our study are presented and discussed in the following subsections. These subsections focus on four topics: (i) participants' background knowledge of CL before the start of the CPD project, (ii) their motivation to join this CPD project, (iii) their pre-project appraisal of the actual or potential application of CL to their teaching, and (iv) the same appraisal in relation to their research practice.

6.1 Previous education on CL

The publicity materials for the CPD project clearly stated that no prior knowledge of CL would be assumed or required from the participants. Instead, everyone with a keen interest in learning about corpus applications to education was welcomed and encouraged to apply. We were interested in finding out whether the call had a circular effect by appealing just to those who already had some knowledge of CL or whether it had been successful in drawing the attention of colleagues who had no or little knowledge of this field.

The findings reveal almost a split in participants' educational background on CL: 15 had studied it while 13 had never done so. Nearly half of the respondents who had previously studied CL (N=7) indicated that they had undertaken a CL module as part of their Master's in TESOL (N=5), Applied Linguistics (N=1) and Linguistics (N=1). Two participants developed their CL knowledge during their PhD since they employed corpus methods in their thesis research. Apart from one participant who learned about CL in his/her Diploma course, all the remaining eight indicated that they learned about CL through routes which did not lead to the award of a formal degree. These routes include a massive open online course, a workshop, and the informal and voluntary auditing of CL modules.

The 13 participants who indicated having had no prior study of CL before the CPD project were asked to explain why this was the case. Five participants referred to their lack of opportunities to do so.

1. "I have not had the opportunity up until now"[3] [F; 23; S; PhD (Social Policy)][4]
2. "I've got limited chance to learn." [F; 27; S; Master's (TESOL)]

As these two examples show, this lack of CL learning opportunities is not confined to non-language-related areas such as Social Policy (cf. Example 1) and History, but it is also observable in the previous experience of participants who specialize in areas like TESOL (cf. Example 2) and Translation (see ASTON, 2009 on the major barriers of CL among translators). Another explanation for the lack of formal CL education was participants' lack of interest in it. This explanation was given only by participants with a language-oriented educational background (i.e. Linguistics and TESOL): they acknowledged that they could have learned about CL during their formal studies, but they decided not to pursue this option. The other reasons mentioned by individual respondents referred to CL being claimed to be underdeveloped in a participant's field of research (i.e. English for Academic Purposes), a Social Work Lecturer's lack of CL awareness and long-standing focus on qualitative methods, and a participant's existing working knowledge on the use of corpora.

All of these responses provide useful pointers to help us pave the way for the future educational CL expansion. It seems vital for us to provide students with the opportunity to learn about CL as part of their degrees and/or in CPD projects like the one reported here. The introduction of CL in formal and informal educational programs is not a new recommendation: Renouf, back in 1997, argued in favor of introducing CL to postgraduate students of applied linguistics in the UK; Römer (2009) further recommended universities to reach out to teachers on problems and needs directly related to teaching through lectures and workshops. However, these recommendations do seem to have been

[3] Participants' responses have not been edited, and they are here reproduced verbatim.
[4] The code adopted in this study consists of four parts: the first letter indicates sex (F=female, M=male); the following numbers reveal participants' ages; the subsequent letter(s) stand for participants' occupations (D=Developer, L=Lecturer, P=Publisher, S=Student, T=Teacher, ?=no answer); and the final part indicates participants' highest degree (either completed or in progress) and its corresponding field. Therefore, "F; 23; S; PhD (Social Policy)" refers to a 23-year-old female student participant who either holds or is studying towards a PhD in Social Policy.

fully implemented. If, for the sake of illustration, we focus on TESOL Master's in the UK (cf. PAPAGEORGIOU *et al.*, 2017), only 35 out of 141 programs offer CL as a standalone module. These 35 programs are found in 17 universities (15 in England and 2 in Scotland), and the CL modules are nearly all optional with only two exceptions (VIANA, 2017). The provision of CL modules in the UK has indeed increased over the years, and this can be seen in the findings from the present study where most participants who already knew about CL had taken a specific module in their Master's. However, there is still scope for further improvement.

Not only should students be provided with opportunities to learn CL, their awareness and interest in it should also be raised. If we return our attention to TESOL Master's in the UK (cf. COPLAND *et al*., 2017), we will see that the finding reported here is not a one-off occurrence. TESOL Master's students were asked to appraise 15 modules on a 6-point Likert scale, ranging from 'not at all important' to 'extremely important'. The results show that CL had a mean of 4.58 with a standard deviation of 1.11 (VIANA, 2017). While this seems a somewhat encouraging result since it is above the 3.5 threshold, the mean for CL is the second last when all the 15 modules are considered (VIANA, 2017).[5]

The provision of more CL opportunities and the increase in their uptake are two related action points. The former is perhaps easier to achieve since it depends primarily on teaching staff to change the curriculum. The latter, however, will take more time because it will possibly require an attitudinal change among future generations.

6.2 Motivational drivers for CPD project participation

In order to help foster engagement with future CPD projects on CL, we must understand the participants' motivational drivers. Table 1 indicates participants' reasons for registering for the focal CPD project.

[5] The module perceived as least important by the participants in Copland *et al*.'s (2017) study is Translation with a mean of 3.83 and a standard deviation of 1.57.

TABLE 1 – Participants' reasons for registering for the CPD project on CL

Category	Frequency	Example
Application of CL	16	3. "With a background in corpus use for a very special purpose (machine translation & lexicography), I wanted to learn how corpora can be used for language teaching." [F; 26; D; Master's (Artificial intelligence)]
Research development	11	4. "Learn new research methods in stylometrics using burrows delta and nearest neighbour drivers" [M; 54; L; PhD (History)]
Knowledge of CL	9	5. "To get a better understanding of the software use in analysing corpora | To get a handle on statistics" [F; 49; S&T; EdD (Education)]
Personal interest	6	6. "I already had an interest but hadn't studied/read much so this project seemed to be something that could develop my interest and help me use corpora practically." [F; 55; T&L; Master's (Chinese)]
Financial reasons	1	7. "To Save money!" [F; 54; L; PhD (Educational Linguistics)]

Participants were primarily drawn to this CPD project as a way of developing their knowledge and understanding of CL applications. Most of the answers refer to educational applications of CL, which is understandable given that this was the main project focus. Although this motivation was observed in previous research (e.g. HEATHER; HELT, 2012; ZAREVA, 2016), participants' educational application needs were varied in this study. In Example 3, the participant's focus lies on the application of her previous knowledge as a corpus developer to the creation of outputs of relevance to language teachers. The applications were not exclusively related to education: there was also a reference to the application of the knowledge acquired in the UK to one participant's home country, for example.

Participants' wish to develop as researchers was the second most frequent reason. This finding should be interpreted alongside the unique population sample for this study (cf. Section 5): several of the participants were experienced professionals who had been working for a considerable number of years and/or who held academic positions. While research-related motivations can be found in the literature (e.g. ZAREVA, 2016), the studies are generally confined to language-related disciplines. Here, we notice that, in addition to the language-related

connections, our interdisciplinary participant sample establishes links to other fields. These areas include the investigation of authorship of historical materials (cf. Example 4) and reports in Accounting. There is an ample area for future exploration of CL: not only do we need to sediment the relationship between CL and language-related areas such as language teaching and translation, but we also need to capitalize on the impact that corpus studies may have outside our field of research. While we cannot claim how widespread the interest in CL across disciplines is or predict if this interest will eventually translate into real applications, our findings suggest that some colleagues from other areas are already (at least initially) willing to learn about CL. This means that they would not need to be convinced to do so and that they are open to this learning, which is an inspiring starting point.

Learning about CL was identified as a motivational driver by nearly one-third of the participants. In Example 5, the participant specifies CL-related matters such as statistical knowledge. This is not always the case, however. Sometimes participants provide a rather general indication of what they would like to learn by just referring to CL as a whole.

As expected, participants' intrinsic motivation played an essential role in their decision to register for the event (see also HEATHER; HELT, 2012; LEŃKO-SZYMAŃSKA, 2014). Example 6 indicates the mutual, two-way relationship between the participant's interest and her project participation. At the same time that the project was a way for this participant to undertake an activity that she was already willing to engage in, her participation also helped to foster her interest in the field.

A final reason mentioned by a single participant had to do with finances. There is not much contextual information to help the interpretation of this reason, but it could be assumed that this was possibly an allusion to the free-of-charge nature of the CPD project to the participants since the British Academy had fully sponsored it.

To further our understanding of the motivational drivers, participants were asked to indicate their three main expectations for the CPD project. This allowed us to check the extent to which their expectations matched their reasons for joining the project and whether any other relevant aspect had not been captured in the previously reported open-ended question. Their expectations were thematically analyzed, and the final categories are presented in Table 2.

TABLE 2 – Participants' expectations for the CPD project on CL

Category	Frequency	Example
Knowledge enhancement	31	8. "understanding technical aspects of building corpora" [F; 55; L; PhD (Linguistics)]
Application of CL	21	9. "to develop a better understanding of how learners can use corpora to improve their own linguistic competencies" [M; 35; T; Master's (International Relations)]
Research development	16	10. "develop ideas in how can incorporate corpus linguistics into an application for a research project" [F; 43; L; PhD (Social Work)]
Networking	6	11. "opportunity to meet new contacts" [F; 46; P; Master's (Languages)]
Practical skills improvement	6	12. "Learn hands-on skills" [F; 54; L; PhD (Educational Linguistics)]
Motivation increment	2	13. "enthuse me into world of corpora which I find quite dry just now" [F; 42; T; Master's (Applied Linguistics)]
CV	1	14. "CV building" [M; 25; ?; Master's (History)]

Participants' top three expectations were coherent with their reasons for having registered for the project – the only difference is the order in which they appear. 'Knowledge enhancement' features as the most frequent expectation while 'knowledge of CL' appeared as the third most frequent reason in Table 1. The former is wider encompassing than the latter; however, most of the expectations included in this category (N=24) referred to CL in general or specific CL matters like corpus compilation (cf. Example 8). The few remaining expectations included in this category (N=7) dealt with learning about quantitative methods and/or language use.

'Application of CL' and 'research development' were both main reasons to participate in the project and top expectations for it (cf. TABLES 1 and 2). Most participants' answers grouped in 'application of CL' refer to Education (cf. Example 9), which was the focus of the CPD project. Although research development has not been identified as a motivational driver in previous studies (see Section 2), it is one of the major driving forces to attend this project. Another interesting point is that, in our study, research development is not restricted to the usual language-oriented knowledge areas: it also concerns the other knowledge areas represented in the study as indicated in Example 10 from a Lecturer in Social Work.

The category of 'motivation increment' in Table 2 could be linked to 'personal interest' in Table 1. Despite the difference in their foci, 'motivation increment' encompassed examples where the participants indicated that the project could help to increase their interest in CL. Participants' reduced motivation level is more explicitly conveyed in the 'motivation increment' category as is evident in Example 13: the teacher participant expresses her lack of excitement with corpus work.

The list in Table 2 contains three new expectations that had not been mentioned in the previous analysis. The need to be in contact with like-minded CL researchers and practitioners was a top expectation to six participants. It is important to note that this expectation came from participants primarily in language-related fields (i.e. Educational Linguistics, Languages, Linguistics, TESOL) where there are notably more corpus experts and where it is easier to establish such networks. A few initiatives on this front can be seen in different parts in the UK such as 'Corpus Linguistics in the South' and 'Corpus Linguistics in Scotland'.

The category of 'practical skills improvement' is linked to the hands-on nature of CL. This can be interpreted in relation to Fligelstone's (1993, p. 98) well-known taxonomy of corpus-related activities: "teaching about", "teaching to exploit" and "exploiting to teach". Participants have shown their willingness to develop their knowledge and understanding of these three categories. They want primarily to be taught about CL (cf. 'knowledge enhancement') and to acquire or sharpen their skills in relation to exploiting to teach (cf. 'application of CL'). Some of them expect the workshop to teach them to exploit corpora (cf. 'practical skills improvement').

Finding the right balance among these three categories must be considered in the planning of pedagogical projects on CL. The data do not show any difference between those who had previously studied CL and those who had not had that experience. This means that the two groups want to learn about CL in the first place and they are also similarly interested in its applications. Learning how to do corpus analysis seems less of a priority for both groups. This may be because they need to understand whether acquiring these skills is a worthwhile investment of their time in the first place. An alternative explanation might be the participants' belief that they can develop their practical skills at a later stage perhaps in a more independent way. The results indicate that, for the type of target participants that we had envisaged (see Section 5), general aspects of CL should be prioritized over corpus practicalities.

The last new category observed in the project expectations relates to a concern with CV building, which was mentioned by a single participant. While pragmatic reasons such as this one are found in research projects examining student perspectives on their education (e.g. COPLAND *et al*., 2017), this was not relevant in this study. We believe that this is probably because of the population sample consisting primarily of academics/professionals working in the UK and of the non-credit-bearing aspect of the CPD project.

The findings indicate a roadmap for future pedagogical CPD projects to upskill academics'/professionals' knowledge and understanding of CL. A focus on both the core theoretical and research content as well as on CL applications seems to be necessary to meet participants' interest and to fulfill their expectations. This way, the CPD projects will act in a two-way relationship: appealing to participants' interest and raising their motivation. Although they do not seem to be essential, providing ways to develop participants' practical skills and networks should be given some consideration as well.

6.3 Appraisal of corpus applications to teaching

Participants were asked to identify the advantages and barriers of applying CL to their teaching practice. In both cases, approximately one-third of the participants (N=8 for the question on advantages and N=9 for the one on barriers) decided not to answer these questions, and they have not been included in the results reported in this section. The blank responses could be interpreted in several ways: the questions were open-ended for which response rates are usually lower than closed questions; participants were provided with the most generous space for their answers in this part of the questionnaire, thus suggesting that these answers would potentially be the longest ones; and/or the participants might be less motivated to complete this question since it was the penultimate one in the research instrument. In addition to these blank responses, there were a couple where the participants declared not to be applicable to them. For instance, one of the participants was a corpus developer and did not have any teaching experience. These answers were not included in the final analysis either.

Table 3 summarizes the categories created after a bottom-up analysis of the empirical data.

TABLE 3 – Advantages of applying CL to teaching practice

Category	Frequency	Example
Language use	7	15. "get students to be exposed to the real language" [M; 29; S; Master's (TESOL)]
Pedagogical improvement	7	16. "I have done some level of autonomy over 'how' I structure my teaching, and CL is going to make me a better practitioner" [M; 45; T; Master's (TESOL)]
Student autonomy	4	17. "students can use their own language as the basis for corpus searches" [M; 35; T; Master's (International Relations)]
Big data	1	18. "A way of analysing large amount of data" [F; 54; L; PhD (Educational Linguistics)]

The advantages can be related to either corpus or educational matters. Answers included in 'language use' or 'big data' reiterate points that are usually associated with corpus work. This is especially the case in relation to the first category, which is one of the main advantages of corpus investigations (e.g. SINCLAIR, 1991; TOGNINI-BONELLI, 2001). The association of CL with the investigation of large textual datasets is not uncommon (e.g. BOWKER; PEARSON, 2002; CONRAD, 2002), but corpus work is not restricted to them. The exploration of small, specialized corpora is also relevant in CL (e.g. FLOWERDEW, 2004; GHADESSY; HENRY; ROSEBERRY, 2001).

The two remaining categories establish a link between CL and Education. Participants believe that the exploration of corpora in their pedagogical practice will contribute to their professional development (cf. Example 16), their materials design skills, their skills in increasing student motivation, and their enhanced language explanations, to cite just some examples. Those advantages are observed in CL-informed classrooms such as the ones investigated by Heather and Helt (2012), Leńko-Szymańska (2014) and Zareva (2016). Participants also foresee a link between corpus work and student autonomy, a point that is recurrent in the literature (e.g. ASTON, 2011; CHARLES, 2014; GAVIOLI, 2009). As Example 17 indicates, the reference to student autonomy development is not restricted to those from a language-oriented educational background; it also made by participants with degrees in other fields.

Our analysis of the barriers of applying CL to teaching practice reveals that the most frequent issue faced by the participants is their lack

of relevant knowledge. This may relate to research (cf. Example 19), CL, IT skills and/or hands-on practice (see HEATHER; HELT, 2012; ZAREVA, 2016 for similar difficulties). This barrier is coherent with participants' reasons for enrolling in the project and their expectations of it: knowledge, research and practical skills development featured as important factors (cf. Section 6.2).

TABLE 4 – Barriers of applying CL to teaching practice

Category	Frequency	Example
Lack of knowledge	6	19. "Knowledge of the research process. I'm also a luddite so not very [unclear word] with computers." [M; 39; S&T; Master's (TESOL)]
Resources	4	20. "too much resources to select and summarise" [F; 27; S; Master's (TESOL)]
Time	4	21. "needing to spend a lot of time explaining what corpus linguistics is to my students before being able to use it in my teaching" [F; 55; L; PhD (Linguistics)]
Student-related issues	3	22. "Can students understand this or will they be interested?" [F; 42; T; Master's (Applied Linguistics)]
Lack of support from colleagues	2	23. "I am the only person in my workplace who engages with CL, and the general attitude is one of skepticism toward it" [M; 45; T; Master's (TESOL)]
Teaching-related challenges	2	24. "the staging of the lesson plan should be carefully prepared" [M; 29; S; Master's (TESOL)]

Two other barriers – 'student-related issues' and 'teaching-related challenges' – could be linked to participants' motivational drivers for participating in the project. As discussed in Section 6.2, learning about the corpus application (especially to education) was among the top factors. Interestingly, one participant questions students' ability to understand CL or to be interested in it (cf. Example 22). While it might be more challenging to address the question about students' interest, there is plenty of evidence in the literature that students are able to profit from data-driven learning (e.g. BOULTON, 2012; CHARLES, 2014; TODD, 2001).

Participants' reported lack of support from colleagues could perhaps explain their expectation to capitalize on their project participation

for networking purposes (see Section 6.2). In this sense, professionals differ from students, who reportedly receive sufficient support from their instructors (see ZAREVA, 2016). Professionals need to make the best use of the opportunity presented at such CL CPD projects to reach out to speakers and other participants for mutual support and development if they cannot find the support they need at their respective workplaces.

Two of the barriers are of a more practical nature: 'resources' and 'time'. The former category includes comments about computer access and burdensome programs. It also encompasses a comment on the large availability of resources (cf. Example 20), which is seen negatively because it requires one to select the most appropriate resource. A similar problem is evident in Frankenberg-Garcia (2015): student translators express difficulties in choosing the right corpora or the most effective query tools. Our research participants also mention the lack of time as a deterrent of corpus use in their teaching practice. This issue is approached from a range of angles: remarks are made in relation to the time required to prepare a corpus-based lesson and to explain some of the CL basics to students. In Example 21, the participant refers to the learning time required before reaping any positive outcomes (see also ASTON, 2009; WILKINSON, 2006).

When Tables 3 and 4 are compared, we notice that there is a similar overall number of advantages and barriers: 19 vs. 21, respectively. These come from the same number of participants. All the participants who identified at least one advantage also listed one barrier. The only exceptions lie with Participant F; 55; T&L; Master's (Chinese), who only focused on the positive side; and with Participant M; 39; S&T; Master's (TESOL), who did the opposite and commented on the challenges of integrating CL to his teaching practice.

6.4 Appraisal of corpus applications to research

Given the project's two foci on corpus applications to teaching and research, participants were additionally asked to identify the advantages and barriers of applying CL to their research practice. Zareva's (2016) study reveals an optimistic level of research enthusiasm (M=3.7 on a five-point Likert scale) among graduate students of TESOL in learning how to do corpus research. However, there seems to be a dearth of studies investigating how academics and/or professionals evaluate the use of CL in their research. The results reported in this section address this gap.

Similar to the procedure described in Section 6.3, the instances where the participants decided not to answer these questions and/or they felt they were unable to answer them were discarded. There were slightly fewer instances of non-completion in this case (N=7 for advantages and N=7 for barriers) when compared to the question on participants' teaching practice (N=8 for advantages and N=9 for barriers). Some participants indicated that this would not apply to their circumstances, and these answers were not included in the results either.

Table 5 summarizes the advantages that participants see in their adoption of a corpus approach to their research practice. Two of the categories are the same from Table 3: 'language use' and 'big data'. These reasons have been mentioned by different participants with a single exception. Participant M; 29; S; Master's (TESOL) referred to 'language use' twice in his answers. However, his answers are clearly distinct. Pedagogically, he commented on the introduction to students to real-life language use; research-wise, he singled out the role of corpora in discourse-related research and investigations of language use in different contexts.

TABLE 5 – Advantages of applying CL to research practice

Category	Frequency	Example
Methodological approach	12	25. "Learning to apply new research methods" [M; 54; L; PhD (History)]
Language use	3	26. "Corpus analysis can help with research of terminologies, collocations." [F; 45; S; Master's (Translation)]
Interdisciplinarity	2	27. "I think it could be a useful method to offer new/ future insights into accounting research" [M; 25; S; Master's (Research)]
Big data	1	28. "Analysis of a lot of data with relative ease" [M; 27; S&T; PhD (Linguistics)]
Collaborative work	1	29. "I am working with a corpus expert on a research." [F; 42; L; PhD (Linguistics)]
None	1	30. "None as yet" [M; 35; T; Master's (International Relations)]

The most frequently mentioned advantage relates to the methodological affordances provided by corpus work. There is a long-standing debate in the field whether CL is a science or a method

(see VIANA; ZYNGIER; BARNBROOK, 2011). It seems that the methodological advantages of CL correspond to its biggest advantage for the participants. For example, they commented on corpus techniques, the complementary/supplementary role that CL may play to qualitative research, and the new perspectives that corpus analysis may open up (see Example 25). We could potentially interpret this result in light of our participants' educational background. Because only five of them have their highest degree in Linguistics (here conceived in a strict way to include 'Linguistics', 'Languages' and 'Chinese'), most of the participants may be primarily more interested in the practical application of CL, thus seeing it as a way of assisting them in their research practice.

Other advantages included 'interdisciplinarity' and 'collaborative work'. The former highlights once more the nature of our participant sample: the project was successful in gathering colleagues from other non-language-related fields, who were interested in learning and applying CL to their research practice (see, for instance, Example 27 about Accounting). The category of 'collaborative work' reinforces points that had been made earlier in this paper: participants value opportunities to network (cf. Section 6.2) and they resent lack of support from teaching colleagues (cf. Section 6.3).

Finally, one participant claimed that CL had no advantages to his research practice. His concise reply does not allow for further discussion of his answer. However, because he gave the same answer as to the barriers (see TABLE 6) and because he provided full answers to the section on teaching practice, he may have meant that he was not able to answer this question due to, for instance, lack of CL research knowledge and/or lack of research experience. Alternatively, he may have meant that this question did not apply to his circumstances since he was working as a teacher at the time.

In relation to the participants' identified barriers as to the use of CL in their research, Table 6 reveals that there is some similarity to the pedagogical barriers (cf. TABLE 4). Three of the categories are the same: 'lack of knowledge', 'time' and 'resources' are equally impeditive to their undertaking of corpus research. Because of the physical proximity of these questions in the questionnaire, we checked whether participants had potentially given the same answers. This occurred in a limited number of cases: only five answers from four participants were equal, suggesting that they did not see any difference in the barriers they faced with the integration of CL in teaching or research.

The category entitled 'none' had also appeared in relation to the advantages (see TABLE 5). One case relates to Participant M; 35; T; Master's (International Relations) and has been discussed earlier in this section. The other instance is from Participant M; 48; T; Master's (TESOL), who has identified advantages and barriers in all the other three cases. This seems to suggest that he holds a positive perspective as to the use of corpus in his research practice.

TABLE 6 – Barriers of applying CL to research practice

Category	Frequency	Example
Lack of knowledge	10	31. "Unfamiliarity with verification procedures of research results" [M; 54; L; PhD (History)]
Time	4	32. "time consuming" [F; 42; T; Master's (Applied Linguistics)]
Copyright	2	33. "I think copyright can be an issue, with some document that I can use for corpus analysis" [F; 45; S; Master's (Translation)]
Lack of patience	2	34. "I'm also quite impatient- a bit problem perhaps for CL" [M; 39; S&T; Master's (TESOL)]
None	2	35. "None as yet" [M; 35; T; Master's (International Relations)]
Resources	2	36. "Computers!" [F; 45; S&L; PhD (Education)]
Avoidance of bias	1	37. "avoid researcher bias" [M; 29; S; Master's (TESOL)]

Table 6 contains three new categories. Two participants mention 'Copyright' as a hurdle that they may have to overcome in order to gain access to the texts for corpus analysis. This issue has been discussed in the literature, especially in relation to corpus compilation (e.g. BOWKER; PEARSON, 2002; MCENERY; XIAO; TONO, 2006; WYNNE, 2004). Example 33 comes from a student participant in Translation where copyright makes it extremely difficult or virtually impossible to investigate different translated versions of the same recent literary work of art, for example.

Interestingly enough, two participants identified their lack of patience to conduct corpus research as a barrier. Their answers to this specific question (see Example 34) do not provide much information for us to understand why their impatience would be an issue. A close investigation of their answers to the questionnaire indicate that they have a similar profile: both are from the UK, hold a Master's in TESOL and worked as language teachers at the time. They were not CL beginners: these two participants had studied CL in their Master's before their CPD project participation. In their responses to the barriers about teaching, both alluded to computer skill issues (see Example 19), a result that is supported by the self-evaluation of their familiarity with computer use (i.e. 'slightly familiar' for one participant and 'moderately familiar' for the other). It could be hypothesized that their unfamiliarity with technology might be a key factor leading to their impatience when using computer tools to undertake corpus research.

The other new category in Table 6 is 'avoidance of bias', which was mentioned by a single participant (cf. Example 37). The lack of any additional information in the questionnaire makes it difficult to make sense out of the participant's answer. It might have been the case that he wanted to include it in the box on the advantages and misplaced it in the barriers box. This misunderstanding would be coherent with the literature in that a corpus approach may reduce researcher bias (e.g. BAKER, 2006).

A comparative analysis of Tables 5 and 6 reveals that, while slightly more barriers than advantages have been identified, the difference is small to support any conclusions. Similar to what was observed in relation to teaching, the same number of participants identified advantages and barriers, meaning that the positive and negative aspects were equally dispersed.

7 Conclusion

This original study examined the perspectives of participants from different disciplinary backgrounds on a CL CPD project. It therefore addressed two research gaps in the educational application of CL: it researched the experience of participants from several disciplines (rather than only those from language studies) and investigated an underexplored educational context – a non-degree-awarding CPD one.

Methodologically, the study drew on a project funded by the British Academy that aimed to introduce academics and/or professionals to corpus research and applications. We administered a questionnaire to the 28 CPD project participants at the first face-to-face event, all of whom voluntarily decided to participate in this study. The rigorous analysis was conducted by integrating qualitative and quantitative approaches with both researchers carefully and thoroughly checking the analysis that had been undertaken by each other.

The findings revealed that the successful nature of the focal CPD project in drawing the attention of two types of participants: those who had never studied CL before, the main target participant group, and those who had already studied it previously. In this sense, the CPD project was successful in that it did not only appeal to those somehow versed in CL. Instead, complete novices in CL were reached, including those from non-language-related educational backgrounds.

Among the participants who had previously studied CL, half of them learned about it formally (primarily through Master's modules), and the other half developed their learning through routes that did not lead to the award of a degree. The former reinforces the relevance of including CL in the curriculum while the latter highlights the importance of informal learning initiatives such as CPD projects in making CL knowledge accessible to a larger number of (current/future) academics/professionals. As the participants indicated that lack of opportunities was one of the top barriers for their prior study of CL, more formal and informal CL learning opportunities should be provided in the years to come.

Increasing CL learning provision does not seem to suffice, though. As some participants pointed out, they had had the chance to learn about CL before this CPD project, but they lacked the interest to engage in it. We should therefore foster academics'/professionals' interest in these learning opportunities. The need to work on this personal aspect is reinforced by participants' indication that one of the drivers for their registration in this CPD project was to increase their motivations to undertake corpus analysis.

Participants' expectations of this CL CPD project foregrounded three main aspects. These academics and/or professionals wanted to (i) enhance their knowledge, primarily in relation to CL matters; (ii) learn about corpus applications, especially to the field of Education, which was the focus of the project; and (iii) further their research development

– mainly with regard to corpus analysis. To a certain extent, these expectations match their perceived barriers on the use of CL in their teaching and research practices: lack of knowledge was the main factor, thus potentially explaining why they had decided to register for this CPD project.

Our findings reveal participants' appraisals of the embedding of CL in their professional practices. When it comes to teaching, they indicate that corpus approaches allow them and their students to analyze language in use and that these approaches improve their pedagogical practice. Research-wise, the participants believe that the main advantage is the methodological approach afforded by CL such as in the case of interdisciplinary research. Naturally, the application of CL to teaching and research is not only seen from a positive angle. In both cases, the participants identify their lack of knowledge as the main barrier to the embedding of corpus approaches in their professional practices.

While the present study is based on a small participant sample, it is similar in size or even more extensive than some of the previous studies on the application of CL in educational contexts (e.g. FARR, 2008; FRANKENBERG-GARCIA, 2015; ZAREVA, 2016). Most importantly, we worked with all the participants who joined a specific CPD project on CL. It would not be sensible to increase the participant sample for research purposes since it would not reflect the real-life educational initiative being investigated here.

The significance of this study is two-fold. Research-wise, it advances our knowledge of academics'/professionals' perceived advantages and barriers of embedding CL in their respective workplaces, a context that is underexplored and differs from the one reported in previous studies (e.g. ZAREVA, 2016; RODRÍGUEZ-INÉS, 2013). The research findings lead to the study's practice-related significance: the need to increase the number of CL modules on offer and to develop more CPD projects like the one funded by the British Academy.

As educational initiatives would generally aim to have a long-lasting impact, a follow-up longitudinal study should be conducted in order to capture any potential changes in participants' perspectives on the use of CL in their teaching and/or research activities. It would also be worthwhile going beyond an investigation of their perspectives and examining their actual practices. However, we acknowledge that, although not impossible, this would be more challenging to accomplish

on many fronts such as having access to the participants' workplaces and ensuring the comparability of observed practices given participants' diverse professional backgrounds and activities.

The present study on the educational pedagogical application of CL has explored the significantly under-researched topic of CPD projects. The investigation of CPD participants' perspectives is much needed for furthering the impact of CL: it helps us understand what can be done to engage participants in CL work. Researching CPD participants' prior knowledge, motivations and appraisals are vital in tailoring future CPD projects accordingly, thus fulfilling their aim of democratizing access to CL education in informal settings.

Acknowledgements

The authors are grateful to the British Academy, which supported this project through a Skills Innovator Award (grant number SK150041), and to the participants, who kindly agreed to contribute to this research.

Authors' contributions

Viana had the idea for the study and was the lead researcher for most tasks. Both authors worked collaboratively in the design of the research instrument, which was administered by Lu. The data were digitized by Lu and analyzed by Viana and Lu. Viana was in charge of the overall structure of the paper, wrote most of the sections and thoroughly revised the entire paper. Lu was responsible for the literature review, drafted an initial version of the section on methods and the conclusion, contributed to the discussion of the findings, and read the entire paper critically.

References

ASTON, G. Foreword. *In*: BEEBY, A.; RODRÍGUEZ-INÉS, P.; SÁNCHEZ-GIJÓN, P. (ed.). *Corpus use and Translating*. Amsterdam/ Philadelphia: John Benjamins, 2009. p. ix-x.

ASTON, G. Applied Corpus Linguistics and the Learning Experience. *In*: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (ed.). *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins, 2011. p. 1-16. DOI: https://doi.org/10.1075/scl.48.01ast

BAKER, P. *Using Corpora in Discourse Analysis*. London: Continuum, 2006.

BIBER, D.; REPPEN, R. *The Cambridge Handbook of English Corpus Linguistics*. Cambridge: Cambridge University Press, 2015. DOI: https://doi.org/10.1017/CBO9781139764377

BOULTON, A. Data-Driven Learning: On Paper, in Practice. *In*: HARRIS, T.; MORENO JAÉN, M. (ed.). *Corpus Linguistics in Language Teaching*. Bern: Peter Lang, 2010. p. 17-52.

BOULTON, A. Corpus Consultation for ESP: A Review of Empirical Research. *In*: BOULTON, A; CARTER-THOMAS, S.; ROWLEY-JOLIVET, E. (ed.). *Corpus-Informed Research and Learning in ESP:* Issues and Applications. Amsterdam: John Benjamins, 2012. p. 261-292. DOI: https://doi.org/10.1075/scl.52.11bou

BOWKER, L. Using Specialized Monolingual Native-Language Corpora as a Translation Resource: A Pilot Study. *Meta,* [*S.l.*], v. 43, n. 4, p. 631-651, 1998. DOI: https://doi.org/10.7202/002134ar

BOWKER, L. Corpus Resources for Translators: Academic Luxury or Professional Necessity? *TradTerm,* São Paulo, n. 10, p. 213-247, 2004. DOI: https://doi.org/10.11606/issn.2317-9511.tradterm.2004.47178

BOWKER, L.; PEARSON, J. *Working with Specialized Language:* A Practical Guide to Using Corpora. London: Routledge, 2002. DOI: https://doi.org/10.4324/9780203469255

BREYER, Y. Learning and Teaching with Corpora: Reflections by Student Teachers. *Computer Assisted Language Learning*, [*S.l.*], v. 22, n. 2, p. 153-172, 2009. DOI: https://doi.org/10.1080/09588220902778328

BUENDÍA-CASTRO, M.; LÓPEZ-RODRÍGUEZ, C. I. The Web for Corpus and the Web as Corpus in Translator Training. *New Voices in Translation Studies*, [*S.l.*], n.10, p. 54–71, 2013.

CALLIES, M. Integrating Corpus Literacy into Language Teacher Education. *In*: GÖTZ, S.; MUKHERJEE, J. (ed.). *Learner Corpora and Language Teaching*. Amsterdam: John Benjamins, 2019. p. 245-263. DOI: https://doi.org/10.1075/slcs.201.12cal

CHAMBERS, A.; O'RIORDAN, S. Learning to Teach through French from a Corpus of Classroom Discourse: Giving Corrective Feedback. *In*: CONACHER, J. E.; KELLY-HOLMES, H. (ed.). *New Learning Environments for Language Learning*. Frankfurt: Peter Lang, 2007. p. 87-101.

CHARLES, M. Getting the Corpus Habit: EAP Students' Long-Term Use of Personal Corpora. *English for Specific Purposes*, [*S.l.*], n. 35, p. 30-40, 2014. DOI: https://doi.org/10.1016/j.esp.2013.11.004

CHEN, M.; FLOWERDEW, J.; ANTHONY, L. Introducing In-Service English Language Teachers to Data-Driven Learning for Academic Writing. *System*, [*S.l.*], n. 87, p. 102-148, 2019. DOI: https://doi.org/10.1016/j.system.2019.102148

CONRAD, S. Corpus Linguistic Approaches for Discourse Analysis. *Annual Review of Applied Linguistics*, Cambridge, n. 22, p. 75-95, 2002. DOI: https://doi.org/10.1017/S0267190502000041

CONRAD, S. Variation in Corpora and Its Pedagogical Implications. *In*: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (ed.). *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins, 2011. p. 47-62. DOI: https://doi.org/10.1075/scl.48.04con

COPLAND, F.; VIANA, V.; BOWKER, D.; MORAN, E.; PAPAGEORGIOU, I.; SHAPIRA, M. *ELT Master's Courses in the UK:* Students' Expectations and Experiences. London: British Council, 2017.

CRAWFORD, P.; BROWN, B. Health Communication: Corpus Linguistics, Data Driven Learning and Education for Health Professionals. *Taiwan International ESP Journal*, Taiwan, v. 2, n. 1, p. 3-28, 2010.

EBRAHIMI, A.; FAGHIH, E. Integrating Corpus Linguistics into Online Language Teacher Education Programs. *ReCALL*, Cambridge, v. 29, n.1, p. 120-135, 2017. DOI: https://doi.org/10.1017/S0958344016000070

FARR, F. Relational Strategies in the Discourse of Professional Performance Review in an Irish Academic Environment: The Case of Language Teacher Education. *In*: SCHNEIDER, K. P.; BARRON, A. (ed.). *Variational Pragmatics:* The Case of English in Ireland. Berlin: Mouton de Gruyter, 2005. p. 203-234.

FARR, F. Reflecting on Reflections: The Spoken Word as a Professional Development Tool in Language Teacher Education. *In*: HUGHES, R. (ed.). *Spoken English, Applied Linguistics and TESOL:* Challenges for Theory and Practice. Hampshire: Palgrave Macmillan, 2006. p. 182-215. DOI: https://doi.org/10.1057/9780230584587_9

FARR, F. Evaluating the Use of Corpus-Based Instruction in a Language Teacher Education Context: Perspectives from the Users. *Language Awareness,* [*S.l.*], v. 17, n. 1, p. 25-43, 2008. DOI: https://doi.org/10.2167/la414.0

FARR, F. How Can Corpora Be Used in Teacher Education? *In*: O'KEEFEE, A.; MCCARTHY, M. (ed.). *The Routledge Handbook of Corpus Linguistics*. London/New York: Routledge, 2010. p. 620-632.

FLIGELSTONE, S. Some Reflections on the Question of Teaching, from a Corpus Linguistics Perspective. *ICAME Journal,* [*S.l.*], n. 17, p. 97-109, 1993.

FLOWERDEW, L. The Argument for Using English Specialized Corpora to Understand Academic and Professional Settings. *In*: CONNOR, U.; UPTON, T. (ed.). *Discourse in the Professions:* Perspectives from Corpus Linguistics. Amsterdam: John Benjamins, 2004. p. 11-33. DOI: https://doi.org/10.1075/scl.16.02flo

FLOWERDEW, L. *Corpora and Language Education*. Basingstoke: Palgrave Macmillan, 2012. DOI: https://doi.org/10.1057/9780230355569

FRANKENBERG-GARCIA, A. Training Translators to Use Corpora Hands-On: Challenges and Reactions by a Group of Thirteen Students at a UK University. *Corpora*, Edinburg, v. 10, n. 3, p. 351-380, 2015. DOI: https://doi.org/10.3366/cor.2015.0081

FRÉROT, C. Corpora and Corpus Technology for Translation Purposes in Professional and Academic Environments: Major Achievements and New Perspectives. *Cadernos de Tradução*, Florianópolis, n. 36, p. 36-61, 2016. DOI: https://doi.org/10.5007/2175-7968.2016v36nesp1p36

GALLEGO-HERNÁNDEZ, D. The Use of Corpora as Translation Resources: A Study Based on a Survey of Spanish Professional Translators. *Perspectives,* [*S.l.*], v. 23, n. 3, p. 375-391, 2015a. DOI: https://doi.org/10.1080/0907676X.2014.964269

GALLEGO-HERNÁNDEZ, D. Business Translation Training and ad Hoc Corpora. *In*: SÁNCHEZ-GIJÓN, P.; TORRES-HOSTENCH, O; MESA-LAO, B. (ed.). *Conducting Research in Translation Technologies*. Oxford/New York: Peter Lang, 2015b. p. 119-140.

GAN, S. L.; LOW, F.; YAAKUB, N. F. Modeling Teaching with a Computer-Based Concordance in a TESL Preservice Teacher Education Program. *Journal of Computing in Teacher Education,* [*S.l.*], v. 12, n. 4, p. 27-32, 1996.

GATTO, M. *The Web as a Corpus:* Theory and Practice. London: Bloomsbury, 2014.

GAVIOLI, L. Corpus Analysis and the Achievement of Learner Autonomy in Interaction. *In*: LOMBARDO, L. (ed.). *Using Corpora to Learn about Language and Discourse*. Bern: Peter Lang, 2009. p. 39-72.

GHADESSY, M.; HENRY, A.; ROSEBERRY, R. L. (ed.). *Small Corpus Studies and ELT:* Theory and Practice. Amsterdam: John Benjamins, 2001. DOI: https://doi.org/10.1075/scl.5

GRANATH, S. Who Benefits from Learning How to Use Corpora? *In*: AIJMER, K. (ed.). *Corpora and Language Teaching*. Amsterdam/ Philadelphia: John Benjamins, 2009. p. 47-65. DOI: https://doi.org/10.1075/scl.33.07gra

HAFNER, C. A.; CANDLIN, C. N. Corpus Tools as an Affordance to Learning in Professional Legal Education. *Journal of English for Academic Purposes*, [*S.l.*], v. 6, n. 4, p. 303-318, 2007. DOI: https://doi.org/10.1016/j.jeap.2007.09.005

HEATHER, J.; HELT, M. Evaluating Corpus Literacy Training for Pre-Service Language Teachers: Six Case Studies. *Journal of Technology and Teacher Education,* [*S.l.*], v. 20, n. 4, p. 415-440, 2012.

HÜTTNER, J.; SMIT, U.; MEHLMAUER-LARCHER, B. ESP Teacher Education at the Interface of Theory and Practice: Introducing a Model of Mediated Corpus-Based Genre Analysis. *System,* [*S.l.*], n. 37, p. 99-109, 2009. DOI: https://doi.org/10.1016/j.system.2008.06.003

JÄÄSKELÄINEN, R.; MAURANEN, A. Translators at Work: A Case Study of Electronic Tools Used by Translators in Industry. *In*: BARNBROOK, G.; PERNILLA, D.; MAHLBERG, M. (ed.). *Meaningful*

Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora. London: Continuum, 2006. p. 48-53.

JOHANSSON, S. Some Aspects of the Development of Corpus Linguistics in the 1970s and 1980s. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics:* An International Handbook. Berlin: Mouton de Gruyter, 2008. p. 33-54.

JOHNS, T. Should You Be Persuaded: Two Examples of Data-Driven Learning. *In*: JOHNS, T.; KING, P. (ed.). *ELR Journal 4:* Classroom Concordancing. Birmingham: CELS, The University of Birmingham, 1991. p. 1-16.

KRÜGER, R. Working with Corpora in the Translation Classroom. *Studies in Second Language Learning and Teaching*, Kalisz, Poland, n. 4, p. 505-525, 2012. DOI: https://doi.org/10.14746/ssllt.2012.2.4.4

LAURSEN, L. A.; PELLÓN, A. I. Text Corpora in Translation Training: A Case Study of the Use of Comparable Corpora in Classroom Teaching. *The Interpreter and Translator Trainer,* [*S.l.*], n. 61, p. 45-70, 2012. DOI: https://doi.org/10.1080/13556509.2012.10798829

LEŃKO-SZYMAŃSKA, A. Is This Enough? A Qualitative Evaluation of the Effectiveness of a Teacher-Training Course on the Use of Corpora in Language Education. *ReCALL*, Cambridge, v. 26, n. 2, p. 260-278, 2014. DOI: https://doi.org/10.1017/S095834401400010X

LEŃKO-SZYMAŃSKA, A. Training Teachers in Data Driven Learning: Tackling the Challenge. *Language Learning & Technology*, Austin, TX, v. 21, n. 3, p. 217-241, 2017.

MCCARTHY, M. J. Accessing and Interpreting Corpus Information in the Teacher Education Context. *Language Teaching,* Cambridge, v. 41, n. 4, p. 563-574, 2008. DOI: https://doi.org/10.1017/S0261444808005247

MCENERY, T.; XIAO, R.; TONO, Y. *Corpus-Based Language Studies:* An Advanced Resource Book. London: Routledge, 2006.

MELLANGE. *Corpora and E-Learning Questionnaire*. Results Summary. 2006. Available on: http://mellange.eila.univ-paris-diderot.fr/Mellange-Results-1.pdf. Retrieved at: July 18, 2018.

MONZÓ-NEBOT, E. Corpus-Based Activities in Legal Translator Training. *The Interpreter and Translator Trainer*, [*S.l.*], n. 22, p. 221-252, 2008. DOI: https://doi.org/10.1080/1750399X.2008.10798775

MUKHERJEE, J. Bridging the Gap Between Applied Corpus Linguistics and the Reality of English Language Teaching in Germany. *In*: CONNOR, U.; UPTON, T. A. (ed.). *Applied Corpus Linguistics:* A Multi-Dimensional Perspective. Amsterdam: Rodopi, 2004. p. 239-250. DOI: https://doi.org/10.1163/9789004333772_014

O'KEEFFE, A.; MCCARTHY, M. (ed.). *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 2010. DOI: https://doi.org/10.4324/9780203856949

O'SULLIVAN, Í.; CHAMBERS, A. Learners' Writing Skills in French: Corpus Consultation and Learner Evaluation. *Journal of Second Language Writing,* [*S.l.*], n. 15, p. 49-68, 2006. DOI: https://doi.org/10.1016/j.jslw.2006.01.002

PAPAGEORGIOU, I.; VIANA, V.; COPLAND F.; BOWKER, D.; MORAN, E. *Master's ELT Audit Document*. 2017. Available on: https://www.teachingenglish.org.uk/sites/teacheng/files/Audit%20Final%2010.pdf. Retrieved on: July 1, 2020.

RENOUF, A. Teaching Corpus Linguistics to Teachers of English. *In*: WICHMANN, A.; FLIGELSTONE, A.; MCENERY, T.; KNOWLES, G. (ed.). *Teaching and Language Corpora.* London: Longman, 1997. p. 255-266. DOI: https://doi.org/10.4324/9781315842677-22

REPPEN, R.; VÁSQUEZ, C. Using Corpus Linguistics to Investigate the Language of Teacher Training. *In*: WALIŃKI, J.; KREDENS, K.; GOŹDŹ-ROSZKOWSKI, S. (ed.). *Corpora and ICT in Language Studies, PALC 2005*. Frankfurt am Main: Peter Lang, 2007. p. 13-29.

RODRÍGUEZ-INÉS, P. Evaluating the Process and Not Just the Product When Using Corpora in Translator Education. *In*: BEEBY, A.; RODRÍGUEZ-INÉS, P.; SÁNCHEZ-GIJÓN, P. (ed.). *Corpus Use and Translating*. Amsterdam/Philadelphia: John Benjamins, 2009. p. 129-150. DOI: https://doi.org/10.1075/btl.82.09rod

RODRÍGUEZ-INÉS, P. Electronic Corpora and Other Information and Communication Technologies Tools: An Integrated Approach to Translation Teaching. *The Interpreter and Translator Trainer*, [*S.l.*], v.

4, n. 2, p. 251-282, 2010. DOI: https://doi.org/10.1080/13556509.2010.10798806

RODRÍGUEZ-INÉS, P. Electronic Target-Language Specialized Corpora in Translator Education. *Babel*, [*S.l.*], v. 59, n. 1, p. 57-75, 2013. DOI: https://doi.org/10.1075/babel.59.1.04rod

RÖMER, U. Corpus Research and Practice: What Help Do Teachers Need and What Can We Offer? *In*: AIJMER, K. (ed.). *Corpora and Language Teaching*. Amsterdam/Philadelphia: John Benjamins, 2009. p. 83-98. DOI: https://doi.org/10.1075/scl.33.09rom

RÖMER, U. Using General and Specialized Corpora in English Language Teaching: Past, Present and Future. *In*: CAMPOY-CUBILLO, M. C.; BELLÉS-FORTUÑO, B.; GEA-VALOR, L. (ed.). *Corpus-based Approaches to English Language Teaching*. London: Continuum, 2010. p. 18-35.

SINCLAIR, J. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press, 1991.

TODD, W. R. Induction from Self-Selected Concordances and Self-Correction. *System,* [*S.l.*], n. 29, p. 91-102, 2001. DOI: https://doi.org/10.1016/S0346-251X(00)00047-6

TOGNINI-BONELLI, E. *Corpus Linguistics at Work*. Amsterdam: John Benjamins, 2001. DOI: https://doi.org/10.1075/scl.6

TSUI, A. B. M. ESL Teachers' Questions and Corpus Evidence. *International Journal of Corpus Linguistics*, [*S.l.*], v. 10, n. 3, p. 335-356, 2005. DOI: https://doi.org/10.1075/ijcl.10.3.03tsu

VARELA-VILA, T. Córpora ad hoc en la práctica traductora especializada: aplicación al ámbito de las enfermedades neuromusculares. *In*: CANTOS-GÓMEZ, P.; SÁNCHEZ-PÉREZ, A. (ed.). *A Survey on Corpus-Based Research*. Panorama de investigaciones basadas en corpus. Murcia: Asociación Española de Lingüística del Corpus, 2009. p. 814-831.

VÁSQUEZ, C.; REPPEN, R. Transforming Practice: Changing Patterns of Interaction in Post-Observation Meetings. *Language Awareness,* [*S.l.*], v. 16, n. 3, p. 153-172, 2007. DOI: https://doi.org/10.2167/la454.0

VAUGHAN, E. "I Think We Should Just Accept Our Horrible Lowly Status": Analysing Teacher-Teacher Talk in the Context of Community of

Practice. *Language Awareness,* [*S.l.*], v. 16, n. 3, p. 173-189, 2007. DOI: https://doi.org/10.2167/la456.0

VIANA, V. The Politics of Corpus Linguistics. *In*: VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (ed.). *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins, 2011. p. 229-245. DOI: https://doi.org/10.1075/scl.48

VIANA, V. Data-Driven Language Learning and Teaching: A Corpus Perspective. *In*: LINGUISTICS AND KNOWLEDGE ABOUT LANGUAGE IN EDUCATION (LKALE) BRITISH ASSOCIATION FOR APPLIED LINGUISTICS (BAAL) SPECIAL INTEREST GROUP (SIG) MEETING, 2017, Sheffield. Sheffield, 2017. [Oral presentation].

VIANA V.; BOCORNY, A.; SARMENTO, S. *Teaching English for Specific Purposes*. Alexandria: TESOL Press, 2018.

VIANA, V.; ZYNGIER, S.; BARNBROOK, G. (ed.). *Perspectives on Corpus Linguistics*. Amsterdam: John Benjamins, 2011. DOI: https://doi.org/10.1075/scl.48

WILKINSON, M. Compiling Corpora for Use as Translation Resources. *Translation Journal,* [*S.l.*], n. 101, [s.p.], 2006. Available on: http://www.translationdirectory.com/article910.htm. Retrieved at: Aug. 21, 2020.

WYNNE, M. Archiving, Distribution and Preservation. *In*: WYNNE, M. (ed.). *Developing Linguistic Corpora:* A Guide to Good Practice. Oxford: Oxbow Books, 2004. p. 71-78.

ZANETTIN, F. Swimming in Words: Corpora, Translation, and Language Learning. *In*: ASTON, G. (ed.). *Learning with Corpora*. Bologna: CLUEB, 2001. p. 177-197.

ZAREVA, A. Incorporating Corpus Literacy Skills into TESOL Teacher Training. *ELT Journal*, Oxford, v. 71, n. 1, p. 69-79, 2016. DOI: https://doi.org/10.1093/elt/ccw045

Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1529-1638, 2021

O desenho de tarefas pedagógicas para o ensino de Inglês para Fins Acadêmicos: conquistas e desafios da Linguística de Corpus

The design of pedagogical tasks for teaching English for Academic Purposes: achievements and challenges of Corpus Linguistics

Ana Eliza Pereira Bocorny
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul / Brasil
ana.bocorny@ufrgs.br
https://orcid.org/0000-0002-0515-9630

Anamaria Welp
Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre, Rio Grande do Sul / Brasil
anamaria.welp@ufrgs.br
https://orcid.org/0000-0002-9015-4761

Resumo: Nas últimas décadas, um grande número de instituições de ensino superior (IES) buscou a internacionalização de suas atividades. Sendo o inglês a língua franca da academia (AMMON, 2011; JENKINS, 2009; TARDY, 2004), publicar nesse idioma facilita a disseminação do conhecimento científico produzido no país e aumenta as chances de citação e de colaboração (BOCORNY *et al*., no prelo; MENEGHINI; PACKER, 2007 *apud* BAUMVOL, 2018*)*. Tendo em vista o contexto descrito, este estudo objetiva propor princípios para a elaboração de tarefas pedagógicas (TPs) com a utilização de dados linguísticos extraídos de um *corpus* especializado e relativos à linguagem convencionalmente usada em artigos de pesquisa. Desse objetivo geral, derivam dois objetivos específicos. O primeiro, de ordem analítica, busca extrair, categorizar e classificar expressões multipalavra a partir de um *corpus* especializado de textos da seção *introdução* de artigos de pesquisa recentes (2003-2019) publicados em inglês em periódicos internacionais da área da Física. O segundo, de ordem pedagógica,

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1529-1638

visa usar os dados linguísticos coletados para informar a construção de TPs para o ensino e a aprendizagem de Inglês para Fins Acadêmicos (IFA). As TPs resultantes deste estudo estão disponibilizadas *on-line* e de forma gratuita no Ambiente Virtual de Aprendizagem LÚMINA Idiomas (BOCORNY, 2017).

Palavras-chave: tarefas de acesso aberto; Linguística de *Corpus*; gêneros acadêmicos; expressões multipalavra; Inglês para Fins Acadêmicos.

Abstract: In the last decades, a large number of higher education institutions (HEIs) sought to internationalize their activities. Since English is the lingua franca of the academy (AMMON, 2011; JENKINS, 2009; TARDY, 2004), publishing in that language facilitates the dissemination of scientific knowledge produced in the country and increases the chances of citation and collaboration (BOCORNY *et al.*, in press; MENEGHINI; PACKER, 2007 apud BAUMVOL, 2018). In view of the described context, this study aims to propose principles for the elaboration of pedagogical tasks (PTs) with the use of linguistic data extracted from a specialized corpus related to the language conventionally used in research articles. From this general objective, two specific objectives are derived. The first, of an analytical nature, seeks to extract, categorize and classify multi-word expressions from a specialized corpus of texts in the introduction section of recent research articles (2003-2019) published in English in international physics journals. The second, of a pedagogical nature, aims to use the collected linguistic data to inform the construction of PTs for teaching and learning English for Academic Purposes (EAP). The PTs resulting from this study are available online and free of charge in the Virtual Learning Environment LÚMINA Idiomas (BOCORNY, 2017).

Keywords: open access tasks; Corpus Linguistics; academic genres; multi-word expressions; English for Academic Purposes.

Recebido em 10 de outubro de 2020

Aceito em 16 de dezembro de 2020

1 Introdução

O ensino superior (ES) mudou substancialmente nas últimas décadas. Uma dessas mudanças está relacionada ao fato de instituições de ensino superior (IES) em todo o mundo buscarem internacionalizar suas atividades (BOCORNY *et al*., no prelo). De acordo com o "Scimago Journal & Country Rank",[1] nos últimos 23 anos (1996-2019), a produção científica brasileira apresentou um crescimento significativo, saindo do

[1] Disponível em: https://www.scimagojr.com/countryrank.php. Acesso em: 21 set. 2020.

21º para o 15º lugar no *ranking* da produção científica internacional, com 973.456 artigos publicados. Apesar de estar à frente de países como Suécia e Finlândia, a nossa produção científica pode beneficiar-se de uma maior qualificação linguística de nossos pesquisadores para a produção e a publicação de artigos em periódicos internacionais de alto fator de impacto. O aumento do volume de publicações de nossa comunidade científica em periódicos internacionais tem o potencial de dar maior visibilidade ao conhecimento construído por nossos pesquisadores. Sendo o inglês a língua franca da academia (AMMON, 2011; JENKINS, 2009; TARDY, 2004), publicar nesse idioma facilita a disseminação do conhecimento científico produzido no país e aumenta as chances de citação e de colaboração (BOCORNY *et al.,* no prelo; MENEGHINI; PACKER, 2007 *apud* BAUMVOL, 2018*)*.

A identificação de padrões lexicais que apresentam certa estabilidade e frequência (HYLAND, 2012), típicos dos gêneros acadêmicos e das áreas de especialidade, tem grande importância pedagógica no âmbito do ensino das linguagens especializadas. Para ensinar escritores menos proficientes a produzir textos conforme os padrões considerados convencionais pelos especialistas das áreas nas quais se inserem, é preciso identificar os padrões linguísticos recorrentes. Os textos especializados comunicam por meio de elementos linguísticos (termos, unidades terminológicas, unidades fraseológicas, pacotes lexicais) e não linguísticos (gráficos, tabelas, imagens). A comunicação via meios linguísticos, segundo o princípio da idiomaticidade de Sinclair (1991), acontece através de porções de língua (*chunks of language*) que o usuário tem a seu dispor e não por meio de palavras individuais. Os padrões que essas unidades estabelecem entre si e com outros elementos do texto se constituem como fios que se entrelaçam em uma trama de sentido. Este trabalho trata da observação e da descrição de um dos elementos linguísticos fraseológicos recorrentes dessa trama e da sua utilização para informar a construção de tarefas pedagógicas (TPs) voltadas para o ensino de Inglês para Fins Acadêmicos (IFA).

Tendo em vista o contexto apresentado, a pesquisa desenvolvida tem como objetivo propor princípios para a elaboração de tarefas pedagógicas (TPs) com a utilização de dados linguísticos extraídos de um *corpus* especializado e relativos à linguagem convencionalmente usada em artigos de pesquisa da área da Física. Desse objetivo geral, derivam dois objetivos específicos. O primeiro, de ordem analítica, busca

extrair, categorizar e classificar expressões multipalavra a partir de um *corpus* especializado de textos da seção *introdução* de artigos de pesquisa recentes (2003-2019) publicados em inglês em periódicos internacionais da área da Física. O segundo, de ordem pedagógica, visa usar os dados linguísticos coletados com o propósito de informar a construção de tarefas pedagógicas no âmbito do ensino e da aprendizagem de Inglês para Fins Acadêmicos (IFA).

Espera-se que tais aplicações pedagógicas auxiliem pesquisadores brasileiros menos proficientes a produzirem artigos de pesquisa em inglês utilizando a linguagem convencionalmente encontrada nos periódicos internacionais de maior prestígio. Pesquisas futuras buscarão ampliar, para outras áreas do conhecimento e para outros gêneros acadêmicos, a descrição linguística e a metodologia de desenvolvimento de TPs voltadas ao ensino de língua para esse segmento.

Com vistas a cumprir os objetivos propostos, o presente artigo organiza-se em cinco seções: introdução, revisão da literatura, metodologia, resultados e, por fim, as considerações finais.

2 Revisão de literatura

O resultado prático que se pretende atingir com este estudo deriva do encontro e do entrelaçamento de pressupostos teóricos oriundos de três áreas do conhecimento: (i) os estudos sobre gêneros do discurso, (ii) os princípios da Linguística de *Corpus* e (iii) as teorias relativas ao ensino e à aprendizagem com base em tarefas.

2.1 Gêneros do discurso

Ao investigar-se a noção de "gêneros", encontra-se na literatura uma diversidade denominativa e conceitual. Alguns autores, como Bakhtin (2010), utilizam o termo "gêneros do discurso"; já outros, como Marcuschi (2002), o termo "gêneros textuais". Nesta pesquisa, não se fará distinção entre ambos. Quanto à heterogeneidade de definições de gênero, destacam-se brevemente apenas as concepções que são mais relevantes para este estudo, segundo Bakhtin (2010), Swales (1990) e Bhatia (2001).

Bakhtin (2010, p. 262) define os gêneros do discurso como "tipos relativamente estáveis de enunciados". Para ele, os gêneros textuais são relativos aos diversos textos que conseguimos identificar pelo fato de

eles se repetirem de um modo razoavelmente regular em determinados contextos (BAKHTIN, 2010). Dessa forma, os gêneros apresentam certas características estáveis relativas à sua macroestrutura, às escolhas linguísticas e ao público ao qual eles se destinam. Foi Bakhtin que passou a entender os gêneros como entidades sociodiscursivas que são criadas em distintas esferas da atividade humana e que refletem as condições específicas de cada uma.

Um dos traços principais da definição de gênero, de acordo com Swales (1990), é a sua caracterização por um propósito comunicativo compartilhado por membros de determinada comunidade discursiva, ou seja, a comunidade da área do conhecimento onde o texto-alvo é produzido. Swales (1990, 2004), ao apresentar seu modelo analítico para descrever a estrutura retórica da introdução de artigos de pesquisa através da análise dos padrões organizacionais e retóricos, chamados por ele de movimentos (*moves*) e passos (*steps*), mostra que tanto os movimentos quanto os passos são unidades retóricas que expressam funções comunicativas no discurso oral ou escrito (SWALES, 2004). O Quadro 1 mostra uma adaptação do modelo *Create a Research Space* (*CARS*) resultante da combinação dos modelos de Swales (1990, 2004).

QUADRO 1 – Adaptação do modelo *CARS* para introduções de artigos de pesquisa resultante da combinação das duas versões do modelo de Swales (1990, 2004)

MOVIMENTO 1: Estabelecendo um território
Passo 1: Defendendo a centralidade do tópico
Passo 2: Fazendo generalizações
Passo 3: Revisando pesquisas prévias
MOVIMENTO 2: Estabelecendo um nicho
Passo 1A: Indicando lacunas ou **Passo 1B:** Adicionando ao que já é sabido
Passo 2: Apresentando justificativas
MOVIMENTO 3: Introduzindo o presente estudo
Passo 1: Anunciando a presente pesquisa de forma descritiva e/ou seus propósitos
Passo 2: Apresentando problemas de pesquisa ou hipóteses
Passo 3: Esclarecendo a terminologia
Passo 4: Descrevendo procedimentos

Passo 5: Apresentando resultados
Passo 6: Estabelecendo o valor da presente pesquisa
Passo 7: Descrevendo a estrutura do trabalho

Fonte: Swales (1990, 2004).[2]

A partir do conceito de gêneros de Swales, Bhatia (2001) propõe a sua própria definição: para ele, um gênero é um conjunto de propósitos comunicativos compartilhados por uma determinada comunidade discursiva. Por consequência, se os propósitos comunicativos mudam, pode haver uma mudança no gênero. Uma das contribuições mais importantes de Bhatia para a proposta desta pesquisa é a sua definição de análise de gênero como sendo o estudo do comportamento linguístico em contextos acadêmicos ou profissionais (BHATIA, 2001).

A concepção de gênero usada nesta pesquisa está alinhada com as três perspectivas apresentadas. Ressalta-se a concepção de gênero do discurso como tipos relativamente estáveis de enunciados (BAKHTIN, 2010). Destaca-se, também, a ideia de que os gêneros são veículos de comunicação através dos quais se busca atingir um objetivo (SWALES, 1990). Por fim, sublinha-se o entendimento de que a análise de gênero é o estudo do comportamento linguístico em contextos acadêmicos ou profissionais (BHATIA, 2001).

2.2 Linguística de *Corpus*

A Linguística de *Corpus* parte de uma perspectiva de descrição da língua, seja tal língua geral ou especializada. Inicialmente, os *corpora* eram coletados e analisados manualmente, armazenados muitas vezes em fichas de papel. Na atualidade, a Linguística de *Corpus* está relacionada à tecnologia, que possibilita o armazenamento e o estudo de textos via ferramentas computacionais desenvolvidas para a análise de *corpora* textuais.

Os estudos a partir de grande volume de dados (*big data*) com a utilização de ferramentas potentes de análise de *corpora*, conduzidos pelos princípios e pelas metodologias propostos pela Linguística de *Corpus*, possibilitam a extração de dados linguísticos que podem

[2] Todas as traduções neste artigo são de nossa autoria.

contribuir com o desenvolvimento de aplicações pedagógicas usadas no ensino de diferentes gêneros textuais.

Elementos linguísticos fraseológicos recorrentes são bastante estudados no âmbito da Linguística de *Corpus* (cf. BIBER; CONRAD, 1999; BIBER *et al*., 2004; BIBER; BARBIERI, 2007; BIBER, 2009; GRAY; BIBER, 2013; STAPLES *et al*., 2013). Tais expressões recebem nomes diferentes conforme suas características. As expressões multipalavra recorrentes e contínuas de três ou mais palavras (por exemplo*, the aim of this paper is*) mais frequentes em determinado registro são denominadas *lexical bundles* (BIBER *et al.,* 1999; BIBER *et al.,* 2004). As expressões multipalavra recorrentes e descontínuas, isto é, estruturas que apresentam palavras gramaticais fixas e *slots* variáveis preenchidos por palavras de conteúdo (por exemplo*, don't * to, it is * to*), recebem o nome de *formulaic frames* (BIBER, 2009) ou *lexical frames* (GRAY; BIBER, 2013).

Cortes (2013) realiza um estudo no qual relaciona a linguagem formulaica, mais especificamente os *lexical bundles*, presente em artigos de pesquisa de diferentes disciplinas aos movimentos retóricos desse gênero. Como Cortes (2013), ao propormos este estudo, temos em vista fornecer os elementos fraseológicos típicos de uma disciplina, relacionando tais elementos linguísticos aos movimentos retóricos e aos passos do gênero estudado. Em especial, temos em vista a extração, a categorização e a classificação de *key lexical bundles* (*KLBs*) típicos de uma área de especialidade, bem como a posterior construção manual de *key lexical frames* (*KLF*), conforme proposto por Borcorny *et al.* (no prelo). Tais dados linguísticos, extraídos de um *corpus* da seção *introdução* de artigos da área da Física, serão utilizados para a elaboração de uma SD voltada ao ensino e à aprendizagem de IFA.

2.3 O trabalho com gêneros do discurso através de tarefas

Segundo Green (2020), por sua natureza plural, é mais adequado se considerar o letramento acadêmico como *letramentos*. Isso porque letramentos acadêmicos se configuram como um conjunto de práticas comunicativas características do discurso acadêmico, as quais são moldadas e desempenhadas pelos propósitos comunicativos particulares desse contexto. Para o autor, o domínio das práticas de letramento no contexto da academia está relacionado à construção de uma nova identidade, que implica se apropriar e se sentir confortável com uma

nova forma de pensar, agir e se comunicar, ou seja, sentir-se pertencendo a uma determinada comunidade.

Como mencionamos anteriormente, com a finalidade de investir na produção acadêmica brasileira, dando visibilidade ao conhecimento científico produzido no país, julga-se oportuno incentivar a publicação de artigos em língua inglesa, em virtude do seu *status* de língua franca da academia (AMMON, 2011; JENKINS, 2009; TARDY, 2004). Nesse sentido, o ensino e a aprendizagem de língua inglesa para fins acadêmicos devem se pautar na ampliação do repertório de práticas de letramento de pesquisadores através da familiarização com os gêneros que circulam nesse contexto. Em outras palavras, é preciso oferecer-lhes oportunidades para que participem de práticas de letramento realizadas em língua inglesa.

As práticas de letramento no ensino de línguas, sobretudo no ensino de línguas para fins acadêmicos, podem ser abordadas à luz da teoria sociocultural de Vygotsky (1998) e à noção de "andaimento" associada a ela. Gibbons (2013) explica que o termo "andaimento" é uma metáfora criada por Wood, Brunner e Ross (1976) para relacionar a estrutura utilizada por trabalhadores na construção civil com o suporte fornecido, no decorrer da interação, por um parceiro mais experiente a um aprendiz que desempenha uma tarefa ou resolve um problema que está acima de sua capacidade. O auxílio promovido durante essa interação, entretanto, é temporário, pois, no curso natural do aprendizado, o andaimento deve ser removido, permitindo que o aprendiz seja capaz de levar a cabo as ações aprendidas de forma independente.

Nesse contexto, com o propósito de possibilitar ao aprendiz a construção de conhecimento disciplinar de forma independente e a mobilização desse conhecimento de maneira adequada, visando à participação legítima em sua comunidade acadêmica, qualquer abordagem de ensino e aprendizagem de língua deve se preocupar em promover a construção de andaimes. Assim, os materiais didáticos para esse fim devem se guiar pela noção de "andaimento" e procurar oportunizar a familiarização do aprendiz com gêneros acadêmicos e suas ferramentas através do uso e da reflexão, permitindo a sua atuação nas práticas de letramento que orbitam tais gêneros.

Por sua natureza dialógica e por refletir a multiforme atividade humana, os gêneros discursivos constituem-se como elementos de socialização e de organização das práticas sociais. Nesse sentido, o

trabalho com gêneros no ensino e na aprendizagem de línguas permite que os indivíduos participem de diferentes práticas mediadas pela linguagem. Segundo Bakhtin (2010, p. 265), "[...] a língua passa a integrar a vida através de enunciados concretos (que a realizam); é igualmente através de enunciados concretos que a vida entra na língua".

Na perspectiva do filósofo russo, os gêneros são sócio-historicamente situados e refletem as condições específicas e as finalidades de cada campo de atividade humana, "não só por seu conteúdo (temático) e pelo estilo da linguagem, ou seja, pela seleção dos recursos lexicais, fraseológicos e gramaticais da língua mas, acima de tudo, por sua construção composicional" (BAKHTIN, 2010, p. 261). Dessa forma, podem ser considerados instrumentos de mediação para o ensino da textualidade (DENARDI, 2017).

Schneuwly e Dolz (2004) afirmam que o trabalho com gêneros é a base para a organização da atividade pedagógica. Assim, se o ensino e a aprendizagem de línguas devem se orientar a partir do trabalho com gêneros, o planejamento das atividades deve priorizar o uso da linguagem para ampliar as práticas de letramento de membros de uma dada comunidade. Os autores sugerem que o ensino de línguas com base em gêneros seja organizado a partir de sequências didáticas (SDs), as quais definem como:

> [...] uma seqüência de módulos de ensino, organizados conjuntamente para melhorar uma determinada prática de linguagem. As sequências didáticas instauram uma primeira relação entre um projeto de apropriação de uma prática de linguagem e os instrumentos que facilitam essa apropriação. Desse ponto de vista, elas buscam confrontar os alunos com práticas de linguagem historicamente construídas, os gêneros textuais, para lhes dar a possibilidade de reconstruí-las e delas se apropriarem. (SCHNEUWLY; DOLZ, 2004, p. 51).

Portanto, o ensino por tarefas, organizadas em SDs, parece ser a alternativa mais apropriada para abordar as práticas de letramento acadêmico, pois, conforme Welp, Didio e Finkler (2019, p. 25), além de oferecer oportunidades de experimentação do uso da língua, com toda a sua complexidade, em situações reais, o ensino e a aprendizagem através de tarefas pedagógicas (TPs) favorecem a interação através da

colaboração e promovem "o trabalho exploratório e a reflexão sobre as formas linguísticas a serem usadas para criar mensagens significativas".

Em consonância com Van Den Branden (2006), neste artigo, definimos tarefa como uma ação desempenhada por um indivíduo, por meio do uso da língua, para alcançar um objetivo. De acordo com o autor, ao trabalhar a partir de TPs, o aluno aprende a língua através do seu uso e, ao mesmo tempo, tenta encontrar soluções para problemas de comunicação reais, com o propósito de alcançar um objetivo não linguístico (VAN DEN BRANDEN, 2016).

Tendo em vista que o ensino e a aprendizagem de línguas para fins acadêmicos têm por finalidade a familiarização, o aprimoramento e a ampliação das práticas sociais realizadas na academia e que, para Vygotsky (1998), a linguagem é um artefato simbólico que medeia a construção do conhecimento, a elaboração de tarefas e a criação de SDs têm o potencial de oferecer andaimento ao aprendiz, guiando o processo de transformação de conhecimento espontâneo em conhecimento científico por meio da linguagem.

3 Metodologia

Os procedimentos metodológicos descritos a seguir foram adotados para atingirem-se os objetivos propostos neste estudo, ou seja, extrair, categorizar e classificar expressões multipalavra de um *subcorpus* da seção *introdução* de artigos de pesquisa da área da Física, bem como para usarem-se os dados linguísticos coletados a fim de informar-se a construção de uma SD.

3.1 Descrição do *subcorpus* especializado usado no estudo

A análise da linguagem da introdução de artigos de pesquisa a partir da Linguística de *Corpus* (LC) inicia com a compilação do *corpus*, realizada por meio da ferramenta *AntCorGen* (ANTHONY, 2019). O *corpus* especializado resultante desse processo de compilação foi chamado de *Corpus of Discipline and Section-Specific Academic English* (*CODISAE*). Conforme Bocorny *et al.* (no prelo), o *CODISAE* é um *corpus* composto de 12 milhões de palavras que reúne textos das quatro seções (*Introdução*, *Materiais e Métodos*, *Resultados* e *Conclusão*) de artigos de pesquisa recentes (2003-2019) das áreas das Ciências da Saúde, da Física e das Ciências da Computação, escritos em inglês e

publicados em periódicos internacionais, revisados por pares e de acesso aberto na plataforma *PLoS ONE*. O *CODISAE* apresenta 12 *subcorpora* de um milhão de palavras. Cada *subcorpus* é composto de textos de uma seção de uma das áreas do conhecimento estudadas. Neste estudo, será utilizado o *subcorpus* da seção *introdução* de artigos de pesquisa da área da Física do *CODISAE*. O Quadro 2 mostra a composição dos *subcorpora* da área da Física.

QUADRO 2 – *Subcorpora* do *CODISAE* com artigos da área da Física

Seção	Número de palavras
Introdução	1 milhão
Materiais e Métodos	1 milhão
Resultados	1 milhão
Conclusão	1 milhão

Fonte: Elaboração própria.

3.2 Procedimentos para a extração, a categorização e a classificação das expressões multipalavra no *subcorpus* especializado

A extração dos *key lexical bundles* (*KLBs*) do *subcorpus* especializado foi feita por meio da ferramenta *Sketch Engine* (KILGARRIFF *et al.*, 2004) conforme os seguintes critérios: deveriam ser compostos de seis unidades lexicais, ter uma frequência mínima de seis ocorrências por milhão de palavras, estar presentes em pelo menos cinco textos diferentes do *subcorpus* e ter um índice de chavicidade – ou seja, um índice que indica o quanto aquela unidade é mais frequente no *corpus* de estudo em relação ao *corpus* de referência – maior que um.

Extraídos os *KLBs*, inicia-se o processo de categorização, que acontece a partir da identificação de palavras-chave comuns a um grupo de *KLBs* ou da identificação de unidades com mesma função comunicativa. A partir da observação das categorias de *KLBs*, é possível identificar, nessas unidades, a existência de palavras gramaticais fixas e de palavras de conteúdo variáveis (*slots*), como pode ser visto no exemplo apresentado no Quadro 3.

QUADRO 3 – Palavras gramaticais fixas (negrito) e palavras de conteúdo variáveis (*slots*) em *KLBs* com a mesma função comunicativa da seção *introdução* de artigos da área da Física

	purpose	***of***	*this*	*paper*	***is***	***to***
	purpose	***of***	*this*	*study*	***is***	***to***
t	*purpose*	***of***	*this*	*study*	***is***	
	aim	***of***	*the present*	*study*		
	aim	***of***	*the present*	*study*	***was***	
	aim	***of***	*the present*	*study*	***was***	
	aim	***of***	*this*	*paper*	***is***	***to***
		of	*the present*	*study*	***is***	***to***
	aim	***of***	*this*	*paper*	***is***	
		of	*the present*	*study*	***was***	***to***

Fonte: Elaboração própria.

A observação dos *KLBs* categorizados e organizados em quadros facilita a identificação da sua função comunicativa. No caso dos *KLBs* dispostos no Quadro 3, é fácil dizer que sua função comunicativa é apresentar o(s) objetivo(s) do estudo. É importante lembrar que a Linguística de *Corpus*, ao assumir uma perspectiva descritiva da linguagem, não estabelece categorias *a priori*. As categorias emergem da observação dos dados. Entretanto, quando a função comunicativa dos *KLBs* não é facilmente identificada, pode-se usar um *framework* como o apresentado por Swales (1990, 2004) para auxiliar nessa tarefa. Um *framework* como o de Swales (1990, 2004), que lista os movimentos retóricos e os passos de um determinado gênero juntamente com suas funções comunicativas, pode servir como referência para a identificação da função comunicativa de alguns *KLBs* que não sejam tão transparentes. Por fim, parte-se para a construção manual dos *key lexical frames* (*KLFs*). Nessa etapa, as palavras gramaticais fixas dos *KLBs* de uma mesma categoria e com a mesma função comunicativa permanecem na primeira linha da estrutura, enquanto as palavras lexicais variáveis são marcadas por asteriscos. Na estrutura construída neste trabalho, as palavras lexicais variáveis, que podem preencher os *slots*, são listadas abaixo dos asteriscos. Tal estrutura é mostrada no Quadro 4.

QUADRO 4 – *KLF* construído manualmente a partir dos *KLBs* com a função comunicativa *apresentar o(s) objetivo(s) do estudo*, da seção *introdução* de artigos da área da Física

the	*	of	*	*	is/was	to
	purpose		this	paper		
	aim		the present	study		

Fonte: Elaboração própria.

A seção seguinte descreve a metodologia utilizada para o desenvolvimento de TPs a partir dos *KLFs* construídos manualmente (*cf.* Quadro 5) com os *KLBs* obtidos do *subcorpus* especializado usado neste estudo.

3.3 Princípios e procedimentos para a construção de TPs

Como já mencionado, do objetivo geral deste estudo, derivam dois objetivos específicos. O primeiro, de ordem analítica, orientou a seção 3.2. O segundo, de ordem pedagógica, consiste em usar os dados linguísticos coletados para informar a construção de TPs. Assim, esta seção apresenta os princípios e descreve os procedimentos envolvidos no processo de elaboração de uma SD para um curso de IFA, tarefa que tem como objetivo a produção da introdução de artigos acadêmicos da área da Física. Nesse sentido, a SD proposta abordará a linguagem utilizada em publicações dessa área do conhecimento.

A partir do arcabouço teórico que fundamentou o trabalho aqui descrito, com base em Welp, Didio e Finkler (2019), foram estabelecidos os seguintes princípios que devem guiar o desenho e a elaboração de TPs voltadas para o ensino e a aprendizagem de línguas para fins acadêmicos (LFA):

1. os objetivos de aprendizagem devem ser estabelecidos a partir da área do conhecimento e das necessidades acadêmicas do grupo de aprendizes para o qual as tarefas serão produzidas;
2. os gêneros discursivos estruturantes escolhidos devem ser academicamente relevantes e coerentes com os objetivos de aprendizagem estabelecidos;
3. os textos selecionados devem ser autênticos e representativos das práticas sociais e dos gêneros que circulam no contexto acadêmico;

4. as tarefas devem oferecer oportunidades de uso da língua próprias aos textos produzidos na área do conhecimento dos aprendizes e devem promover reflexões sobre tal uso de forma contextualizada;
5. as tarefas que tratam dos recursos linguísticos devem levar em consideração a frequência dos itens lexicais e discursivos presentes em textos acadêmicos da área de conhecimento do aprendiz;
6. a ordem e os enunciados das tarefas devem ser organizados de forma que ofereçam andaimento, oportunizando, assim, o aprendizado;
7. as tarefas devem provocar interações relevantes entre alunos e textos, alunos e alunos e alunos e professor;
8. a realização das tarefas deve oferecer oportunidades de aprendizado significativo e deve alcançar resultados para além da sala de aula.

Assim como em Welp, Didio e Finkler (2019), após estabelecidos os princípios, passou-se à estruturação da metodologia, descrita a seguir:

1. definição dos objetivos de aprendizagem, considerando-se a área do conhecimento do grupo de alunos para os quais as tarefas serão produzidas;
2. definição do gênero discursivo da produção que resultará da sequência de tarefas;
3. compilação de um *corpus* de textos do gênero estruturante;
4. extração dos dados linguísticos do *corpus* relevantes para o gênero;
5. elaboração das tarefas.

4 Resultados

Os resultados descritos a seguir dizem respeito (i) à extração, à categorização e à classificação das expressões multipalavra da seção *introdução* de artigos de pesquisa da área da Física; e (ii) ao uso, na construção de TPs, dos dados linguísticos coletados em (i).

4.1 Extração, categorização e classificação das expressões multipalavra

A extração dos *KLBs,* conforme os critérios apresentados em 3.2, resultou em um total de 17 unidades (cf. Apêndice 1). Como descrito em Bocorny *et al*. (no prelo), a classificação dos *KLBs* foi um processo

manual que iniciou com a organização de todas as unidades extraídas do *corpus* de estudo em uma tabela (cf. Apêndice 1). Com as unidades organizadas, foi iniciado o processo de categorização. Nessa etapa, as unidades com estrutura semelhante (*the * of this study was to **), alguma palavra lexical comum (*aim, study*) ou palavras lexicais que expressassem a mesma função comunicativa foram marcadas com a mesma cor (cf. Apêndice 2). O Quadro 5 mostra as cinco categorias identificadas a partir dos 17 *KLBs* extraídos.

QUADRO 5 – *KLFs* construídos a partir dos *KLBs* extraídos da seção *introdução* de artigos da área da Física que expressam as funções comunicativas descritas no modelo proposto por Swales (1990, 2004)

	KLBs	***KLFs***
MOVIMENTO 1: Estabelecendo um território		
Passo 1: Defendendo a centralidade do tópico	*plays an important role in the* *play an important role in the*	*(play/plays) an important role in the*
Passo 2: Fazendo generalizações	*it is well known that the* *it has been shown that the*	*it (is well known/has been shown) that the*
Passo 3: Revisando pesquisas prévias		
MOVIMENTO 2: Estabelecendo um nicho		
Passo 1A: Indicando lacunas ou **Passo 1B:** Adicionando ao que já é sabido	*to the best of our knowledge*	
Passo 2: Apresentando justificativas		
MOVIMENTO 3: Introduzindo o presente estudo		
Passo 1: Anunciando a presente pesquisa de forma descritiva e/ou seus propósitos	*of the present study is to* *purpose of this paper is to* *purpose of this study is to the* *purpose of this study is the aim of the present study aim of the present study was aim of this paper is to the aim of this paper is of the present study was to*	*the (purpose/aim) of (this /the present) (study/paper) (is/was) to*
Passo 2: Apresentando problemas de pesquisa ou hipóteses		

Passo 3: Esclarecendo a terminologia		
Passo 4: Descrevendo procedimentos		
Passo 5: Apresentando resultados		
Passo 6: Estabelecendo o valor da presente pesquisa		
Passo 7: Descrevendo a estrutura do trabalho	*this paper is organized as follows the rest of the paper is the paper is organized as follows*	*(the rest of the/this) paper is organized as follows*

Fonte: Elaboração própria.

A identificação das funções comunicativas exercidas por cada uma das cinco categorias de *KLBs* se deu pela observação das unidades, tendo-se como parâmetro o modelo apresentado por Swales (1990, 2004). Como pode ser observado no Quadro 5, a maior incidência de *KLBs* acontece no movimento retórico 3 (*introduzindo o presente estudo*), passo 1 (*anunciando a presente pesquisa de forma descritiva e/ou seus propósitos*), no qual se encontram 09 dos 17 *KLBs* extraídos (53%). Ainda no movimento 3, passo 7 (*descrevendo a estrutura do trabalho*), há 03 das 17 unidades (17%). Encontram-se no movimento 3, portanto, 70% dos *KLBs* extraídos do *corpus* de estudo. A incidência de *KLBs* no movimento 1 (*estabelecendo um território*) diminui bastante. No passo 1 (*defendendo a centralidade do tópico*), encontram-se 02 dos 17 *KLBs* (12%), enquanto no passo 2 (*fazendo generalizações*) há 02 unidades (12%). No passo 3 (*revisando pesquisas prévias*) do movimento 1, não há incidência de *KLBs*. Temos no movimento 1, desse modo, 30% dos *KLBs* extraídos do *corpus* de estudo. Por fim, no movimento 2 (*estabelecendo um nicho*), passo 1A (indicando lacunas) há a incidência de 01 dos 17 *KLBs* (6%) que trata da identificação de lacunas em estudos prévios. Finalmente, na coluna *KLFs* do Quadro 5, encontramos as estruturas lexicais construídas manualmente a partir do agrupamento dos *KLBs* de um mesmo passo com a mesma função comunicativa, conforme sugerido por Bocorny *et al.* (no prelo). Tal agrupamento resulta em quatro *KLFs* (cf. QUADRO 5) que serão utilizados como dados linguísticos para a construção da SD proposta.

4.2 Uso dos dados linguísticos coletados na construção de TPs

Nesta seção, apresentamos o contexto para o qual a SD aqui apresentada foi produzida e descrevemos as referidas TPs.

4.2.1 Contextualização

A SD apresentada a seguir foi desenhada para a disciplina de Inglês Instrumental I da Universidade Federal do Rio Grande do Sul, cujo programa é organizado a partir de gêneros acadêmicos estruturantes. A disciplina, que tem quatro créditos e é ministrada por uma das autoras, é presencial, tem como um de seus gêneros estruturantes o artigo de pesquisa, exige um conhecimento pré-intermediário de língua inglesa e é de caráter optativo para alunos de todos os cursos da universidade. Na próxima subseção, descrevemos o processo de elaboração da SD produzida para alunos da área da Física da disciplina de Inglês Instrumental I.

4.2.2 A SD proposta

Considerando os princípios e os procedimentos elencados (cf. seção 3.3), iniciamos a descrição da SD proposta tratando do gênero discursivo alvo, dos objetivos de aprendizagem, do *corpus* de textos do gênero-alvo, da extração dos dados linguísticos e da elaboração das tarefas propriamente ditas.

O passo inicial no desenho dos materiais didáticos a serem usados na disciplina é o conhecimento do perfil e das necessidades dos alunos. O nível de proficiência dos alunos, conforme o Quadro Comum Europeu (QCE), é B1-B2, o que permite que a SD seja escrita em inglês. Ressalta-se que, embora a SD tenha sido produzida especificamente para ser trabalhada com os alunos que cursam a disciplina, ela ainda não havia sido utilizada no momento de produção deste artigo. Como já mencionado, a disciplina tem como gênero estruturante o artigo de pesquisa. O objetivo de aprendizagem, definido tendo-se em vista a área do conhecimento do grupo de alunos (a Física), foi a produção da introdução de um artigo de pesquisa da área em questão. Definidos o gênero estruturante e o objetivo de aprendizagem, um *corpus* de estudo previamente compilado (cf. seção 3.1) foi usado para a extração dos dados linguísticos conforme já descrito (cf. seção 4.1). De posse dos dados linguísticos (*KLFs*) extraídos do *corpus* de estudo, iniciou-se a elaboração das tarefas. A SD, que contém as TPs propostas neste estudo,

foi estruturada em cinco partes: (i) sondagem, (ii) contextualização, (iii) estrutura retórica, (iv) elementos linguísticos, (v) produção. A referida SD está disponível em sua totalidade no Apêndice 3.

A sondagem tem por objetivo verificar o conhecimento prévio do aluno a respeito do gênero estruturante.

WHAT DO YOU KNOW ABOUT IT?

1. Look at the texts below:

   e. What type of texts can you see? What do you know about them?
   f. What are the parts of this type of text?
   g. What is usually the first part? What information should it contain?
   h. Do you think this information can vary from area to area?

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges

Chih-Cheng Lai[a], Tzu-Ping Shih[b], Wen-Chien Ko[c], Hung-Jen Tang[d], Po-Ren Hsueh[e,f,*]

[a] Department of Internal Medicine, Kaohsiung Veterans General Hospital, Tainan Branch, Tainan, Taiwan
[b] Department of Family Medicine, Kaohsiung Veterans General Hospital, Tainan Branch, Tainan, Taiwan
[c] Department of Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[d] Department of Medicine, Chi Mei Medical Center, Tainan 71004, Taiwan
[e] Department of Laboratory Medicine, National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan
[f] Department of Internal Medicine, National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 11 February 2020
Accepted 12 February 2020

Editor: Jean-Marc Rolain

Keywords:
2019-nCoV
SARS-CoV-2
COVID-19
China
Epidemic
Remdesivir

ABSTRACT

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; previously provisionally named 2019 novel coronavirus or 2019-nCoV) disease (COVID-19) in China at the end of 2019 has caused a large global outbreak and is a major public health issue. As of 11 February 2020, data from the World Health Organization (WHO) have shown that more than 43 000 confirmed cases have been identified in 28 countries/regions, with >99% of cases being detected in China. On 30 January 2020, the WHO declared COVID-19 as the sixth public health emergency of international concern. SARS-CoV-2 is closely related to two bat-derived severe acute respiratory syndrome-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21. It is spread by human-to-human transmission via droplets or direct contact, and infection has been estimated to have mean incubation period of 6.4 days and a basic reproduction number of 2.24–3.58. Among patients with pneumonia caused by SARS-CoV-2 (novel coronavirus pneumonia or Wuhan pneumonia), fever was the most common symptom, followed by cough. Bilateral lung involvement with ground-glass opacity was the most common finding from computed tomography images of the chest. The one case of SARS-CoV-2 pneumonia in the USA is responding well to remdesivir, which is now undergoing a clinical trial in China. Currently, controlling infection to prevent the spread of SARS-CoV-2 is the primary intervention being used. However, public health authorities should keep monitoring the situation closely, as the more we can learn about this novel virus and its associated outbreak, the better we can respond.

© 2020 Elsevier B.V. and International Society of Chemotherapy. All rights reserved.

1. Introduction

Since the emergence of the 2019 novel coronavirus (2019-nCoV) infection in Wuhan, China, in December 2019 [1], it has rapidly spread across China and many other countries [2–8]. So far, 2019-nCoV has affected more than 43 000 patients in 28 countries/regions and has became a major global health concern (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=6f80d1b9_4). On 11 February 2020, the World Health Organization (WHO) announced a new name for the epidemic disease caused by 2019-nCoV: coronavirus disease (COVID-19). Regarding the virus itself, the International Committee on Taxonomy of Viruses has renamed the previously provisionally named 2019-nCoV as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) [3].

Although early studies reported a link between a single local fish and wild animal market and most cases of infection, indicating possible animal-to-human transmission, studies have increasingly demonstrated human-to-human transmission of SARS-CoV-2 through droplets or direct contact [2,8–10]. Moreover, according to

A contextualização promove um primeiro contato do aluno com o gênero-alvo, já buscando construir o conhecimento a respeito de seu conceito e de suas características.

WARM UP QUESTIONS

4. Are you familiar with the genre Research Article? Do you ever read Research Articles from your area?
5. What is the purpose of this genre? What elements do you expect to find in these texts?
6. What do you think is the difference between a Research Article, a Review Article and an Opinion Article ?

AN INTRODUCTION TO A RESEARCH ARTICLE

1. Get in groups of three. Each member of the group will read the introduction of a different Research Article from a well-known academic journal about a topic of interest in your area. Decide within the group who is going to read which Research Article introduction and, during your first reading, skim the text and fill out the chart below.

What is the name of the journal?
When was the article published?
What is the title of the Research Article? What do you think the article is about?
What are the names of the authors?
How many sections does the article have?
What is the first section?
What information do you expect to find in the first section of the Research Articles?
How many words does the first section have?

2. Read the introduction of the Research Article assigned to you thoroughly and answer the questions below.

e. By reading the introduction, do you know what the article is about? Is it consistent with the title of the article?
f. Did you find all the elements you expected to find in the introduction of the Research Article? Did you find any others? If so, which ones?
g. List the relevant information presented in the introduction.
h. Does the introduction make you interested in reading the rest of the article? If so, identify the passages of the text that were important to create expectation about the rest.

As atividades relativas à estrutura retórica procuram deixar claro para o aluno quais as partes que constituem o gênero-alvo e qual sua função comunicativa. A extensão e a ordem das partes também são aspectos tratados nesta etapa da SD. Além disso, as tarefas desta SD buscam provocar a reflexão dos alunos em relação à qualidade do texto através do que se espera da seção *introdução* de um artigo científico, estimulando, assim, o pensamento crítico quanto às informações relevantes que devem estar contidas nessa seção. As tarefas se propõem não somente à compreensão da seção, mas também à formação de leitores de textos acadêmicos que utilizam estratégias e desenvolvem posicionamentos em relação ao que leem.

FINDING REGULARITIES

1. Read the text again, this time focusing on the structure of the introduction of the Research Article. Search for specific information to fill the chart below. Be prepared to report it to your classmates. This activity will help you prepare to later write and assess the introduction of your own article. An example is presented below:

RESEARCH ARTICLE INTRODUCTION COMMUNICATIVE FUNCTIONS	ARE THESE COMMUNICATIVE FUNCTIONS PRESENT IN THE INTRODUCTION YOU READ?	EXAMPLE OF HOW THE COMMUNICATIVE FUNCTION CAN BE EXPRESSED LINGUISTICALLY IN THE INTRODUCTION
Establishing a territory		
Step 1: Claiming centrality	( )yes ( )no	**plays an important role in the** play an important role in the **Ex:** "Here we provide an experimental proof that the light intensity **plays an important role in the** vertical distribution of seven Synechococcus spp. strains isolated from the littoral zone of Lake Constance in Germany."
Step 2: Making topic generalization/s	( )yes ( )no	it is well known that the **it has been shown that the** **Ex:** "These parameters vary substantially between different studies although **it has been shown that the** method used to test the materials influences the measured P binding capacity and thus the predicted performance of the filters as well as their predicted lifetime."
Step 3: Reviewing items of previous literature	( )yes ( )no	

Establishing a niche		
Step 1A: Indicating a gap or **Step 1B:** Adding to what is known	( )yes ( )no	**to the best of our knowledge** **Ex:** "Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and **to the best of our knowledge** there is no comprehensive model of cellular biomechanics."
Introducing the Present Work		
Step 1: Announcing present research descriptively and/or purposively	( )yes ( )no	of the present study is to purpose of this paper is to purpose of this study is to **the purpose of this study is** the aim of the present study aim of the present study was aim of this paper is to the aim of this paper is of the present study was to Ex: "**The purpose of this study is** to investigate the effects of the operating parameters on natural gas supersonic separation process, including the back pressure, inlet mass flow rates, inlet pressures and inlet temperatures."
Step 2: Presenting research questions or hypotheses	( )yes ( )no	
Step 3: Definitional clarifications	( )yes ( )no	
Step 4: Summarizing methods	( )yes ( )no	
Step 5: Announcing principal outcomes	( )yes ( )no	
Step 6: Stating the value of the present research	( )yes ( )no	
Step 7: Outlining the structure of the paper	( )yes ( )no	**this paper is organized as follows** the rest of the paper is the paper is organized as follows **Ex:** "**This paper is organized as follows** : First, we provide detailed explanation of the methodology of our LV shape restoration algorithm. Next we describe the experiments done on the 30 simulated samples and the 20 in vivo patient-specific models to test the performance of the algorithm, followed by a discussion on the implications of the experimental results."

Source: Based on Swales (1990, 2004) framework.

2. Now let's take a general look at the structure of the sub-genre "introduction of a Research Article". According to Bakhtin (2010, p. 262), "every particular utterance is individual, but every field of language use elaborates their relatively stable kinds of utterances, which we call discourse genres." Considering this statement, get together with the other members of your group and compare the table each one of you filled out in 3. Analyze the similarities and differences among the three introductions read. Fill a new table that summarizes the general structure.

RESEARCH ARTICLE INTRODUCTION COMMUNICATIVE FUNCTIONS	IS THIS COMMUNICATIVE FUNCTION PRESENT IN INTRODUCTION...			DOES THE GROUP FIND THIS COMMUNICATIVE FUNCTION RELEVANT TO COMPOSE AN INTRODUCTION? WHY?
	...1?	...2?	...3?	
Establishing a territory				
1: Claiming centrality				
2: Making topic generalization/s				
3: Reviewing items of previous literature				
Establishing a niche				
1A: Indicating a gap or **1B:** Adding to what is known				
Introducing the Present Work				
1: Announcing present research descriptively and/or purposively (presenting the aim of the study)				
2: Presenting research questions or hypotheses				
3: Definitional clarifications				
4: Summarizing methods				
5: Announcing principal outcomes				
6: Stating the value of the present research				
7: Outlining the structure of the paper				

Source: Based on Swales (1990, 2004) framework.

3. So far, we have become acquainted with Research Article introductions published in academic journals. Now you are going to select a research project you are involved with so that you can write the introduction of a Research Article about it.

Conhecendo a estrutura retórica do gênero-alvo na sua área de especialidade, inicia-se o trabalho com os elementos linguísticos usados para expressar as funções comunicativas de cada um dos movimentos retóricos (e passos) da seção do gênero em questão. A partir dos recursos linguísticos extraídos do *corpus*, a tarefa conduz o aluno a inferir a função de cada forma linguística no texto e a refletir sobre seu uso para mais adiante se colocar na posição de autor de textos no gênero abordado.

LANGUAGE ELEMENTS

1. You are now going to read two introductions of Research Articles from the area of Physics published in the PLOS ONE platform. Before you do so, discuss:

a. Do you know the PLOS ONE platform? Have you ever visited it? Tell your classmates what you know about it.
b. Open the platform website and check if it offers any guidelines to authors.

2. Read the introductions below and, using different colors, highlight the parts of the text that represent the communicative functions listed in item 3. In pairs, discuss the differences and similarities you identify. Check your answers with your other classmates afterwards.

Introduction A

Introduction

The deep sea, under 1000-m depth, is characterized by a high hydrostatic pressure (≈10 MPa), with, generally, a low temperature and a low organic-matter concentration. Laboratory experiments using pure cultures of piezophilic bacteria have highlighted microbial adaptations to high hydrostatic pressure. The adaptive traits include those related to growth [1,2], membrane [3] and storage lipids [4], membrane and soluble proteins [5,6], the respiratory-chain complexes [7,8], replication, transcription and translation [9,10]. Most isolated piezophilic bacteria belong to the genera: *Carnobacterium*, *Desulfovibrio*, *Marinitoga*, *Shewanella*, *Photobacterium*, *Colwellia*, *Moritella*, and *Psychromonas* within the Gammaproteobacteria subclass reviewed by Bartlett *et al.* [11].

Darkness is another major characteristic of this deep-sea environment that can be disturbed by a biological phenomenon named bioluminescence. Bioluminescence is the process by which living micro- or macro-organisms emit light. Amongst the bioluminescent organisms, marine luminous bacteria are ecologically versatile and can be found as free-living forms, epiphytes, saprophytes, parasites, symbionts in the light organs of fishes and squids, and commensals in the gut of various marine organisms [12,13,14]. Metagenomic analysis from deep eastern-Mediterranean water samples shows a surprising high number of *lux* genes directly involved in bioluminescence [15]. As far as we know, all-known marine bioluminescent bacteria are phylogenetically affiliated to the *Vibrio*, *Photobacterium* and *Shewanella* genera within the Gammaproteobacteria subclass [16]. Amongst them, *Photobacterium phosphoreum* is the predominant species found in the Mediterranean Sea [17].

Those of the most studied micro-organisms are, for piezophily, *Photobacterium profundum* SS9 (e.g. [18]), not known as luminous, and for bioluminescence, *P. phosphoreum* (e.g. [19]). Up to date, little information is available concerning potential physiological-adaptation mechanisms of luminous bacteria to hydrostatic pressure,

especially for both piezophily and bioluminescence. In this study, we used a bioluminescent strain isolated from Mediterranean deep-sea waters (sampled at 2200-m depth) and identified as *Photobacterium phosphoreum* ANT-2200 [20]. At this depth, the *in situ* conditions of pressure and temperature are about 22 MPa and 13°C, respectively. The purpose of this study is (1) to define temperature and pressure optima for growth and (2) to study pressure effect (0.1 versus 22 MPa, 13°C) on growth and bioluminescence activities of *P. phosphoreum* ANT-2200 using a new laboratory controlled hyperbaric system dedicated to high-pressure and bioluminescence studies.

Traditionally, a linear regression is used to determine the growth rate of a strain during the logarithmic phase. The logistic (or Verhulst) model [26] was used in this study to determine both the growth rate (r) and the maximum population density (K). This model gives a continuous function of optical density, fitting discrete experimental data measured during the bacterial growth. Its hypotheses take into account limited resources in the medium and are defined as:

The birth rate:

$$n(x) = \alpha - \beta x$$

Introduction B

Increasing interference due to multiple users and other signal sources is one of the fundamental problems in wireless communication and has been extensively studied for many years. Smart antenna systems decrease interference by adaptive beamforming techniques like minimum variance distortionless response (MVDR). It is one of the commonly utilized adaptive array beamforming techniques [1], but it is often not able to form nulls towards any nearby interference sources satisfactorily. Consequently, MVDR may lead to significant performance degradation in the case of unexpected interfering signals [2]. It is difficult and time consuming to solve these problems through conventional empirical approach, and sometimes, in the applied cases, is impractical. Recently, the employment of meta-heuristics algorithms has been growing instead of exhaustive and exact procedures in similar applications [3–7]. Consequently, meta-heuristics and exploratory methods need to provide mathematically reliable solutions for this complicated class of optimization problems. However, the performance of these algorithms is often unsatisfactory for cases with three or more interference sources due to issues such as premature convergence and lack of sufficient exploration. Several methods are suggested for increasing the search diversity of SGSA, such as increasing the initial number of leaders (initial kbest). However, this significantly increases computational complexity of the force equation in GSA without properly addressing the key issue of dominant agents, with large masses, causing premature convergence. Primary reason for the search pattern domination is because agents are allowed to exert a force proportional to their performance and most SGSA variants allow the best agents to consistently influence all agents. Therefore, this paper suggests a stochastic leader gravitational search algorithm (SL-GSA) to enhance MVDR beamforming performance by preventing premature convergence and improving overall exploration. Standard gravitational search algorithm (SGSA) [8] was proposed as a global optimization method for computationally complex real world problems. In SGSA, the particles, called agents, move based on Newton's law of universal gravitation. The search space is represented as an 'n' dimensional space and the position of each agent is represented by a coordinate vector of length n. The mass of these agents are determined based on their fitness. The performance of each agent is calculated using the fitness function and their positions are updated accordingly. All the SGSA search agents (individuals) globally move toward the agents with heavier masses due to their gravitational force. Hence, superior solutions of the problems are represented by the heavier masses. The global search ability and high performance of SGSA in solving several nonlinear functions have been confirmed previously [8]. The balance between exploration and exploitation is critical for heuristic algorithms to achieve robust and reliable performance. In SGSA, this balance is achieved using the time variant linearly decreasing kbest parameter, which determines the number of agents that are allowed to exert force on the others in a given iteration. Thus, the parameter kbest is initially large and linearly reduced to provide some protection from premature convergence. This technique still allows the optimization process to be heavily influenced by agents with superior fitness resulting in poor exploration properties. As kbest agents are chosen based on their current fitness, it allows agents with superior fitness to attract the others towards optimal solutions. Thus, the algorithm is highly dependent on the best performing agents. However, if the kbest agents stagnate at a local optimum, the other agents become practically helpless to prevent premature convergence. The SGSA agents gravitate towards 'kbest' optimum agents. This allows convergence towards superior solutions but also allows the search to stagnate at local optima. In this paper, SL-GSA randomly selects agents from a gradually reducing set that removes agents with inferior performance based on the adaptive parameter, γ. This directly prevents the domination of the search pattern by any individual agent. Thus, SL-GSA is far less likely to stagnate in a local optimum because it randomly ignores the best particles sometimes. This allows more efficient exploration before final convergence. The proposed new parameter, γ, prevents selection of the agents with the worse fitness in the later part of the optimization. This, in conjunction with the linear decrease of the parameter k, allows SL-GSA to converge faster than SGSA. This is verified by applying the proposed algorithm to six benchmark functions and two case studies of MVDR beamforming technique. High performance of convergence and quality of final solution compared to original algorithm is achieved as discussed in simulation results. The rest of this paper is organized as follows: Section 2 introduces the brief review of SGSA. The proposed SL-GSA is presented in section 3. The basics of adaptive beamforming and the conventional MVDR technique are explained in section 4 and 5, respectively. The testing of the proposed SL-GSA via benchmark functions and the simulation results obtained via SGSA and its variants are reported in section 6. Section 7 shows the incorporation of MVDR in SL-GSA. The efficiency of SL-GSA for different interferences in two case studies is also reported in this section. Finally, Section 8 concludes this investigation.

3. The key lexical bundles (*KLBs*) below were extracted from a corpus of Physics Research Article Introductions compiled from PLOS ONE platform. They were some of the most frequent in the corpus. With a classmate, complete the chart below according to the examples given. You can choose from the communicative functions listed below:

a) **Establishing a territory**: claiming centrality
b) **Establishing a territory**: making topic generalization/s
c) **Establishing a niche:** indicating a gap or adding to what is known
d) **Introducing the present work:** stating the purpose of the study
e) **Introducing the Present Work:** outlining the structure of the paper

Key Lexical Bundles	Likely communicative function in the text	Likely location in the text
...it is well known that the...		
...purpose of this paper is to...		
...this paper is organized as follows...		
...plays an important role in the...		
...of the present study was to...		
...the rest of the paper is...		
it has been shown that the...		
...the aim of this paper is...		

Source: Based on Swales (1990, 2004) framework.

YOUR TURN

1. You are going to write an introduction for a Research Article that has to do with your research project in the area of Physics. Before that, get in groups and make a list of the indispensable elements to write a good introduction. You may consult the table under the **Finding Regularities** section to help build your list.

a. _____
b. _____
c. _____
d. _____
e. _____
f. _____
g. _____
h. _____
i. _____
j. _____

4. All the sentences below were taken from research articles from the area of Physics. See which KLB best completes each sentence.

	Examples		KLBs
1	"Here we provide an experimental proof that the light intensity is _____ vertical distribution of seven Synechococcus spp. strains isolated from the littoral zone of Lake Constance in Germany."[2]	( )	to the best of our knowledge
2	"These parameters vary substantially between different studies although _____ method used to test the materials influences the measured P binding capacity and thus the predicted performance of the filters as well as their predicted lifetime."	( )	plays an important role in the
3	"Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and _____ there is no comprehensive model of cellular biomechanics."	( )	The purpose of this study is
4	"_____ to investigate the effects of the operating parameters on natural gas supersonic separation process, including the back pressure, inlet mass flow rates, inlet pressures and inlet temperatures."	( )	This paper is organized as follows
5	"_____ First, we provide detailed explanation of the methodology of our LV shape restoration algorithm. Next we describe the experiments done on the 30 simulated samples and the 20 in vivo patient-specific models to test the performance of the algorithm, followed by a discussion on the implications of the experimental results."	( )	it has been shown that the

Por fim, o aluno é convidado a produzir um primeiro rascunho do gênero-alvo. Esse rascunho poderá receber *feedback* dos pares e do professor tendo em vista uma rubrica de avaliação previamente preparada e discutida (WELP; DIDIO; FINKLER, 2019). Depois de receber o *feedback* dos pares e do professor, o aluno poderá fazer a reescrita de seu texto.

YOUR TURN

1. You are going to write an introduction for a Research Article that has to do with your research project in the area of Physics. Before that, get in groups and make a list of the indispensable elements to write a good introduction. You may consult the table under the **Finding Regularities** section to help build your list.

a. _____
b. _____
c. _____
d. _____
e. _____
f. _____
g. _____
h. _____
i. _____
j. _____

2. Next, decide with the whole class which elements are going to be part of the assessment criteria of your introductions.

a. _____
b. _____
c. _____
d. _____
e. _____
f. _____
g. _____
h. _____
i. _____
j. _____

3. Write the first version of your introduction and bear in mind the following:
   a. What you are writing about
   b. Who you are writing to
   c. How you are organizing your text
   d. What language you are using
   e. Where you are publishing it

4. Look at the table below. When writing the first version of your introduction, answer the questions below about your project using the Key Lexical Frames (KLFs) suggested having the examples provided as a reference.

Questions	KLF	Example	Your sentence
What is the importance of the present study?	**____(play/plays) an important role in the ____**	**Ex:** "Here we provide experimental proof that the light intensity **plays an important role in the** vertical distribution of seven Synechococcus spp. strains isolated from the littoral zone of Lake Constance in Germany."[2]	
What other studies have shown?	**_____ it (is well known/has been shown) that the ____**	**Ex:** "These parameters vary substantially between different studies although **it has been shown that the** method used to test the materials influences the measured P binding capacity and thus the predicted performance of the filters as well as their predicted lifetime."	
Is there any gap in the present studies?	**_____ to the best of our knowledge _____**	**Ex:** "Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and **to the best of our knowledge** there is no comprehensive model of cellular biomechanics."	

Is there any gap in the present studies?	**_____ to the best of our knowledge _____**	**Ex:** "Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and **to the best of our knowledge** there is no comprehensive model of cellular biomechanics."	
What is the purpose of the study?	**The (purpose/aim) of (this /the present) (study/paper) (is/was) to _____**	**Ex:** "**The purpose of this study is** to investigate the effects of the operating parameters on natural gas supersonic separation process, including the back pressure, inlet mass flow rates, inlet pressures and inlet temperatures."	
What is the structure of the paper?	**(The rest of the/This) paper is organized as follows _____**	**Ex:** "**This paper is organized as follows** : First, we provide a detailed explanation of the methodology of our LV shape restoration algorithm. Next we describe the experiments done on the 30 simulated	

5. After you write the first version of your introduction, exchange it with a classmate and use the rubric built by the group to give suggestions and recommendations to help them improve their text.

Por fim, observa-se que a SD propicia a mobilização dos conhecimentos necessários para a produção textual, incluindo as características do gênero alvo e os recursos lexicais e fraseológicos próprios dele. Ainda, a produção escrita dos alunos somente é considerada finalizada após o aprimoramento do texto através de mais de uma oportunidade de reescrita, as quais oferecem momentos de aprendizagem por meio de reflexão sobre os recursos linguísticos utilizados.

5 Considerações finais

Este estudo teve dois objetivos específicos. O primeiro, de ordem analítica, consistiu em extrair, categorizar e classificar expressões multipalavra nos *corpora* especializados de textos da seção *introdução* dos artigos de pesquisa compilados. O segundo, de ordem pedagógica, foi usar os dados linguísticos coletados para informar a construção de TPs voltadas para o ensino de IFA.

As TPs resultantes deste estudo, disponibilizadas *on-line* e de forma gratuita no Ambiente Virtual de Aprendizagem LÚMINA Idiomas (BOCORNY, 2017), podem ser utilizadas pela comunidade acadêmica de forma autônoma ou em aulas presenciais, em disciplinas de Inglês Instrumental, nos cursos de IFA do Centro de Línguas para Fins Acadêmicos da UFRGS (CLA-UFRGS), ou, ainda, por professores de disciplinas de diferentes áreas do conhecimento e de diferentes instituições que utilizem o inglês como meio de instrução ou que desejem incentivar seus alunos a produzirem artigos em inglês durante seus cursos.

Espera-se que a utilização das TPs construídas com dados linguísticos obtidos neste estudo, aliada a outras iniciativas, levem ao aprimoramento de pesquisadores brasileiros e a um consequente aumento no impacto dos artigos produzidos no Brasil. Por fim, pesquisas futuras buscarão ampliar, para outras áreas do conhecimento e para outros gêneros acadêmicos, a descrição linguística e a construção de TPs propostas no presente estudo.

Agradecimentos

Este trabalho foi conduzido durante o período de concessão da bolsa de Professor Visitante no Exterior na Universidade do Norte do Arizona (EUA) e financiado pelo Programa Institucional de Internacionalização da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES/PRINT/UFRGS – Edital nº 003/2019, no âmbito do Ministério da Educação do Brasil.

Declaração das contribuições de cada autora

As autoras Ana Eliza Pereira Bocorny e Anamaria Welp produziram colaborativamente este artigo.

Referências

AMMON, U. (ed.). *The Dominance of English as a Language of Science*: Effects on Other Languages and Language Communities. Berlin: Walter de Gruyter, 2011.

ANTHONY, L. *AntCorGen (Version 1.1.2) [Computer Software]*. Tokyo: Waseda University, 2019. Disponível em: https://www.laurenceanthony.net/software. Acesso em: 10 out. 2020.

BAKHTIN, M. *Estética da criação verbal*. 5. ed. São Paulo: Martins Fontes, 2010.

BAUMVOL, L. K. *Language Practices for Knowledge Production and Dissemination*: The Case of Brazil. 2018. 270f. Tese (Doutorado em Estudos da Linguagem) – Instituto de Letras, Universidade Federal do Rio Grande do Sul, 2018. Disponível em: https://lume.ufrgs.br/bitstream/handle/10183/189174/001088580.pdf?sequence=1&isAllowed=y. Acesso em: 10 out. 2020.

BHATIA, V. *Analysing Genre:* Language Use in Professional Settings. London: Longman, 2001.

BIBER, D. A Corpus-Driven Approach to Formulaic Language in English. *International Journal of Corpus Linguistics*, [*S.l.*], v. 14, n. 3, p. 275-311, 2009. DOI: https://doi.org/10.1075/ijcl.14.3.08bib

BIBER, D; CONRAD, S. Lexical Bundles in Conversation and Academic Prose. *Language and Computers*, [*S.l.*], v. 26, p. 181-190, 1999.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S; FINEGAN, E. *Longman Grammar of Spoken and Written English.* Harlow: Pearson, 1999.

BIBER, D.; CONRAD, S.; REPPEN, R.; BYRD, P.; HELT, M. Speaking and Writing in the University: A Multidimensional Comparison. *Tesol Quarterly,* [*S.l.*], v. 36, n. 1, p. 9-48, 2002. DOI: https://doi.org/10.2307/3588359

BIBER, D.; CONRAD, S.; CORTES, V. If You Look at…: Lexical Bundles in University Teaching and Textbooks. *Applied Linguistics*, [*S.l.*], v. 25, n. 3, p. 371-405, 2004. DOI: https://doi.org/10.1093/applin/25.3.371

BIBER, D.; BARBIERI, F. Lexical Bundles in University Spoken and Written Registers. *English for Specific Purposes*, [*S.l.*], v. 26, n. 3, p. 263-286, 2007. DOI: https://doi.org/10.1016/j.esp.2006.08.003

BOCORNY, A. *LUMINA Idiomas*, 2017. Página inicial. Disponível em: https://www.ufrgs.br/luminaidiomas/. Acesso em: 12 jan. 2021.

BOCORNY, A. E. P.; REBECHI, R.; REPPEN, R.; DELFINO, M. C. N.; LAMEIRA, V. A produção de artigos da área médica e das ciências da

saúde com o auxílio de *key lexical bundles*: um estudo direcionado por *corpus*. *DELTA,* São Paulo. No prelo.

BOCORNY, A. E. P.; KILLIAN, C. K. Contexto e pressupostos teóricos para a elaboração de uma plataforma online amigável, de livre acesso e com recurso multimídia em uma universidade brasileira. *Trama*, Marechal Cândido Rondon, PR, v. 13, n. 28, p. 4-28, 2017. DOI: https://doi.org/10.48075/rt.v13i28.15585

CORTES, V. The Purpose of this Study Is to: Connecting Lexical Bundles and Moves in Research Article Introductions. *Journal of English for Academic Purposes*, [*S.l.*], v. 12, n. 1, p. 33-43, 2013. DOI: https://doi.org/10.1016/j.jeap.2012.11.002

DENARDI, D. A. C. Didactic Sequence: A Dialetic Mechanism for Language Teaching and Learning. *Revista Brasileira de Linguística Aplicada*, Belo Horizonte, v. 17, n. 1, p. 163-184, 2017. DOI: https://doi.org/10.1590/1984-6398201610012

DOMINGUES, M. L.; FAVERO, E. L.; MEDEIROS, I. P. Etiquetagem de palavras para o português do Brasil. *In:* TIL –WORKSHOP EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, V., 2007, Rio de Janeiro. *Anais* [...]. Rio de Janeiro: SBC, 2007. p. 1721-1724.

FRANKENBERG-GARCIA; A., BOCORNY, A. E. P.; TAVARES-PINTO, P.; SARMENTO, S. *Supporting the Internationalisation of Brazilian Research*: Curso oferecido via financiamento *Capes:Print* para as Universidade Federal do Rio Grande do Sul e Universidade Estadual Paulista, 04-06 de jun.de 2019. 30f. Notas de aula.

GIBBONS, P. Scaffolding. *In:* ROBINSON, P. (ed.). *The Routledge Encyclopedia of Second Language Acquisition*. London: Routledge, 2013. p. 563-564.

GRAY, B.; BIBER, D. Lexical Frames in Academic Prose and Conversation. *International Journal of Corpus Linguistics*, [*S.l.*], v. 18, n. 1, p. 109-136, 2013. DOI: https://doi.org/10.1075/ijcl.18.1.08gra

GREEN, S. *Scaffolding Academic Literacy with Low-Proficiency Users of English*. Londres: Palgrave Macmillan, 2020. DOI: https://doi.org/10.1007/978-3-030-39095-2_3

HYLAND, K. *Disciplinary Identities:* Individuality and Community in Academic Discourse. Munique: Ernst Klett Sprachen, 2012.

JENKINS, J. English as a lingua franca: Interpretations and Attitudes. *World Englishes*, [*S.l.*], v. 28, n. 2, p. 200-207, 2009. DOI: 10.1111/j.1467-971X.2009.01582.x

KILGARRIFF, A.; RYCHLY, P.; SMRZ, P.; TUGWELL, D. The Sketch Engine. *In*: EURALEX INTERNACIONAL CONGREE, 11th., 2004, Bretagne-Sud. *Proceedings* […]. Bretagne-Sud: Université de Bretagne-Sud, 2004. p. 105-116.

KOSTLA, I.; BUNNING, L. Curriculum Design in English Language Teaching. *ELT Development Series*. Alexandria, VA: TESOL Press, 2017.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. *In:* DIONÍSIO, A. P.; MACHADO, A. R.; BEZERRA, M. A. (org.). *Gêneros textuais & ensino*. Rio de Janeiro: Editora Lucerna, 2002. p. 19-36.

MENEGHINI, R.; PACKER, A. L. Is There Science Beyond English? *EMBO Reports*, [*S.l.*], v. 8, n. 2, p. 112-116, 2007. DOI: https://doi.org/10.1038/sj.embor.7400906

MORLEY, J. *Academic Phrasebank*. Manchester: University of Manchester, 2014. Disponível em: http://www.phrasebank.manchester.ac.uk/. Acesso em: 10 out. 2020.

MORLEY, J. *The Academic Phrasebank:* An Academic Writing Resource for Students and Researchers. Manchester: The University of Manchester, 2017.

SANTIN, D. M.; VANZ, S. A. S.; STUMPF, I. R. C. Internacionalização da produção científica brasileira: políticas, estratégias e medidas de avaliação. *Revista Brasileira de Pós-Graduação*, Brasília, DF, v. 13, n. 30, p. 81-100, 2016. DOI: https://doi.org/10.21713/2358-2332.0.923

SARDINHA, T. B. Análise multidimensional. *DELTA,* São Paulo, v. 16, n. 1, p. 99-127, 2004. DOI: https://dx.doi.org/10.1590/S0102-44502000000100005.

SCHNEUWLY, B.; DOLZ, J. (org.). *Gêneros orais e escritos na escola.* Campinas: Mercado de Letras, 2004, 278p.

SINCLAIR, J. M. H. *Corpus, Concordance, Collocation*. Oxford: Oxford Univ. Press, 1991.

STAPLES, S.; EGBERT, J.; BIBER, D. Formulaic Sequences and EAP Writing Development: Lexical Bundles in the TOEFL iBT Writing Section. *Journal of English for Academic Purposes*, [*S.l.*], v. 12, n. 3, p. 214-225, 2013. DOI: https://doi.org/10.1016/j.jeap.2013.05.002

SWALES, J. *Research Genres:* Explorations and Applications. Cambridge: Cambridge University Press, 2004. DOI: https://doi.org/10.1017/CBO9781139524827

SWALES, J. *Genre Analysis:* English in Academic and Research Settings. Cambridge: Cambridge University Press, 1990.

TARDY, C. The Role of English in Scientific Communication: Lingua Franca or Tyrannosaurus Rex? *Journal of English for Academic Purposes*, [*S.l.*], v. 3, n. 3, p. 247-269, 2004. DOI: https://doi.org.10.1016/j.jeap.2003.10.001

VAN DEN BRANDEN, K. (ed.). *Task-Based Language Education*: From Theory to Practice. Cambridge: Cambridge University Press, 2006. DOI: https://doi.org/10.1017/CBO9780511667282

VAN DEN BRANDEN, K. Task-Based Language Teaching. *In:* HALL, G. (ed.). *The Routledge Handbook of English Language Teaching*. New York: Routledge, 2016. p. 238-251.

VYGOTSKY, L. S. *A formação social da mente*. São Paulo: Martins Fontes, 1998.

WELP, A. K. S.; DIDIO, Á. R.; FINKLER, B. Questões contemporâneas no cinema e na literatura: o desenho de uma sequência didática para o ensino de inglês como língua adicional. *Brazilian English Language Teaching Journal*, Porto Alegre, v. 10, n. 2, p. 1-25, 2019. DOI: https://doi.org/10.15448/2178-3640.2019.2.35861

WOOD, D.; BRUNER, J. S.; ROSS, G. The Role of Tutoring in Problem Solving. *Journal of Child Psychology and Psychiatry*, London, v. 17, n. 2, p. 89-100, 1976. DOI: https://doi.org/10.1111/j.1469-7610.1976.tb00381.x

APÊNDICES

APÊNDICE 1 – Resultado da extração de *KLBs* com 6 palavras da seção *introdução* de artigos de pesquisa da área da Física, realizada com base nos critérios apresentados em 3.2

KLB **(6 palavras)**	***Corpus* de estudo**		***Corpus* de referência**		**IC**
	Frequência Absoluta	**Frequência Normalizada**	**Frequência Absoluta**	**Frequência Normalizada**	
it is well known that the	8	6.4	6.0	1.0	3.8
it has been shown that the	11	8.9	12.0	1.9	3.4
of the present study is to	14	11.3	26.0	4.1	2.4
purpose of this paper is to	8	6.4	21.0	3.3	1.7
purpose of this study is to	9	7.2	25.0	4.0	1.7
this paper is organized as follows	21	16.9	62.0	9.9	1.6
the purpose of this study is	8	6.4	23.0	3.7	1.6
the aim of the present study	12	9.7	36.0	5.7	1.6
to the best of our knowledge	41	33.0	129.0	20.5	1.6
plays an important role in the	12	9.7	38.0	6.0	1.5
the rest of the paper is	9	7.2	32.0	5.1	1.4
play an important role in the	16	12.9	62.0	9.9	1.3
aim of the present study was	9	7.2	36.0	5.7	1.2
the paper is organized as follows	13	10.5	56.0	8.9	1.2
aim of this paper is to	10	8.0	45.0	7.2	1.1
the aim of this paper is	9	7.2	42.0	6.7	1.1
of the present study was to	15	12.1	72.0	11.4	1.1

APÊNDICE 2 – Resultado da categorização de *KLBs* com 6 palavras da seção *introdução* de artigos de pesquisa da área da Física, realizada com base nos critérios apresentados em 3.2

KLB **(6 palavras)**	***Corpus* de estudo**		***Corpus* de referência**		**IC**
	Frequência Absoluta	**Frequência Normalizada**	**Frequência Absoluta**	**Frequência Normalizada**	
it is well known that the	8	6.4	6.0	1.0	3.8
it has been shown that the	11	8.9	12.0	1.9	3.4
of the present study is to	14	11.3	26.0	4.1	2.4
purpose of this paper is to	8	6.4	21.0	3.3	1.7
purpose of this study is to	9	7.2	25.0	4.0	1.7
this paper is organized as follows	21	16.9	62.0	9.9	1.6
the purpose of this study is	8	6.4	23.0	3.7	1.6
the aim of the present study	12	9.7	36.0	5.7	1.6
to the best of our knowledge	41	33.0	129.0	20.5	1.6
plays an important role in the	12	9.7	38.0	6.0	1.5
the rest of the paper is	9	7.2	32.0	5.1	1.4
play an important role in the	16	12.9	62.0	9.9	1.3
aim of the present study was	9	7.2	36.0	5.7	1.2
the paper is organized as follows	13	10.5	56.0	8.9	1.2
aim of this paper is to	10	8.0	45.0	7.2	1.1
the aim of this paper is	9	7.2	42.0	6.7	1.1
of the present study was to	15	12.1	72.0	11.4	1.1

APÊNDICE 3 – Tarefa proposta

Inglês para Fins Acadêmicos (IFA)

Introdução de artigo de pesquisa

Nível QCE B1 - B2	Tempo 20 - 30 Minutos	Habilidades Leitura e Escrita

Objetivo
Aprender a escrever a introdução de um artigo de pesquisa da área da Física

Área
Física

WHAT DO YOU KNOW ABOUT IT?

1. Look at the texts below:

a. What type of texts can you see? What do you know about them?
b. What are the parts of this type of text?
c. What is usually the first part? What information should it contain?
d. Do you think this information can vary from area to area?

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges

Chih-Cheng Lai[a], Tzu-Ping Shih[b], Wen-Chien Ko[c], Hung-Jen Tang[d], Po-Ren Hsueh[e,f,*]

[a] Department of Internal Medicine, Kaohsiung Veterans General Hospital, Tainan Branch, Tainan, Taiwan
[b] Department of Family Medicine, Kaohsiung Veterans General Hospital, Tainan Branch, Tainan, Taiwan
[c] Department of Medicine, College of Medicine, National Cheng Kung University, Tainan, Taiwan
[d] Department of Medicine, Chi Mei Medical Center, Tainan 71004, Taiwan
[e] Department of Laboratory Medicine, National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan
[f] Department of Internal Medicine, National Taiwan University Hospital, National Taiwan University College of Medicine, Taipei, Taiwan

ARTICLE INFO

Article history:
Received 11 February 2020
Accepted 12 February 2020

Editor: Jean-Marc Rolain

Keywords:
2019-nCoV
SARS-CoV-2
COVID-19
China
Epidemic
Remdesivir

ABSTRACT

The emergence of severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2; previously provisionally named 2019 novel coronavirus or 2019-nCoV) disease (COVID-19) in China at the end of 2019 has caused a large global outbreak and is a major public health issue. As of 11 February 2020, data from the World Health Organization (WHO) have shown that more than 43 000 confirmed cases have been identified in 28 countries/regions, with >99% of cases being detected in China. On 30 January 2020, the WHO declared COVID-19 as the sixth public health emergency of international concern. SARS-CoV-2 is closely related to two bat-derived severe acute respiratory syndrome-like coronaviruses, bat-SL-CoVZC45 and bat-SL-CoVZXC21. It is spread by human-to-human transmission via droplets or direct contact, and infection has been estimated to have mean incubation period of 6.4 days and a basic reproduction number of 2.24–3.58. Among patients with pneumonia caused by SARS-CoV-2 (novel coronavirus pneumonia or Wuhan pneumonia), fever was the most common symptom, followed by cough. Bilateral lung involvement with ground-glass opacity was the most common finding from computed tomography images of the chest. The one case of SARS-CoV-2 pneumonia in the USA is responding well to remdesivir, which is now undergoing a clinical trial in China. Currently, controlling infection to prevent the spread of SARS-CoV-2 is the primary intervention being used. However, public health authorities should keep monitoring the situation closely, as the more we can learn about this novel virus and its associated outbreak, the better we can respond.

© 2020 Elsevier B.V. and International Society of Chemotherapy. All rights reserved.

1. Introduction

Since the emergence of the 2019 novel coronavirus (2019-nCoV) infection in Wuhan, China, in December 2019 [1], it has rapidly spread across China and many other countries [2–8]. So far, 2019-nCoV has affected more than 43 000 patients in 28 countries/regions and has became a major global health concern (https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200211-sitrep-22-ncov.pdf?sfvrsn=6f80d1b9_4). On 11 February 2020, the World Health Organization (WHO) announced a new name for the epidemic disease caused by 2019-nCoV: coronavirus disease (COVID-19). Regarding the virus itself, the International Committee on Taxonomy of Viruses has renamed the previously provisionally named 2019-nCoV as severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) [3].

Although early studies reported a link between a single local fish and wild animal market and most cases of infection, indicating possible animal-to-human transmission, studies have increasingly demonstrated human-to-human transmission of SARS-CoV-2 through droplets or direct contact [2,8–10]. Moreover, according to

Corpus linguistics, newspaper archives and historical research methods

Chinmay Tumbe
Economics Area, Indian Institute of Management, Ahmedabad, India

Received 31 January 2018
Revised 10 July 2018
4 November 2018
12 February 2019
Accepted 27 February 2019

Abstract

Purpose – The purpose of this paper is to demonstrate the utility of corpus linguistics and digitised newspaper archives in management and organisational history.

Design/methodology/approach – The paper draws its inferences from Google NGram Viewer and five digitised historical newspaper databases – The Times of India, The Financial Times, The Economist, The New York Times and The Wall Street Journal – that contain prints from the nineteenth century.

Findings – The paper argues that corpus linguistics or the quantitative and qualitative analysis of large-scale real-world machine-readable text can be an important method of historical research in management studies, especially for discourse analysis. It shows how this method can be fruitfully used for research in management and organisational history, using term count and cluster analysis. In particular, historical databases of digitised newspapers serve as important corpora to understand the evolution of specific words and concepts. Corpus linguistics using newspaper archives can potentially serve as a method for periodisation and triangulation in corporate, analytically structured and serial histories and also foster cross-country comparisons in the evolution of management concepts.

Research limitations/implications – The paper also shows the limitation of the research method and potential robustness checks while using the method.

Practical implications – Findings of this paper can stimulate new ways of conducting research in management history.

Originality/value – The paper for the first time introduces corpus linguistics as a research method in management history.

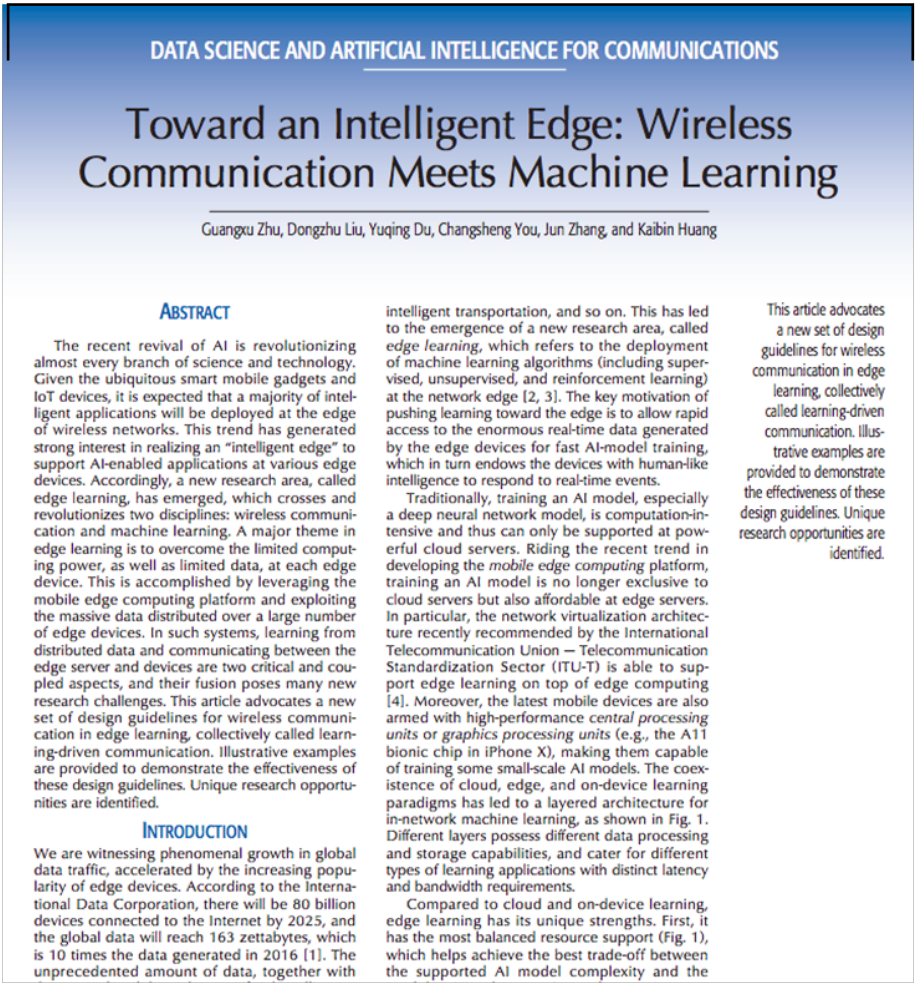
Keywords Management history, Corpus linguistics, Text analysis, Research methodology, Newspaper

Paper type Research paper

Introduction

With the advent of mass digitisation of historical prints over the past two decades, researchers now have access to a unique but under-appreciated source of data – Words. Billions of words have entered online repositories as hundreds of millions of pages of old books and newspapers have been digitised by firms and organisations. Words in these real-world texts form a "corpus" and corpus linguistics refers to a branch of linguistic studies that systematically analyses them (Biber *et al.*, 1998; Oakes, 1998; McEnery and Wilson, 2001). At its simplest, it can be defined as a "methodology that uses computer support – in particular, software called 'concordance programs'- to analyse authentic, and usually very large, volumes of textual data" (Mautner, 2009, p. 122). Alternatively, it may be defined as "the study of language data on a large scale – the computer-aided analysis of very extensive collections of transcribed utterances or written texts" (McEnery and Hardie, 2012). Both these definitions highlight the existence of a corpus of text that is sufficiently large in scale and machine-readable in nature. Scale is important to differentiate it from standard

Journal of Management History
Vol. 25 No. 4, 2019
pp. 533-549
© Emerald Publishing Limited
1751-1348
DOI 10.1108/JMH-01-2018-0009



DATA SCIENCE AND ARTIFICIAL INTELLIGENCE FOR COMMUNICATIONS

Toward an Intelligent Edge: Wireless Communication Meets Machine Learning

Guangxu Zhu, Dongzhu Liu, Yuqing Du, Changsheng You, Jun Zhang, and Kaibin Huang

ABSTRACT

The recent revival of AI is revolutionizing almost every branch of science and technology. Given the ubiquitous smart mobile gadgets and IoT devices, it is expected that a majority of intelligent applications will be deployed at the edge of wireless networks. This trend has generated strong interest in realizing an "intelligent edge" to support AI-enabled applications at various edge devices. Accordingly, a new research area, called edge learning, has emerged, which crosses and revolutionizes two disciplines: wireless communication and machine learning. A major theme in edge learning is to overcome the limited computing power, as well as limited data, at each edge device. This is accomplished by leveraging the mobile edge computing platform and exploiting the massive data distributed over a large number of edge devices. In such systems, learning from distributed data and communicating between the edge server and devices are two critical and coupled aspects, and their fusion poses many new research challenges. This article advocates a new set of design guidelines for wireless communication in edge learning, collectively called learning-driven communication. Illustrative examples are provided to demonstrate the effectiveness of these design guidelines. Unique research opportunities are identified.

INTRODUCTION

We are witnessing phenomenal growth in global data traffic, accelerated by the increasing popularity of edge devices. According to the International Data Corporation, there will be 80 billion devices connected to the Internet by 2025, and the global data will reach 163 zettabytes, which is 10 times the data generated in 2016 [1]. The unprecedented amount of data, together with intelligent transportation, and so on. This has led to the emergence of a new research area, called *edge learning*, which refers to the deployment of machine learning algorithms (including supervised, unsupervised, and reinforcement learning) at the network edge [2, 3]. The key motivation of pushing learning toward the edge is to allow rapid access to the enormous real-time data generated by the edge devices for fast AI-model training, which in turn endows the devices with human-like intelligence to respond to real-time events.

Traditionally, training an AI model, especially a deep neural network model, is computation-intensive and thus can only be supported at powerful cloud servers. Riding the recent trend in developing the *mobile edge computing* platform, training an AI model is no longer exclusive to cloud servers but also affordable at edge servers. In particular, the network virtualization architecture recently recommended by the International Telecommunication Union — Telecommunication Standardization Sector (ITU-T) is able to support edge learning on top of edge computing [4]. Moreover, the latest mobile devices are also armed with high-performance *central processing units* or *graphics processing units* (e.g., the A11 bionic chip in iPhone X), making them capable of training some small-scale AI models. The coexistence of cloud, edge, and on-device learning paradigms has led to a layered architecture for in-network machine learning, as shown in Fig. 1. Different layers possess different data processing and storage capabilities, and cater for different types of learning applications with distinct latency and bandwidth requirements.

Compared to cloud and on-device learning, edge learning has its unique strengths. First, it has the most balanced resource support (Fig. 1), which helps achieve the best trade-off between the supported AI model complexity and the

This article advocates a new set of design guidelines for wireless communication in edge learning, collectively called learning-driven communication. Illustrative examples are provided to demonstrate the effectiveness of these design guidelines. Unique research opportunities are identified.

WARM UP QUESTIONS

1. Are you familiar with the genre Research Article? Do you ever read Research Articles from your area?
2. What is the purpose of this genre? What elements do you expect to find in these texts?
3. What do you think is the difference between a Research Article, a Review Article and an Opinion Article ?

AN INTRODUCTION TO A RESEARCH ARTICLE

1. Get in groups of three. Each member of the group will read the introduction of a different Research Article from a well-known academic journal about a topic of interest in your area. Decide within the group who is going to read which Research Article introduction and, during your first reading, skim the text and fill out the chart below.

What is the name of the journal?
When was the article published?
What is the title of the Research Article? What do you think the article is about?
What are the names of the authors?
How many sections does the article have?
What is the first section?
What information do you expect to find in the first section of the Research Articles?
How many words does the first section have?

2. Read the introduction of the Research Article assigned to you thoroughly and answer the questions below.

a. By reading the introduction, do you know what the article is about? Is it consistent with the title of the article?

b. Did you find all the elements you expected to find in the introduction of the Research Article? Did you find any others? If so, which ones?

c. List the relevant information presented in the introduction.

d. Does the introduction make you interested in reading the rest of the article? If so, identify the passages of the text that were important to create expectation about the rest.

FINDING REGULARITIES

1. Read the text again, this time focusing on the structure of the introduction of the Research Article. Search for specific information to fill the chart below. Be prepared to report it to your classmates. This activity will help you prepare to later write and assess the introduction of your own article. An example is presented below:

RESEARCH ARTICLE INTRODUCTION COMMUNICATIVE FUNCTIONS	ARE THESE COMMUNICATIVE FUNCTIONS PRESENT IN THE INTRODUCTION YOU READ?	EXAMPLE OF HOW THE COMMUNICATIVE FUNCTION CAN BE EXPRESSED LINGUISTICALLY IN THE INTRODUCTION
Establishing a territory		
Step 1: Claiming centrality	( )yes ( )no	**plays an important role in the** play an important role in the **Ex:** "Here we provide an experimental proof that the light intensity **plays an important role in the** vertical distribution of seven Synechococcus spp. strains isolated from the littoral zone of Lake Constance in Germany."
Step 2: Making topic generalization/s	( )yes ( )no	it is well known that the **it has been shown that the** **Ex: "**These parameters vary substantially between different studies although **it has been shown that the** method used to test the materials influences the measured P binding capacity and thus the predicted performance of the filters as well as their predicted lifetime."
Step 3: Reviewing items of previous literature	( )yes ( )no	
Establishing a niche		
Step 1A: Indicating a gap or **Step 1B:** Adding to what is known	( )yes ( )no	**to the best of our knowledge** **Ex:** "Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and **to the best of our knowledge** there is no comprehensive model of cellular biomechanics."

Introducing the Present Work		
Step 1: Announcing present research descriptively and/or purposively	( )yes ( )no	of the present study is to purpose of this paper is to purpose of this study is to **the purpose of this study is** the aim of the present study aim of the present study was aim of this paper is to the aim of this paper is of the present study was to Ex: "**The purpose of this study is** to investigate the effects of the operating parameters on natural gas supersonic separation process, including the back pressure, inlet mass flow rates, inlet pressures and inlet temperatures."
Step 2: Presenting research questions or hypotheses	( )yes ( )no	
Step 3: Definitional clarifications	( )yes ( )no	
Step 4: Summarizing methods	( )yes ( )no	
Step 5: Announcing principal outcomes	( )yes ( )no	
Step 6: Stating the value of the present research	( )yes ( )no	
Step 7: Outlining the structure of the paper	( )yes ( )no	**this paper is organized as follows** the rest of the paper is the paper is organized as follows **Ex:** "**This paper is organized as follows :** First, we provide detailed explanation of the methodology of our LV shape restoration algorithm. Next we describe the experiments done on the 30 simulated samples and the 20 in vivo patient-specific models to test the performance of the algorithm, followed by a discussion on the implications of the experimental results."

Source: Based on Swales (1990, 2004) framework.

2. Now let's take a general look at the structure of the sub-genre "introduction of a Research Article". According to Bakhtin (2010, p. 262), "every particular utterance is individual, but every field of language use elaborates their relatively stable kinds of utterances, which we call discourse genres." Considering this statement, get together with the other members of your group and compare the table each one of you filled out in 3. Analyze the similarities and differences among the three introductions read. Fill a new table that summarizes the general structure.

RESEARCH ARTICLE INTRODUCTION COMMUNICATIVE FUNCTIONS	**IS THIS COMMUNICATIVE FUNCTION PRESENT IN INTRODUCTION...**			**DOES THE GROUP FIND THIS COMMUNICATIVE FUNCTION RELEVANT TO COMPOSE AN INTRODUCTION? WHY?**
	...1?	**...2?**	**...3?**	
Establishing a territory				
1: Claiming centrality				
2: Making topic generalization/s				
3: Reviewing items of previous literature				
Establishing a niche				
1A: Indicating a gap or **1B:** Adding to what is known				
Introducing the Present Work				
1: Announcing present research descriptively and/or purposively (presenting the aim of the study)				
2: Presenting research questions or hypotheses				
3: Definitional clarifications				
4: Summarizing methods				
5: Announcing principal outcomes				
6: Stating the value of the present research				
7: Outlining the structure of the paper				

Source: Based on Swales (1990, 2004) framework.

3. So far, we have become acquainted with Research Article introductions published in academic journals. Now you are going to select a research project you are involved with so that you can write the introduction of a Research Article about it.

LANGUAGE ELEMENTS

1. You are now going to read two introductions of Research Articles from the area of Physics published in the PLOS ONE platform. Before you do so, discuss:

   a. Do you know the PLOS ONE platform? Have you ever visited it? Tell your classmates what you know about it.
   b. Open the platform website and check if it offers any guidelines to authors.

2. Read the introductions below and, using different colors, highlight the parts of the text that represent the communicative functions listed in item 3. In pairs, discuss the differences and similarities you identify. Check your answers with your other classmates afterwards.

Introduction A

Introduction

The deep sea, under 1000-m depth, is characterized by a high hydrostatic pressure (≥10 MPa), with, generally, a low temperature and a low organic-matter concentration. Laboratory experiments using pure cultures of piezophilic bacteria have highlighted microbial adaptations to high hydrostatic pressure. The adaptive traits include those related to growth [1,2], membrane [3] and storage lipids [4], membrane and soluble proteins [5,6], the respiratory-chain complexes [7,8], replication, transcription and translation [9,10]. Most isolated piezophilic bacteria belong to the genera: *Carnobacterium*, *Desulfovibrio*, *Marinitoga*, *Shewanella*, *Photobacterium*, *Colwellia*, *Moritella*, and *Psychromonas* within the Gammaproteobacteria subclass reviewed by Bartlett *et al.* [11].

Darkness is another major characteristic of this deep-sea environment that can be disturbed by a biological phenomenon named bioluminescence. Bioluminescence is the process by which living micro- or macro-organisms emit light. Amongst the bioluminescent organisms, marine luminous bacteria are ecologically versatile and can be found as free-living forms, epiphytes, saprophytes, parasites, symbionts in the light organs of fishes and squids, and commensals in the gut of various marine organisms [12,13,14]. Metagenomic analysis from deep eastern-Mediterranean water samples shows a surprising high number of *lux* genes directly involved in bioluminescence [15]. As far as we know, all-known marine bioluminescent bacteria are phylogenetically affiliated to the *Vibrio*, *Photobacterium* and *Shewanella* genera within the Gammaproteobacteria subclass [16]. Amongst them, *Photobacterium phosphoreum* is the predominant species found in the Mediterranean Sea [17].

Those of the most studied micro-organisms are, for piezophily, *Photobacterium profundum* SS9 (e.g. [18]), not known as luminous, and for bioluminescence, *P. phosphoreum* (e.g. [19]). Up to date, little information is available concerning potential physiological-adaptation mechanisms of luminous bacteria to hydrostatic pressure,

especially for both piezophily and bioluminescence. In this study, we used a bioluminescent strain isolated from Mediterranean deep-sea waters (sampled at 2200-m depth) and identified as *Photobacterium phosphoreum* ANT-2200 [20]. At this depth, the *in situ* conditions of pressure and temperature are about 22 MPa and 13°C, respectively. The purpose of this study is (1) to define temperature and pressure optima for growth and (2) to study pressure effect (0.1 versus 22 MPa, 13°C) on growth and bioluminescence activities of *P. phosphoreum* ANT-2200 using a new laboratory controlled hyperbaric system dedicated to high-pressure and bioluminescence studies.

Traditionally, a linear regression is used to determine the growth rate of a strain during the logarithmic phase. The logistic (or Verhulst) model [26] was used in this study to determine both the growth rate (r) and the maximum population density (K). This model gives a continuous function of optical density, fitting discrete experimental data measured during the bacterial growth. Its hypotheses take into account limited resources in the medium and are defined as:

The birth rate:

$$n(x) = \alpha - \beta x$$

Introduction B

Increasing interference due to multiple users and other signal sources is one of the fundamental problems in wireless communication and has been extensively studied for many years. Smart antenna systems decrease interference by adaptive beamforming techniques like minimum variance distortionless response (MVDR). It is one of the commonly utilized adaptive array beamforming techniques [1], but it is often not able to form nulls towards any nearby interference sources satisfactorily. Consequently, MVDR may lead to significant performance degradation in the case of unexpected interfering signals [2]. It is difficult and time consuming to solve these problems through conventional empirical approach, and sometimes, in the applied cases, is impractical. Recently, the employment of meta-heuristics algorithms has been growing instead of exhaustive and exact procedures in similar applications [3–7]. Consequently, meta-heuristics and exploratory methods need to provide mathematically reliable solutions for this complicated class of optimization problems. However, the performance of these algorithms is often unsatisfactory for cases with three or more interference sources due to issues such as premature convergence and lack of sufficient exploration. Several methods are suggested for increasing the search diversity of SGSA, such as increasing the initial number of leaders (initial kbest). However, this significantly increases computational complexity of the force equation in GSA without properly addressing the key issue of dominant agents, with large masses, causing premature convergence. Primary reason for the search pattern domination is because agents are allowed to exert a force proportional to their performance and most SGSA variants allow

the best agents to consistently influence all agents. Therefore, this paper suggests a stochastic leader gravitational search algorithm (SL-GSA) to enhance MVDR beamforming performance by preventing premature convergence and improving overall exploration. Standard gravitational search algorithm (SGSA) [8] was proposed as a global optimization method for computationally complex real world problems. In SGSA, the particles, called agents, move based on Newton's law of universal gravitation. The search space is represented as an 'n' dimensional space and the position of each agent is represented by a coordinate vector of length n. The mass of these agents are determined based on their fitness. The performance of each agent is calculated using the fitness function and their positions are updated accordingly. All the SGSA search agents (individuals) globally move toward the agents with heavier masses due to their gravitational force. Hence, superior solutions of the problems are represented by the heavier masses. The global search ability and high performance of SGSA in solving several nonlinear functions have been confirmed previously [8]. The balance between exploration and exploitation is critical for heuristic algorithms to achieve robust and reliable performance. In SGSA, this balance is achieved using the time variant linearly decreasing kbest parameter, which determines the number of agents that are allowed to exert force on the others in a given iteration. Thus, the parameter kbest is initially large and linearly reduced to provide some protection from premature convergence. This technique still allows the optimization process to be heavily influenced by agents with superior fitness resulting in poor exploration properties. As kbest agents are chosen based on their current fitness, it allows agents with superior fitness to attract the others towards optimal solutions. Thus, the algorithm is highly dependent on the best performing agents. However, if the kbest agents stagnate at a local optimum, the other agents become practically helpless to prevent premature convergence. The SGSA agents gravitate towards 'kbest' optimum agents. This allows convergence towards superior solutions but also allows the search to stagnate at local optima. In this paper, SL-GSA randomly selects agents from a gradually reducing set that removes agents with inferior performance based on the adaptive parameter, γ. This directly prevents the domination of the search pattern by any individual agent. Thus, SL-GSA is far less likely to stagnate in a local optimum because it randomly ignores the best particles sometimes. This allows more efficient exploration before

final convergence. The proposed new parameter, γ, prevents selection of the agents with the worse fitness in the later part of the optimization. This, in conjunction with the linear decrease of the parameter k, allows SL-GSA to converge faster than SGSA. This is verified by applying the proposed algorithm to six benchmark functions and two case studies of MVDR beamforming technique. High performance of convergence and quality of final solution compared to original algorithm is achieved as discussed in simulation results. The rest of this paper is organized as follows: Section 2 introduces the brief review of SGSA. The proposed SL-GSA is presented in section 3. The basics of adaptive beamforming and the conventional MVDR technique are explained in section 4 and 5, respectively. The testing of the proposed SL-GSA via benchmark functions and the simulation results obtained via SGSA and its variants are reported in section 6. Section 7 shows the incorporation of MVDR in SL-GSA. The efficiency of SL-GSA for different interferences in two case studies is also reported in this section. Finally, Section 8 concludes this investigation.

3. The key lexical bundles (*KLBs*) below were extracted from a corpus of Physics Research Article Introductions compiled from PLOS ONE platform. They were some of the most frequent in the corpus. With a classmate, complete the chart below according to the examples given. You can choose from the communicative functions listed below:

a) Establishing a territory: claiming centrality
b) Establishing a territory: making topic generalization/s
c) Establishing a niche: indicating a gap or adding to what is known
d) Introducing the present work: stating the purpose of the study
e) Introducing the Present Work: outlining the structure of the paper

Key Lexical Bundles	Likely communicative function in the text	Likely location in the text
...it is well known that the...		
...purpose of this paper is to...		
...this paper is organized as follows...		
...plays an important role in the...		
...of the present study was to…		
...the rest of the paper is...		
it has been shown that the...		
...the aim of this paper is…		

Source: Based on Swales (1990, 2004) framework.

4. All the sentences below were taken from research articles from the area of Physics. See which KLB best completes each sentence.

	Examples		KLBs
1	"Here we provide an experimental proof that the light intensity is _____ vertical distribution of seven Synechococcus spp. strains isolated from the littoral zone of Lake Constance in Germany."[3]	( )	**to the best of our knowledge**
2	**"**These parameters vary substantially between different studies although _____method used to test the materials influences the measured P binding capacity and thus the predicted performance of the filters as well as their predicted lifetime."	( )	**plays an important role in the**
3	"Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and _____ there is no comprehensive model of cellular biomechanics."	( )	**The purpose of this study is**
4	"_____ to investigate the effects of the operating parameters on natural gas supersonic separation process, including the back pressure, inlet mass flow rates, inlet pressures and inlet temperatures."	( )	**This paper is organized as follows**
5	"_____ First, we provide detailed explanation of the methodology of our LV shape restoration algorithm. Next we describe the experiments done on the 30 simulated samples and the 20 in vivo patient-specific models to test the performance of the algorithm, followed by a discussion on the implications of the experimental results."	( )	**it has been shown that the**

[3] Todos os exemplos foram retirados do *corpus* de estudo.

YOUR TURN

1. You are going to write an introduction for a Research Article that has to do with your research project in the area of Physics. Before that, get in groups and make a list of the indispensable elements to write a good introduction. You may consult the table under the **Finding Regularities** section to help build your list.

a. _____
b. _____
c. _____
d. _____
e. _____
f. _____
g. _____
h. _____
i. _____
j. _____

2. Next, decide with the whole class which elements are going to be part of the assessment criteria of your introductions.

a. _____
b. _____
c. _____
d. _____
e. _____
f. _____
g. _____
h. _____
i. _____
j. _____

3. Write the first version of your introduction and bear in mind the following:

a. What you are writing about
b. Who you are writing to
c. How you are organizing your text
d. What language you are using
e. Where you are publishing it

4. Look at the table below. When writing the first version of your introduction, answer the questions below about your project using the Key Lexical Frames (KLFs) suggested having the examples provided as a reference.

Questions	KLF	Example	Your sentence
What is the importance of the present study?	**____(play/plays) an important role in the ____**	**Ex:** "Here we provide an experimental proof that the light intensity **plays an important role in the** vertical distribution of seven Synechococcus spp. strains isolated from the littoral zone of Lake Constance in Germany."[4]	
What other studies have shown?	**_____ it (is well known/has been shown) that the ____**	**Ex: "**These parameters vary substantially between different studies although **it has been shown that the** method used to test the materials influences the measured P binding capacity and thus the predicted performance of the filters as well as their predicted lifetime."	
Is there any gap in the present studies?	**_____ to the best of our knowledge _____**	**Ex:** "Biomechanical evolution of the simulated MTS Real cells have passive viscoelastic mechanical features, but they also move actively under the pushes of their own cytoskeleton, and **to the best of our knowledge** there is no comprehensive model of cellular biomechanics."	
What is the purpose of the study?	**The (purpose/aim) of (this /the present) (study/ paper) (is/was) to _____**	**Ex:** "**The purpose of this study is** to investigate the effects of the operating parameters on natural gas supersonic separation process, including the back pressure, inlet mass flow rates, inlet pressures and inlet temperatures."	
What is the structure of the paper?	**(The rest of the/This) paper is organized as follows _____**	**Ex:** "**This paper is organized as follows :** First, we provide detailed explanation of the methodology of our LV shape restoration algorithm. Next we describe the experiments done on the 30 simulated samples and the 20 in vivo patient-specific models to test the performance of the algorithm, followed by a discussion on the implications of the experimental results."	

5. After you write the first version of your introduction, exchange it with a classmate and use the rubric built by the group to give suggestions and recommendations to help them improve their text.

[4] Todos os exemplos foram retirados do *corpus* de estudo.