Representação dos dados estruturados do gênero textual como técnica para o processamento automático de texto

Representation of structured data of the text genre as a technique for automatic text processing

Claudia Aparecida Fonseca [*1], **Marcus Vinícius Carvalho Guelpeli** [†2] **e Rafael Santiago de Souza Netto** [‡3]

[1] *Universidade Federal dos Vales do Jequitinhonha e Mucuri, Departamento de Letras, Diamantina, MG, Brasil.*
[2] *Universidade Federal dos Vales do Jequitinhonha e Mucuri, Departamento de sistema de informação, Diamantina, MG, Brasil.*
[3] *Centro Universitário de Barra Mansa, Departamento de Ciência da Computação, Barra Mansa, RJ, Brasil.*

Resumo

O presente trabalho foi desenvolvido na área de Processamento de Linguagem Natural (PLN) e Estudos Linguísticos baseados em *corpus* compilado por ferramentas computacionais. Este trabalho parte do princípio de que é necessário assinalar uma estreita relação entre anotação e geração de *corpus* com a análise dos elementos constitutivos do gênero do texto-base. A proposta visa demonstrar, por via específica do estudo dos dados estruturados do gênero textual artigo científico, uma opção de técnica de processamento automático de texto. Para alcançar os objetivos propostos, criou-se um modelo computacional necessário para a compilação de um *corpus* linguístico, especializado, representativo do gênero Artigo Científico - *CorpACE*. O projeto teve como objeto de estudo os elementos constitutivos do gênero textual artigo científico, marcados em XML, extraídos e coletados do banco de dados da SciELO-*Scientific Electronic Library On-line*. Como produto final, obteve-se uma base de dados com as informações extraídas e estruturadas no formato XML, que delimitam e identificam as marcações do gênero em análise, disponível para várias ferramentas e aplicações. Os resultados demonstram como a representação dos elementos constitutivos do gênero pode condensar as informações disponíveis de forma hierarquizada e dinâmica, construídas durante a compilação. Ao final da pesquisa, presume-se que se fazem necessárias mais pesquisas que aproximem a Ciência da Linguagem da Ciência da Computação com ênfase em PLN na tentativa de representar e manipular os conhecimentos linguísticos em seus vários níveis morfológico, sintático, semântico e discursivo, para a melhoria na implementação e manipulação do processamento automático do texto.

Palavras-chave: Linguística de *corpus*. Processamento de linguagem natural. Artigo científico. Gênero textual. Anotação de *corpora*.

Abstract

The present article was developed in the field of Natural Language Processing (NLP) and Language Studies based on a *corpus* compiled by computational tools. This study is based on the assumption that it is helpful to trace a close relationship between *corpus* generation/annotation and the assessment of the constitutive elements of the text genre source. It aims to demonstrate, through specific studies of structured data from the text genre 'scientific article', alternatives to automatic text processing techniques. In order to reach the intended goal, the authors created a computational model for the compilation of a linguistic, specialized *corpus*, representative of the genre Scientific Article - *CorpACE*. The object of study includes the constitutive elements of scientific articles, marked in XML, extracted and collected from the SciELO-*Scientific Electronic Library On-line database*. The final product was a database obtained with information extracted and structured in XML format, which designate and identify the markups of the genre being analyzed and is available for many tools and applications. The results demonstrate how the representation of constitutive elements of the genre can condense available information with hierarchical and dynamic processes built during the compilation. At the end of the study, it is believed that more research will be required for bringing Language Science and Computer

Textoline
Linguagem e Tecnologia

DOI: 10.35699/1983-3652.2022.35445

Seção:
Artigos

Autor Correspondente:
Cláudia Aparecida Fonseca

Editor de seção:
Daniervelin Pereira
Editor de layout:
Daniervelin Pereira

Recebido em:
29 de julho de 2021
Aceito em:
18 de outubro de 2021
Publicado em:
27 de janeiro de 2022

Essa obra tem a licença "CC BY 4.0".

*Email: claudia.fonseca@ufvjm.edu.br
†Email: marcus.guelpeli@ufvjm.edu.br
‡Email: rafael.santiago@tutanota.com

Science closer with emphasis on NLP in the attempt to represent and manipulate linguistic knowledge in its many levels – morphological, syntactic, semantic and discursive – in order to improve implementation and manipulation of automatic text processing.

Keywords: *Corpus* linguistics. Natural language processing. Scientific article. Text genre. *Corpora* annotation.

1 Introdução

Com a revolução da informática e dos meios de comunicação, os documentos eletrônicos estão se tornando uma das principais mídias de leitura, divulgação acadêmica e fonte de informação para grande parte da população mundial, em função do uso generalizado da Internet. Milhares de documentos eletrônicos são produzidos e disponibilizados na Internet todos os dias. Entretanto, essa grande quantidade de informação é disponibilizada, principalmente, em formato desestruturado, ou seja, é produzida para consumo humano e, dessa forma, não necessariamente processável pelas máquinas (CAMBRIA; WHITE, 2014). Uma vez que o objetivo do armazenamento da informação é a sua posterior recuperação, a consulta sempre será formulada em função da necessidade de informação indicada pelo usuário.

A fim de utilizar plenamente esses documentos *on-line*, de forma eficaz, é imprescindível ser capaz de localizar e extrair a essência do conteúdo da informação que compõe esses arquivos. Entretanto, o grande fluxo de documentos disponibilizados, sobretudo no meio acadêmico, torna impossível a leitura de todos os textos encontrados, já que a capacidade de alcance de uma pessoa é limitada, além de ser necessário muito tempo e esforço para isso. Uma alternativa que vem sendo investigada para sanar essa limitação é o processamento automático do texto, que envolve um conjunto de técnicas computacionais para a análise e representação da linguagem humana (CAMBRIA; WHITE, 2014).

O processamento automático de textos é uma atividade proporcionada pela inovação tecnológica e pela necessidade humana de acesso à informação, o que significa que as tecnologias servem para ampliar a comunicação natural dos indivíduos, que utilizam a linguagem natural por meio dos seus códigos verbal (língua) e não verbal (gestos, imagens etc). É um estudo interessante não só do ponto de vista prático, pois ajuda o usuário a acessar, recuperar e usar, por meio de ferramentas computacionais, quantidades cada vez maiores de informações *on-line*, mas também do ponto de vista teórico científico, porque exige uma profunda compreensão da linguagem natural pelas máquinas, conhecimento esse que está associado aos processos de leitura, compreensão, apresentação, avaliação e produção de textos.

Além disso, destaca-se que, atualmente, o acesso e a recuperação da informação seguem as regras de algoritmos que dependem, principalmente, da representação textual de páginas da web. Isso significa que esses algoritmos são eficientes na execução de tarefas como recuperação de textos, fragmentando-os em *tokens*, verificação da ortografia e contagem de palavras. Entretanto, em se tratando da interpretação de sentenças e extração de informação significativa, suas habilidades ainda são muito limitadas.

Agrega-se a isso o fato de que, do ponto de vista da Linguística, o processo oferece muitos desafios, como, por exemplo, buscar soluções sobre como reformular e combinar fragmentos de texto para gerar um discurso com coesão e coerência (JONES, 2007). Ou seja, a linguagem não pode ser entendida, apenas, como a representação de regras gramaticais da língua, uma vez que ela se manifesta nas mais variadas formas, níveis e modalidades. A linguagem verbal envolve, além de semânticos e morfossintáticos, aspectos do discurso que podem transmitir informações, por exemplo, de tempo e espaço. Tudo isso gera um desafio para campos de conhecimento como da Ciência da Linguagem e Ciência da Computação com ênfase em Processamento de Linguagem Natural (PLN), no sentido de criar representações padronizadas para metadados, que precisam ser normalizadas e padronizadas, de modo que fiquem num formato legível por máquina.

Certos do grande desafio, neste trabalho é apresentado um modelo computacional para a área de PLN, com interface com a Linguística de Corpus (LC), sustentado pela compilação e anotação

de metadados[1] em corpus, utilizando conceitos da Linguística Textual (LT), que sejam capazes de filtrar os recursos estruturados do gênero textual artigo científico como técnica de PLN. A proposta de um modelo computacional juntamente com a compilação de um corpus é útil tanto para propósitos científicos, uma vez que explora a natureza da comunicação linguística, como também para propósitos práticos, uma vez que permite a interação efetiva de homem e máquina. Esse tipo de conhecimento pode, então, ser usado pelos pesquisadores para desenvolver novas estratégias pedagógicas de uso e aplicação dos recursos de linguagem estruturada.

2 O Processamento de Linguagem Natural

A linguagem natural pode ser definida, dentre outras maneiras, como uma das "faculdades cognitivas mais flexíveis e plásticas adaptáveis às mudanças comportamentais e a responsável pela disseminação das constantes transformações sociais, políticas, culturais geradas pela criatividade do ser humano" (MARCUSCHI, 2004, p. 7). Dessa forma, é por meio dela que os indivíduos se comunicam fazendo uso de um conjunto de símbolos e signos verbais e não verbais. Por esse motivo, é importante ressaltar que um fundamento básico da linguagem natural é estar sempre relacionado à situação de uso, não podendo ser dissociado de quem a usa ou de como é usada.

Neste trabalho, utiliza-se o entendimento de língua natural para a expressão linguagem natural, uma vez que as línguas são objetos mais passíveis de estudo e sistematização em relação à linguagem que envolve aspectos mais complexos como percepção, inteligência, consciência e variáveis como domínio relacionado ao convívio social, dentre outros. A língua natural é, de certo modo, o contrário de língua artificial ou de programação de computador. A linguagem artificial é totalmente sensível ao contexto, que nesse caso tem suas regras definidas por uma metalinguagem usada para expressar gramáticas livres de contexto, isto é, um modo formal de descrever linguagens formais, por meio de um conjunto de instruções e protocolos de comunicação. Além disso, por texto natural, ou seja, texto produzido por humano, entende-se que é aquele que existe e circula em algum meio social e que não foi criado com o propósito exclusivo de compor um *corpus* de estudo. Nesse contexto, a língua natural é repleta de regras, variações e ambiguidades, que dependem da análise dos fatores do contexto de produção e circulação, dentre outros aspectos, para ser processada por máquinas.

Nesse sentido, o desafio do PLN é conseguir desenvolver sistemas computacionais que sejam capazes de entender a linguagem humana de acordo com os parâmetros de uma língua formal, sistematizada com propósitos e finalidades bem definidos quanto ao seu uso. Pois numa era altamente tecnológica, como a que se vivencia, é necessário aperfeiçoar as características linguísticas dos *softwares*, para, assim, tornar mais eficiente a comunicação entre homens e máquinas.

A fim de alcançar esse desafio, o PLN precisou envolver e unir diferentes áreas para a realização de estudos e entendimento interdisciplinar que fossem capazes de abordar e incluir múltiplos conhecimentos, tais como: a Engenharia Computacional, que fornece métodos para ilustração do modelo, algoritmo, capaz de projetar e construir *hardware* e *software*; a Linguística, que se ocupa de estudar as características de linguagem humana, categorizando formas e práticas linguísticas; a Matemática, que fornece modelos e métodos formais, lança mão da lógica para a resolução de problemas e desenvolvimento de teses e hipóteses; a Psicologia, que estuda modelos e teorias que motivam o comportamento humano e seus processos mentais; a Estatística, que oferece procedimentos para prever medidas com base em registros de amostra; e a Biologia, que percorre em torno da arquitetura subjacente dos processos linguísticos no cérebro humano (MANARIS, 1998).

Apesar de serem áreas distintas do conhecimento, é por meio da união de seus saberes que os sistemas de geração de linguagem natural convertem informação de bancos de dados de computadores em linguagem compreensível ao ser humano. E sistemas de compreensão de linguagem natural convertem ocorrências de linguagem humana em representações mais formais, mais facilmente manipuláveis por programas de computador. Sendo assim, Silva (2006) alerta que os precursores dos estudos em PLN já indicavam que um computador não poderia emular a linguagem humana satisfatoriamente se não

1 O prefixo "Meta" vem do grego e significa "além de". Metadados são informações que acrescem aos dados e que têm como objetivo informar sobre eles para tornar mais fácil a sua organização e recuperação em banco de dados. Existem muitas definições do termo metadados. Literalmente, significa informação sobre dados, a descrição mais simplificada e referida é o dado de dados (KUCUK; OLGUN; SEVER, 2000).

conseguisse compreender o contexto do assunto em discussão. Seria necessário, para isso, fornecer ao programa um modelo detalhado do domínio específico do discurso em questão. Esse argumento converge com os objetivos da presente pesquisa, cujo princípio acredita na observação da anotação do gênero textual como característica relevante na implementação de sistemas de PLN. Para corroborar, Silva (2006) diz ser fundamental que:

> para emular aspectos de língua natural pressupõe equipar um sistema de PLN com vários sistemas de conhecimento e fazê-lo emular uma série de atividades cognitivas: - possuir um "modelo simples de sua própria mentalidade"; - possuir um "modelo detalhado do domínio específico do discurso"; - possuir um modelo que represente "informações morfológicas, sintáticas, semânticas, contextuais e do conhecimento de mundo físico"; - "compreender o assunto que está em discussão"; - "lembrar, discutir, executar seus planos e ações"; - participar de um diálogo, respondendo, com ações e frases, às frases digitadas pelo usuário; - solicitar esclarecimentos quando seus programas heurísticos não conseguirem compreender uma frase. (SILVA, 2006, p. 122)

Desde seu surgimento em 1950, as pesquisas em PLN têm se dedicado principalmente às tarefas de tradução automática, recuperação da informação, sumarização de textos, modelagem de tópicos e, mais recentemente, mineração de opiniões (CAMBRIA; WHITE, 2014). De um modo geral, o PLN preocupa-se diretamente com o estudo da linguagem natural voltado para a construção de *software* especializado. Sendo assim, para a sua implementação são necessários vários subsistemas complexos para representar os diversos aspectos da linguagem como: sons, palavras, sentenças e discurso nos mais variados níveis estruturais, de significado e de uso (VIEIRA; LIMA, 2001). Mesmo com muitos avanços na área, até o momento, não há disponível para uso um *software* que seja capaz de combinar todas as abordagens e de gerar uma base de conhecimento, que armazene informações que descrevam os recursos de linguagem, necessários para o desenvolvimento de aplicações de PLN, eficientemente. Isso se deve à complexidade, à riqueza de detalhes e às variações da linguagem humana, que dificilmente se adéquam à formalização do computador.

Atualmente, o estudo do PLN utiliza técnicas e abordagens linguísticas para anotação de *corpora*, basicamente, em três grandes níveis de análise: morfológica, sintática e semântica. Entretanto, essas três principais abordagens não conseguem abranger a complexidade dos recursos da linguagem que precisam ser normalizados, padronizados e transformados em linguagem de marcação.

Nesse sentido, existe uma necessidade de realização de um quarto nível de análise dos recursos discursivos da linguagem, conforme já alertava (JONES; WALKER; ROBERTSON, 2000) que, para o desenvolvimento de um sistema de PLN útil, seria preciso dar um enfoque metodológico aos fatores do contexto de produção e circulação, tanto do texto-base (texto de entrada) como do texto a ser gerado pela máquina (texto de saída).

2.1 Principais técnicas em PLN

Com base na premissa de trabalhos já realizados, quanto à abordagem, as técnicas de PLN podem ser classificadas em três classes, a saber: estatística, linguística e híbrida (BHARTI; BABU, 2017). Sobre as principais técnicas e abordagens de PLN que são direcionadas para o processamento de texto, ou seja, que podem ser usadas em anotadores de corpus, é exibido um detalhamento na Tabela 1 com destaque para quatro grandes níveis de análise linguística: morfológica, sintática, semântica e discursiva.

Tabela 1. Principais Técnicas em PLN.

Técnica	Abordagem	Descrição	Trabalhos
Remoção de **Stopwords** – (Filtragem de Palavras de Parada)	Linguística	Consiste em um processo de filtragem para remoção de palavras de pouca relevância, na tentativa de dimensionar todas as informações que não constituem conhecimento no texto. A ideia dessa filtragem é remover palavras que contêm pouca ou nenhuma informação de conteúdo, como artigos, preposições, pronomes, conjunções, advérbios, numerais e interjeições. Além disso, termos que ocorrem com alta frequência ou raramente ocorrem provavelmente não são de grande relevância e podem ser removidos.	Luhn (1958), Salton e McGill (1983), Frakes e Baeza-Yates (1992), Lui, Li e Choy (2007) e De Oliveira Júnior e Esmin (2012).
TF-IDF – (Frequência de Termo - Frequência de Documento Inverso)	Estatística	O *Term Frequency* (TF): baseia-se no pressuposto de que o peso de um termo que ocorre em um documento é diretamente proporcional à sua frequência. *Inverse Document Frequency* (IDF): baseia-se no pressuposto de que a especificidade de um termo pode ser medida por uma função inversa do número de documentos em que ocorre. Sendo assim, essa técnica consiste em ponderar a importância de cada termo dentro de um *corpus* de fundo, normalmente, constituído por documentos pertencentes ao mesmo domínio e da eliminação de uma lista de palavras muito comuns.	Luhn (1958), Jones (1972), Bhatia e Jaiswal (2015), Liu, Li e Feng (2017) e Rocha e Guelpeli (2017).
Latent Semantic Analysis (LSA) (Análise semântica latente)	Híbrida	Consiste em um método, que utiliza a sinonímia e a polissemia, para extração e representação do significado semântico de palavras em um contexto. Essa representação é obtida por meio de cálculos e aplicações matemáticas que analisam a relação entre termos e documentos, decompondo-os em vetor de índice.	Landauer, Foltz e Laham (1998) e Scarton e Aluísio (2010).
N-grams	Estatística	Essa técnica consiste na coocorrência de palavras e permite fazer uma predição estatística de dois, ou mais, termos de um texto que aparecem em uma certa sequência. Um *n-gram* é uma subsequência contígua de n itens de uma determinada sequência de texto ou fala.	Cohen (1995), Liu, Webster e Kit (2009), L. F de Alencar (2010), A. F. de Alencar (2013a) e Tonelli e Pianta (2011).
Segmentation – (Segmentação de texto em frases)	Híbrida	Consiste na segmentação do conteúdo do texto em sentenças individualizadas, representativas de um conjunto semântico mínimo para definição de uma proposição.	Lin, Hsieh e Chuang (2009), SOUSA, KEPLER e FARIA (2010) e A. F. de Alencar (2013b).
Tokenization (Segmentação de texto em palavras)	Híbrida	Consiste no processo que segmenta uma sequência de caracteres do texto em uma sequência de unidades de significado (palavras) que compõem o texto. Os espaços e pontuação são geralmente adotados como *tokens* delimitadores para idiomas ocidentais.	Webster e Kit (1992), SOUSA, KEPLER e FARIA (2010), A. F. de Alencar (2013b) e Silva, Trindade et al. (2015)

Stemming (Lematização e radicalização)	Linguística	A Lematização consiste na representação de cada palavra do texto de entrada em sua forma primitiva (*lemma*). O processo de radicalização das palavras tem como finalidade a remoção de sufixos e prefixos de um termo, para que este seja reduzido ao seu radical (*stem*).	Lovins (1968), SOUSA, KEPLER e FARIA (2010) e Rolim, Ferreira e Costa (2016).
Part-of-Speech (POS) Tagging (Etiquetagem morfossintática)	Linguística	Consiste em etiquetar as palavras do texto de entrada com suas respectivas classes gramaticais e distribuições sintáticas. Algumas das principais técnicas de etiquetagem morfossintática são: A *Baseada em regras* que faz uso de regras de etiquetagem codificadas manualmente por linguistas; A *Probabilística* que faz uso de métodos de etiquetagem estatística em que cada palavra possui um conjunto finito de etiquetas possíveis, e é rotulada com suas etiquetas mais prováveis; e, A *Híbrida* que envolve a combinação das técnicas baseadas em regras e probabilística.	Lau et al. (2008), Domingues, Favero e De Medeiros (2008), SOUSA, KEPLER e FARIA (2010), A. F. de Alencar (2013b) e Santos e Zadrozny (2014).
Etiquetagem do **Gênero Textual**	Linguística	Consiste em etiquetar as principais características do gênero do texto de entrada. Essa técnica possibilita a construção do modelo estrutural em formato arbóreo e permite acrescentar dados linguísticos; informações sobre as relações entre elementos do contexto de produção, ou sentenças ou fragmentos de sentenças da infraestrutura geral do texto; e a visualização das dimensões constitutivas do gênero base. Essa etiquetagem pode delimitar os mais variados elementos constitutivos do gênero textual como: referências bibliográficas, seções, resumo, parágrafos, tabelas, figuras, financiamento, título, subtítulos, autoria, palavras-chave, dentre muitas outras. A aplicação dessa técnica pode recuperar a estrutura básica do texto de entrada, por meio da planificação dos nós raiz e suas possíveis afiliações, que representam a infraestrutura textual. O pré-processamento de um gênero vai ser, de alguma forma, influenciado pelo reconhecimento da superestrutura e da infraestrutura de sua organização composicional.	Fonseca (2018).

Fonte: Elaborada pelos autores.

Todas essas técnicas em PLN têm aplicações variadas nas áreas de resposta automática a perguntas, recuperação de informação, sumarização ou tradução automática de textos, classificação de textos, geração de dicionários, análise de sentimentos, dentre outras (FIALHO et al., 2016). Para cada tipo de anotação, entretanto, são necessárias ferramentas de busca restritas, o que revela a necessidade de sistemas com funcionalidades mais eficientes e abrangentes, para classificação e reconhecimento do maior número de elementos textuais possíveis, inclusive aos referentes à estrutura do gênero textual. Essa possibilidade de abordagem linguística para a representação computacional que inclui a influência do contexto de produção e circulação do texto é uma técnica que pode ser explorada pela utilização dos métodos estatísticos, linguísticos ou por ambos no PLN.

2.2 O artigo científico como texto-base (ou texto de entrada)

Na esfera de circulação acadêmica, é produzida uma gama de textos com propósitos científicos, dentre eles podemos destacar vários gêneros como monografias, dissertações, teses, relatórios, etc. Um desses gêneros é o artigo científico objeto de estudo dessa pesquisa.

A saber, destaca-se que foi adotada a concepção dinâmica de gênero, defendida por Marcuschi (2002, p. 21), inspirado em Bakhtin (1997). Para os autores, um gênero textual se caracteriza como formas de enunciados, com padrões relativamente estáveis. Esses enunciados, por sua vez, têm conteúdo temático, estilo e construção composicional constituídos historicamente pelo trabalho linguístico dos sujeitos nas diferentes esferas e na diversidade da atividade humana. Constituição essa que cumpre determinadas finalidades em determinadas circunstâncias, típicas da comunicação em um dado meio social. Todo esse dinamismo confere o estatuto privilegiado para o estudo e organização dos diversos campos da ciência que utilizam os gêneros discursivos como base de suas pesquisas.

O artigo científico constitui um gênero complexo, polifônico, cujo diálogo com os outros textos precisa ser dominado pelo retextualizador[2], para que o seu aproveitamento no processamento automático seja adequado para atender a uma finalidade de uso. Segundo Bakhtin (1997), os gêneros textuais são usados e determinados por uma comunidade sócio-histórica e refletem a menor mudança na vida social de seus sujeitos. Sendo assim, os artigos científicos carregam características específicas da vida social. São gêneros marcados por um momento histórico e trazem marcas do estilo do autor que "manifesta sua individualidade, sua visão de mundo, em cada um dos elementos estilísticos" (BAKHTIN, 1997, p. 298) na criação de seu texto.

2.2.1 Conceituação e caracterização do gênero artigo científico

O gênero textual artigo científico é caracterizado, segundo Marcantonio, Santos e Lehfeld (1993, p. 71), como "resultados completos de um objeto de pesquisa. Não chegam a constituir-se em matéria para dissertações, tese ou livros. Apresentam as pesquisas realizadas e são publicados em revistas ou periódicos especializados". Para corroborar o entendimento anterior, Marconi e Lakatos (2010, p. 84) classificam o gênero como "pequenos estudos, porém completos, que tratam de uma questão verdadeiramente científica, mas que não se constituem em matéria de um livro." Para as autoras, pelo rigor científico exigido pelas pesquisas, o artigo científico deve apresentar a mesma estrutura composicional exigida pelos trabalhos científicos. Entretanto, distingue-se dos demais tipos de trabalhos científicos em função de seu reduzido tamanho e conteúdo.

O artigo científico é um gênero acadêmico que tem por finalidade apresentar resultados sucintos de uma pesquisa realizada de acordo com o método científico aceito por uma comunidade de pesquisadores. Além disso, tanto as características pertinentes a sua estrutura esquemática quanto seu estilo composicional constituem elementos interessantes para compor um modelo representativo do gênero.

Assim, trabalhos relevantes sobre escrita acadêmica, como o de Swales (1990), preocupou-se com o exame dos movimentos textuais que focalizam a esquematização do gênero. No modelo proposto pelo autor, são postulados os movimentos de texto constituídos por passos: justifica-se e fundamenta-se o trabalho empreendido, singularizam-se os movimentos gerais e específicos de Resumo, Introdução, Revisão de Literatura ou Pressupostos Teóricos, Materiais, Métodos ou Metodologia, Resultados e Discussão, Conclusão ou Considerações Finais. Para o referido autor, o que particulariza o gênero em questão é a sua composição de oito seções que explicitam os movimentos ou ações necessárias para que o artigo cumpra sua finalidade comunicativa (SWALES, 1990).

Por outro lado, estudos sobre o tema, no Brasil, como o de Marconi e Lakatos (2010), são importante base teórica para a fundamentação do estudo. Segundo as referidas autoras, a estrutura de um artigo científico deve apresentar quatro partes essenciais assim caracterizadas:

1. *Preliminares*: Cabeçalho, título (e subtítulo) do trabalho, cujo objetivo é dar conhecimento do conteúdo essencial do artigo ao leitor; Autor(es); Local de atividade: Endereço(s) que devem

2 Quem conduz o processo relacionado ao fato de um gênero surgir da necessidade da produção de um novo texto, a partir de um ou mais textos-base, e implicar a seleção das principais macroestruturas do texto-base de acordo com os propósitos da retextualização (MATENCIO, 2002)

aparecer de acordo com o veículo de publicação.

2. *Sinopse*: o objetivo do resumo é descrever o artigo com frases coerentes, concisas e fidedignas com os objetivos pretendidos e as conclusões alcançadas no trabalho. Como por exemplo: O que o autor fez? Como o autor fez? O que o autor encontrou? O que o autor concluiu?
3. 3 *Corpo do artigo*: deve descrever todo o processo da pesquisa e, geralmente, subdividido em a) Introdução: apresentação do assunto, objetivo, metodologia, limitações e proposições; b) Texto: exposição, explicação e demonstração do material; c) avaliação dos resultados e comparação com as obras anteriores; e, d) Comentários e conclusões: inferência, baseada e fundamentada no texto, de forma resumida.
4. *Parte referencial*: Bibliografia; Apêndices ou anexos (se houver); Data.

Dessa forma, o artigo científico por ser um texto que circula na esfera acadêmica, por guardar muitas características formais em sua estrutura discursiva e esquemática, permite pouca variação em suas configurações, podendo servir como uma base textual interessante para ser investigada.

2.2.2 Planificação do gênero artigo científico

Seguindo um modelo de análise textual do artigo científico, baseando-se no quadro teórico do gênero proposto de Bronckart (1999), podem-se pressupor dois níveis organizacionais, compostos pela infraestrutura textual e pelos mecanismos textuais. Segundo o autor, o indivíduo deve mobilizar algumas de suas representações sobre o mundo para produzir um texto. Dentre elas, duas influenciam diretamente sobre a forma como o texto é organizado. São elas: o contexto de produção e o conteúdo temático.

Em seguida, é apresentada a planificação para análise do primeiro nível organizacional do artigo científico, conforme demonstração do contexto de produção na Figura 1:

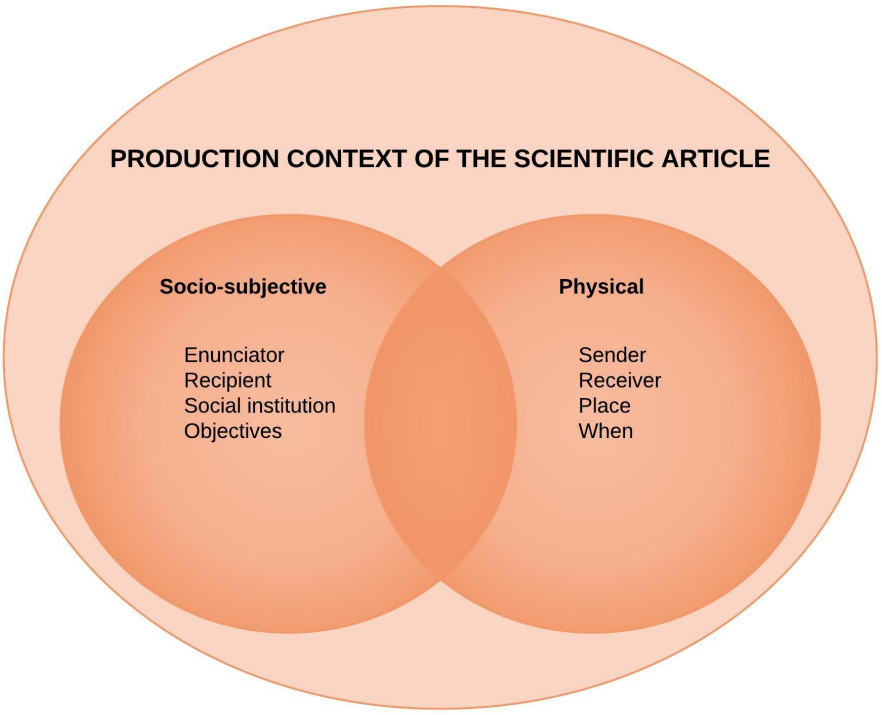

Figura 1. Planificação do contexto de produção do artigo científico.
Fonte: Adaptado do modelo de análise de textos de Bronckart (1999).

O primeiro nível da infraestrutura textual ou contexto de produção é o mais profundo e é constituído pelo plano geral que se refere à organização do conjunto de conteúdo temático. Nesse contexto, Bronckart (1999, p. 97/98) considera que "trata-se de conhecimentos que variam em função da experiência e do nível de desenvolvimento do agente e que estão estocados e organizados em sua

memória, previamente, antes do desencadear da ação de linguagem". Sendo assim, o conteúdo temático de um texto são as representações construídas pelo agente-produtor do texto, baseadas tanto no mundo físico e real quanto no mundo sociossubjetivo desse. Ademais, deve-se considerar o lugar e o papel social ocupado pelo produtor, assim como o momento histórico da produção e a posição social do receptor. Por exemplo, o plano geral de um artigo científico, geralmente, é constituído por: introdução, pressupostos teóricos, metodologia, resultados das análises e conclusões, sendo que cada uma dessas seções tem seus conteúdos e objetivos específicos.

Entre suas principais características estão a estrutura composicional e o estilo da textualização, visto que seu público é um grupo de especialistas na área, conhecedor do assunto, métodos e interesse na pesquisa divulgada. Assim, essas características são decisivas na construção do texto, ou seja, na superestrutura do artigo científico, assim como a seleção lexical das estruturas sintáticas, isto é, na composição do estilo.

Em seguida, na Figura 2, além da apresentação da composição da infraestrutura geral do artigo, é apresentado um modelo de planificação para análise do segundo nível organizacional do artigo científico referente à composição do estilo, ou seja, as capacidades linguístico-discursiva:

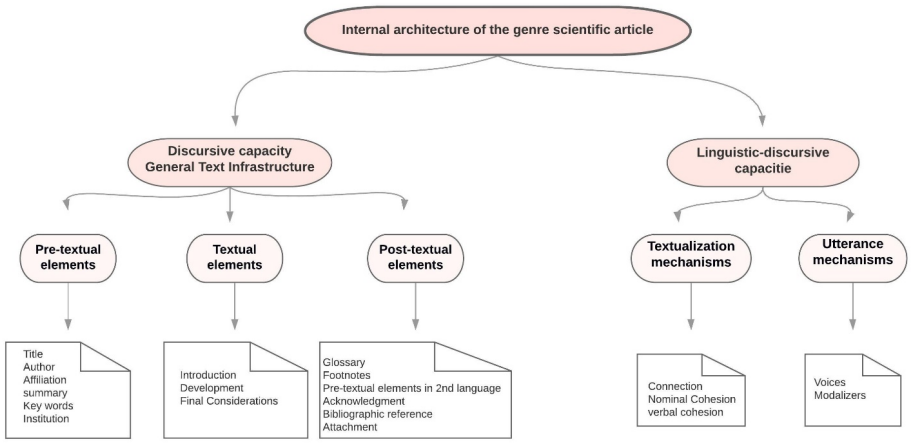

Figura 2. Planificação da Superestrutura do artigo científico.
Fonte: Adaptado do modelo de análise de textos de Bronckart (1999).

A planificação da superestrutura do artigo científico representa o segundo nível organizacional e se refere aos mecanismos textuais, que, por sua vez, dividem-se em *mecanismos de textualização* e*mecanismos enunciativos*. A função desses mecanismos é contribuir para dar ao texto sua coerência linear ou temática, tornando explícitas as grandes articulações hierárquicas, lógicas, temporais e espaciais.

Os mecanismos de textualização garantem a conexão, a coesão nominal e a coesão verbal. Já os mecanismos enunciativos garantem a progressão do "conteúdo temático", dizem respeito às vozes assumidas no interior do texto e as modalizações avaliativas que compõem o enunciado. Bronckart (1999) afirma que:

> qualquer que seja a diversidade e a heterogeneidade dos componentes da infra-estrutura de um texto empírico, ele constitui um todo coerente, uma unidade comunicativa articulada a uma situação de ação e destinada a ser compreendida e interpretada como tal por seus destinatários. Essa coerência geral procede (...) dos mecanismos de textualização e (...) dos mecanismos enunciativos (BRONCKART, 1999, p. 259).

Segundo Silva, Pereira e Bueno (2014), muitos estudiosos, como Schneuwly e Dolz (2004), retomam o modelo de Bronckart (1999), para fins didáticos, ressaltando que, antes de o gênero ser levado para a sala de aula, é necessário que seja feito o seu modelo didático. Entretanto, as autoras salientam que existe uma diversidade de mecanismos que norteiam a elaboração desse modelo, sintetizados em

três princípios de legitimidade, pertinência e solidarização. Para as autoras, esses princípios relacionados às capacidades de reconhecimento da infraestrutura textual e dos mecanismos de textualização podem servir de subsídios para a criação de um modelo didático do gênero artigo científico a fim de subsidiar professores e alunos de pós-graduação *stricto sensu* a desenvolverem atividades para se apropriarem desse gênero.

Assim como na pós-graduação, esses princípios relacionados às capacidades de reconhecimento da infraestrutura textual e dos mecanismos de textualização podem servir de subsídios para a criação de um modelo representativo do gênero artigo científico a fim de subsidiar as aplicações no processamento automático de texto por meio da criação de marcações que recuperam as características próprias desse gênero.

2.2.3 Planificação da representação em linguagem de marcação do gênero artigo científico

É demonstrado na Figura 3 o modelo estrutural em forma de árvore com marcações em XML dos principais elementos constitutivos da representação da superestrutura do artigo científico. Esses elementos podem influenciar no armazenamento, na organização e na recuperação da informação no processamento automático do gênero:

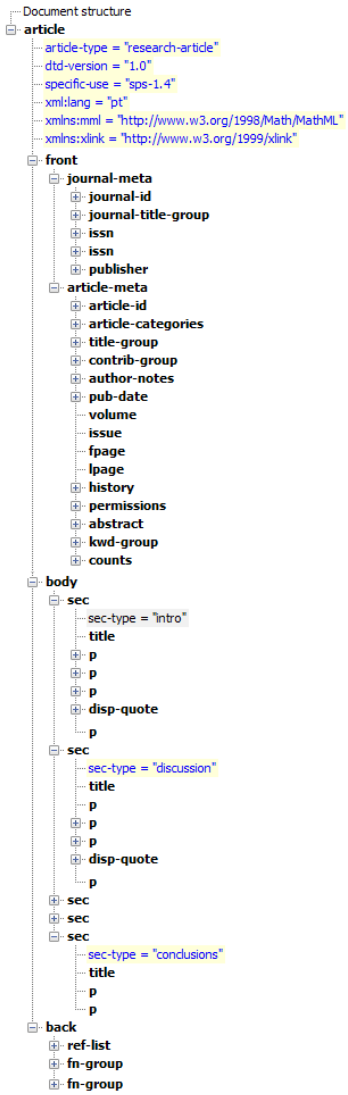

Figura 3. Modelo estrutural em forma de árvore dos principais elementos constitutivos da representação da superestrutura do artigo científico.
Fonte: Adaptado pela autora do banco de dados da SciELO.

Por meio da Figura 3, pode-se observar que a estrutura básica do gênero artigo científico é repre-

sentada pelo nó raiz (<article>) e este possui como filiação mais três nós, representativos de elementos pré-textuais (<front>), elementos textuais (<body>) e elementos pós-textuais (<back>). Os elementos pré-textuais do <front> são caracterizados e descrevem os metadados do periódico, título, autoria, afiliação, resumo, palavras-chave, DOI, volume, número, suplemento, paginação, indicação da licença Creative Commons, data de publicação, seção de cabeçalho, histórico de datas, dados de correspondência, notas de autor, informações de resenhas de livros, contagem de elementos e dados de financiamento (se houver). Os elementos textuais do <body> são caracterizados e descrevem o corpo textual do artigo que pode ser constituído por seções. Cada uma delas possui um elemento <title> seguido de um ou mais parágrafos <p>. Seções de primeiro nível podem ser qualificadas de acordo com seu tipo por meio do atributo @sec-type, cujos valores possíveis, segundo o Guia de uso de elementos e atributos XML para documentos que seguem a implementação GUIA... (2016), são os constantes da Tabela 2.

Tabela 2. Tipo de seções.

Valor	**Descrição**
Cases	Relatos/estudos de caso
Conclusions	Conclusões/considerações finais/Final Remarkes
Discussion	Discussões
Introduction	Introdução/sinopse
Materials	Materiais
Methods	Metodologia/método
Results	Resultados

Fonte: Guia de uso de elementos e atributos XML para documentos que seguem a implementação (GUIA..., 2016).

Os elementos pós-textuais do <back> são caracterizados e descrevem a parte final do documento que compreende lista de referências e demais dados referentes à pesquisa como: notas de rodapé, agradecimentos, apêndice, material suplementar, anexos e glossário.

A representação do modelo didático do artigo científico, em formato arbóreo, permite acrescentar aos dados linguísticos brutos informações sobre as características e a visualização das dimensões constitutivas do gênero, delimitar os objetivos a serem atingidos em relação aos diferentes propósitos de uso do texto no processamento automático. O enriquecimento do corpus com informações da composição do gênero, na forma de representações arbóreas em que se indicam as relações entre elementos do contexto de produção (constantes dos nós <front> <article-meta> </article-meta> </front>), ou sentenças ou fragmentos de sentenças da infraestrutura geral do texto que podem aplicar qualquer formalismo estrutural imposto pelo XML de origem (compreendidos em <body> <sec> </sec> </body> e <back> <ref-list> </ref-list> </back>) atribuído ao gênero. Essa configuração em que consiste na transformação do texto puro em texto estruturado, indica as principais características do gênero que podem ser usadas em aplicações da linguística computacional, ou seja, no pré-processamento automático para resolver um dado problema ou auxiliar em uma dada atividade de localização, armazenamento e/ou recuperação da informação.

Acredita-se que, a partir da estrutura representativa de determinado gênero textual, é possível organizar e preparar a recepção e interpretação de determinados conteúdos para o processamento automático do texto. O pré-processamento de um gênero vai ser, de alguma forma, influenciado pelo reconhecimento da superestrutura e da infraestrutura de sua organização composicional.

Sendo assim, a construção e disponibilidade de bancos de árvores que representem os recursos de linguagem constitutivos do gênero textual é fundamental para fomentar as variadas aplicações em PLN. Essas informações estruturadas são importantes, pois sinalizam o contexto de produção e alguns organizadores textuais centrais que podem se revelar boas estratégias para o pré-processamento. Além disso, constituem bases de dados sobre os quais se pode efetuar análises qualitativas e quantitativas por variadas técnicas, que podem complementar outras abordagens.

2.3 Trabalhos correlatos

Alguns anotadores textuais disponíveis para análise fazem uso de abordagens complexas para a categorização de seus recursos de dados e exploração de correspondência para seus padrões. Dentre elas, as mais utilizadas são: a abordagem morfossintática que especifica o modo como os grupos de elementos devem se organizar; e, a abordagem semântica que especifica o que o grupo de elementos deve significar. A seguir, serão apresentadas as ferramentas computacionais eDictor, Aelius e COMEDI para anotação e processamento de texto, que utilizam abordagens morfossintática e semântica.

2.3.1 eDictor

O eDictor[3] é uma ferramenta utilizada para auxiliar a edição eletrônica em XML de textos antigos para fins de análise filológica e a codificação linguística automática (SOUSA, 2014). Essa ferramenta foi idealizada para a criação do *Corpus* Anotado do Português Tycho Brahe[4] (CTB), cujas principais funcionalidades são: I) flexibilidade dos formatos gerados, permitindo tanto a leitura humana como a leitura automática; II) garantia da qualidade filológica da edição por se tratar de um editor especializado; III) possibilidade de operar com vários níveis de edição: IV) Junção, Segmentação, Grafia, Modernização, Expansão, Correção, Pontuação; V) possibilidade de criar novos níveis de edição de acordo com a necessidade do pesquisador.

Dentre seus vários níveis de edição, destacam-se suas principais funcionalidades que são: junção e segmentação. A primeira é utilizada para unir trechos do texto como palavras quebradas, enquanto a segunda faz o oposto, separa trechos indevidamente unidos. Tais edições são feitas com anotação XML de forma a manter o texto original disponível para consulta. Essas edições nos textos podem incluir outros níveis de trabalho como modernização, expansão, grafia e pontuação (SOUSA, 2014; SOUSA; KEPLER; FARIA, 2010, 2016).

2.3.2 Aelius

O etiquetador Aelius[5] é um *software* livre em desenvolvimento, para análise superficial do Português Brasileiro, que faz parte do projeto *Aelius Brazilian Portuguese POSTagger*[6] e está registrado no SourceForge.net[7] (ALENCAR, L. F. d., 2010; ALENCAR, A. F. d., 2013a,b). Por possuir uma arquitetura híbrida, recorre às abordagens baseadas em regras formuladas manualmente e ao sistema estatístico estocástico baseado em n-gramas. O Aelius foi projetado para etiquetar morfologicamente textos escritos de maneira automática. Para tanto, esse editor desempenha as seguintes tarefas: I) pré-processamento de corpora; II) construção de *language models* e etiquetadores com base num corpus anotado; III) avaliação do desempenho de um etiquetador; IV) comparação entre diferentes anotações de um texto; V) realização de anotação de *corpora* e auxílio na revisão humana de anotação automática.

Essa ferramenta também inclui recursos de idioma, como modelos de linguagem, textos de amostra e padrões de ouro[8]. Atualmente, a Aelius já oferece recursos para *corpora* e corporação de POS e gera anotações em diferentes formatos, como em XML no esquema de codificação TEI[9] P5.

3 Disponível em: https://humanidadesdigitais.org/edictor/. Acesso em: 19 de abr. 2018.
4 Disponível em: http://www.tycho.iel.unicamp.br/corpus/. Acesso em: 19 de abr. 2018.
5 Disponível em: http://aelius.sourceforge.net/. Acesso em 19 de abr. 2018.
6 Disponível em: Aelius Brazilian Portuguese Pos-tagger - http://sourceforge.net/projects/aelius/files/. Acesso em 19 de abr. 2018.
7 Disponível em: Sourcefoge.Net - Maior hospedagem mundial de software de código aberto. http://sourceforge.net/. Acesso em 19 de abr. 2018.
8 Em PLN "padrão de ouro" significa que as avaliações mais fortes e significativas são baseadas em resultados do mundo real, em que um sistema é implantado operacionalmente e é medido seu impacto nos resultados gerados por usuários do mundo real (REITER, 2018).
9 Text Encoding and Interchange (TEI) define um conjunto vasto de elementos e atributos na linguagem XML que permitem representar características estruturais, conceituais e de visualização dos textos.

2.3.3 COMEDI

O *COmponent Metadata EDItor* (COMEDI)[10] é um editor de componentes para metadados baseado na Web, em conformidade com qualquer perfil CMDI[11], e que oferece suporte atualizado para recursos adotado pela CLARIN[12]. No COMEDI, é possível criar um registro de metadados a partir do zero, ou carregar, editar e baixar qualquer arquivo XML CMDI. Seus componentes poderem ser usados independentemente do editor, ou podem ser usados por meio de uma interface web, em que o usuário seleciona um perfil CMDI para iniciar, e o editor exibe o perfil como um formulário *on-line* simples que oculta o código XML. Além disso, o editor pode funcionar como um servidor completo para armazenar, pesquisar, visualizar e gerenciar, por meio da administração de grupos e de usuários, o controle sobre o direito de acesso aos metadados individuais. O gerenciamento de usuários do editor é feito pela autenticação do *login* e opera em dois níveis: o de usuário e o de grupo (LYSE; MEURER; DE SMEDT, 2015).

A Um perfil CMDI consiste em *componentes* e *elementos*. Os elementos são os nós terminais, nos quais são atribuídos um valor, digitado em um campo de texto do formulário, ou selecionado em um menu suspenso. Os elementos que pertencem aos terminais geralmente são agrupados em componentes, por exemplo, um componente pessoa (com elementos como sobrenome, nome) ou um componente licença (com elementos como nome da licença, URL da licença). Em resumo, os perfis recomendados, e que são encontrados no menu suspenso COMEDI, são: I) *corpusProfile*: descreve *corpora* de todos os tipos e modalidades; II) *lexicalProfile*: descreve recursos lexicais; III) o usuário é livre para desenvolver o seu próprio perfil, mas nesse caso, deve reutilizar os componentes existentes, na medida do possível. Essas recomendações têm como objetivo auxiliar na descrição dos recursos linguísticos com metadados de acordo com a estrutura do CMDI. Para isso, o editor precisa distinguir campos obrigatórios de campos opcionais e impor que as entradas obrigatórias sejam preenchidas (DIMA et al., 2012).

As ferramentas computacionais eDictor, Aelius e COMEDI são utilizadas para anotação e processamento de texto, baseadas na utilização de abordagens morfossintática e semântica. Entretanto, não foi observada, nos anotadores analisados, a utilização do gênero textual como recurso de dado a ser explorado para anotação e processamento de texto, revelando a necessidade de uma abordagem discursiva.

Em seguida, será apresentado um modelo computacional que, por meio da abordagem discursiva, filtra e categoriza os recursos de dados contextuais do artigo científico, referente à sua construção composicional, demonstrando a relevância desse tipo de anotação para a compreensão e categorização de padrões do contexto de produção desse gênero. Segundo Cambria e White (2014), trabalhos mais recentes reconhecem a necessidade de categorizar os recursos de linguagem que expressam o conhecimento externo do texto na interpretação e resposta à entrada de linguagem. Esse conhecimento está relacionado à abordagem *pragmática* que especifica como as informações contextuais podem ser aproveitadas para fornecer melhor correlação entre as variadas abordagens (morfossintática e semântica) já existentes e combiná-las entre si.

3 Modelo computacional – AnoTex

O AnoTex[13] é um anotador textual em desenvolvimento, cuja principal funcionalidade é extrair as características constitutivas do gênero artigo científico por meio das marcações em XML enriquecidas com informações linguísticas sobre sua construção composicional (FONSECA et al., 2018). Sua principal atribuição é processar os elementos do contexto de produção e da infraestrutura geral do

10 Disponível em: http://clarino.uib.no/comedi/page?page-id=repository-main-page. Acesso em 19 de abr. 2018.

11 Component MetaData Infrastructure (CMDI), fornece uma estrutura para descrever e reutilizar um conjunto de metadados. Disponível em: https://www.clarin.eu/content/component-metadata. Acesso em 19 de abr. 2018.

12 Common Language Resources and Technology Infrastructure (CLARIN), é uma infraestrutura de pesquisa que foi iniciada a partir da visão de que todos os recursos e ferramentas de linguagem digital de toda a Europa e além são acessíveis através de um ambiente *on-line* de *logon único* para o apoio de pesquisadores nas ciências humanas e sociais. Disponível em: https://www.clarin.eu/. Acesso em 19 de abr. 2018.

13 AnoTex - para a construção automática de *corpus* partir do gênero do texto-base. Ferramenta computacional versão 0.1 beta, linha de comando, escrita em ANSI C, desenvolvida para funcionar em qualquer sistema que disponha de uma libc, contém bibliotecas para extração de informações disponíveis em formato XML e PDF, além de serviços disponíveis para consulta e ampliação (FONSECA, 2018).

texto, para a compilação do corpus de estudo. Essa ferramenta é capaz de filtrar e exportar os elementos anotados em XML e do seu PDF relacionado, extraídos de banco de dados disponíveis na web, como por exemplo da Biblioteca Eletrônica SciELO-*Scientific Electronic Library On-line*[14].

O conjunto de aplicações desempenhado pelo AnoTex está demonstrado na Figura 4 que, em síntese, corresponde ao delineamento do modelo computacional adequado aos propósitos da pesquisa, dividido em quatro etapas denominadas: 1 seleção, 2 compilação, 3 processamento e 4 exportação dos dados.

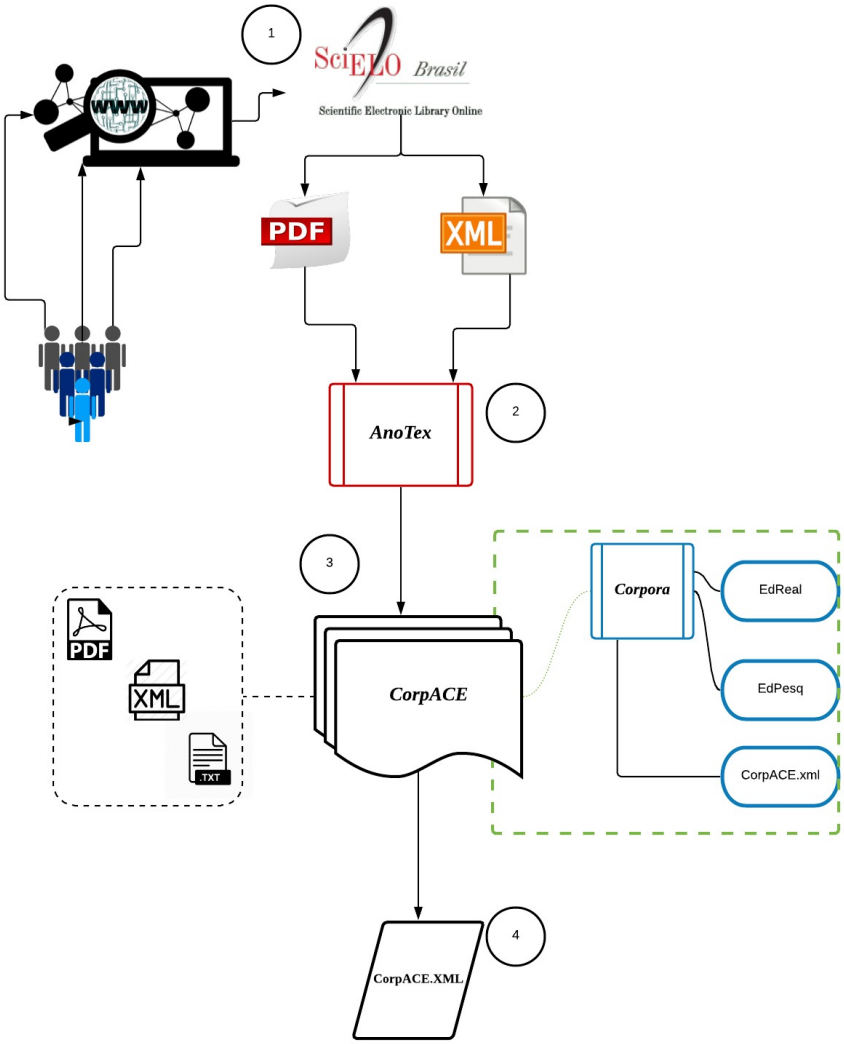

Figura 4. Modelo de compilação do AnoTex.
Fonte: (FONSECA et al., 2018).

As três primeiras etapas ilustradas no modelo dizem respeito à qualidade dos dados, sendo elas:

1. *Seleção*: Escolha do conjunto de requisitos do gênero artigo científico, enfatizado pelo uso de etiquetas XML, necessário para impactar na validade e confiabilidade do corpus de análise, adequando-o aos propósitos da investigação, isso é, que permita destacar as características e a visualização das dimensões constitutivas do gênero;
2. *Compilação*: Captura do conjunto ou amostra de dados constitutivos do gênero textual a serem manipulados. Em outras palavras, é nessa etapa que são organizados e gerados os corpora compostos de corpus com características específicas para a análise desejada;
3. *Processamento*: Transformação e/ou tratamento dos dados, tais como ruídos, inconsistências, dados inexistentes ou incompletos, que podem gerar padrões distorcidos, eliminação de atributos redundantes, padronização do conjunto de valores dos elementos selecionados dos artigos. Atual-

14 Disponível em: http://www.scielo.br. Acesso em 12 jan 2018.

mente, o AnoTex filtra 06 elementos pré-textuais, compreendidos no <front> com categorização e marcação com informações sobre: Nome da revista: <journal-title> </journal-title>; Título do artigo: <article-title> </article-title>; Nome do autor: <contrib contrib-type="author"> <name> <surname> </surname> <given-names> </given-names> </name>; Instituição: <institution content-type=""> </institution>; Resumo:<abstract> <title></title> </abstract>; Palavras-Chave: <kwd-group xml:lang=""> <title> </title> <kwd></kwd> </kwd-group>. Dentre os elementos textuais, compreendidos no <body>, são filtradas as seções e subseções com categorização de marcações com informações sobre: Introdução: <sec sec-type="intro"> <title></title> </sec>; Desenvolvimento: <sec sec-type="discussion"> <title></title> </sec>; Considerações finais: <sec sec-type="conclusions"> <title></title> </sec>. Além disso, o AnoTex filtra as referências bibliográficas dos elementos pós-textuais, compreendidos no <back> com categorização e marcação com informações sobre: Referências bibliográficas: <ref-list> <title> </title> <ref> </ref> </ref-list>;

4. *Exportação*: Saída do corpus em um arquivo XML, para interpretação e avaliação dos padrões identificados em função dos objetivos iniciais. Estas etapas convergem para a formação de uma base de dados concisa, em que dados de diferentes periódicos sejam coletados e armazenados em um único conjunto de dados.

A Figura 5, a seguir, apresenta uma amostra da configuração do *CorpACE* gerada pelo AnoTex. Os elementos pré-textuais demonstrados em 1, 2, 3, 4 5, 8 e 10 correspondem a representação do contexto de produção; os elementos textuais demonstrados em 6 e 9 correspondem a representação da arquitetura geral do texto, assim como os elementos pós-textuais demonstrados em 7.

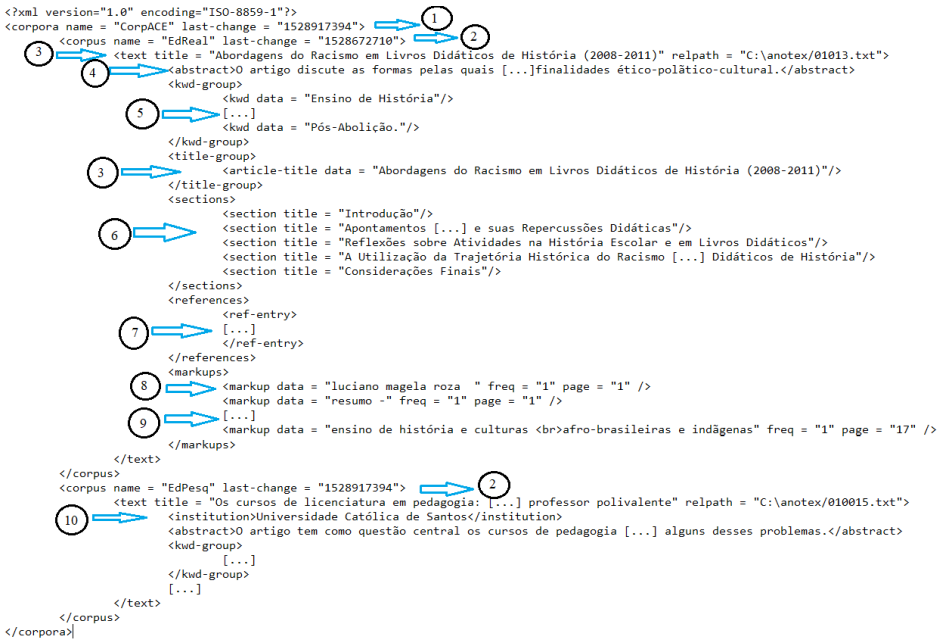

Figura 5. Amostra da configuração do *CorpACE*.
Fonte: Adaptado pela autora do AnoTex v0.1b.

Por meio da Figura 5, pode-se observar que após a filtragem de cada elemento, o AnoTex cria o *CorpACE* gerando um arquivo de saída XML (baseado na entrada dos dados do XML e do PDF). Essa base textual servirá para análise linguística, treinamento e análise de ferramentas computacionais cuja finalidade seja o processamento de texto. A separação entre a estrutura e apresentação dos dados dão maior maleabilidade à informação (a partir de um mesmo XML é possível apresentar a informação de formas distintas além de carregar metadados em uma base de dados). Os dados representados pelo XML são estruturados de forma arbórea, e cada tag ou marca representa um nó ou elemento na árvore.

Para a saída dos corpora em um arquivo XML, o AnoTex precisou receber cinco argumentos obrigatórios: o arquivo em PDF, o arquivo XML, a denominação dos corpora, a denominação do

arquivo de saída e em qual corpus, desse arquivo de saída, a marcação foi adicionada. Uma vantagem, desse processo, é que um arquivo XML de saída pode reunir diversos corpora, requisito muito útil se, por exemplo, for necessária a criação de diferentes versões do mesmo corpus. O AnoTex permite a criação de vários corpora, adicionados dentro do mesmo arquivo, esse conjunto de requisito constitui os corpora do arquivo. Todas essas funcionalidades puderam ser visualizadas nas Figura 4 e Figura 5.

4 Resultados e discussão

Para obter os resultados, foram utilizados 87 artigos de Qualis A1[15], publicados pelas revistas "Educação & Realidade"[16] e "Educação e Pesquisa"[17], no ano de 2017. Os artigos foram coletados do *site* SciELO e gravados em arquivos individuais no formato PDF, XML e TXT, formatos que são suportados pelo AnoTex. A compilação desses arquivos gerou o *CorpACE* que é caracterizado como *corpora* especializados, uma vez que são compostos por textos de uma única área de especialidade: Ensino e representativo de um único gênero textual, o artigo científico. Conforme esquematização de Sardinha (2004), quanto ao tamanho, os *corpora* se enquadram como de médio porte. Em seguida, apresentamos a Tabela 3 com os principais dados numéricos que caracterizam o *CorpACE*:

Tabela 3. Resultados estatísticos do *CorpACE*.

Estatísticas	**Total**
Número de artigos	87
Total de palavras	750380
Número de palavras do menor artigo	6144
Número de palavras do maior artigo	18758
Número médio de palavras por artigo	8625
Número de palavras filtradas dos elementos constitutivos do gênero	130128
Número de palavras do *Corpus* EdReal	383506
Número de palavras do *Corpus* EdPesq	366874

Fonte: Elaborada pelos autores.

O *CorpACE* foi submetido a um tratamento computacional possibilitado pela LC, que se faz presente metodologicamente, nesta pesquisa, por meio da ferramenta AnoTex que gerou os dados da Tabela 3. Foram utilizadas duas bibliotecas principais da ferramenta para a filtragem dos dados em XML e do PDF dos arquivos dos artigos. A primeira foi utilizada para organizar os *corpora* em bancos de árvore com informações do contexto de produção e da infraestrutura geral do texto. A segunda foi utilizada para obter dados das expressões em negrito e itálico com informações estatísticas de localização e frequência nos textos. A ferramenta ajudou tanto na organização dos dados como na análise dos elementos mais frequentes utilizados nos artigos científicos.

O *CorpACE* é constituído por três *corpora*, que possuem um total de 750380 palavras, distribuídas em 87 artigos científicos. Esses *corpora* estão assim subdivididos: *corpus* EdReal (Revista Educação e Realidade), com 383505 palavras distribuídas em 45 artigos científicos; *corpus* EdPesq (Revista Educação e Pesquisa) com 366874 palavras distribuídas em 42 artigos científicos; e *corpus* CorpACE.xml com 130128 palavras dos elementos filtrados (EF) dos 87 artigos.

O primeiro passo para a análise dos dados levantados é, no arquivo XML gerado na saída do AnoTex, observar e realizar a organização e sistematização do resultado das marcações dos elementos constitutivos do gênero que foram selecionados e especificados na etapa 3 de processamento.

15 O Qualis constitui-se um sistema brasileiro de avaliação de periódicos, mantido pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior). Relaciona e classifica os veículos utilizados para a divulgação da produção intelectual dos programas de pós-graduação "*stricto sensu*" (mestrado e doutorado), quanto ao âmbito da circulação (local, nacional ou internacional) e à qualidade (A, B, C), por área de avaliação. Os estratos estão divididos em 8 níveis, em ordem de qualidade.

16 Disponível em: http://www.seer.ufrgs.br/index.php/educacaoerealidade/index. Acesso em 12 jan 2018.

17 Disponível em: http://www.educacaoepesquisa.fe.usp.br/. Acesso em 12 jan 2018.

Em seguida, feitas as delimitações de cada *corpus* apresentamos a contagem e a sistematização dos *corpora* EdReal e EdPesq, por elemento filtrado, separadamente na Figura 6.

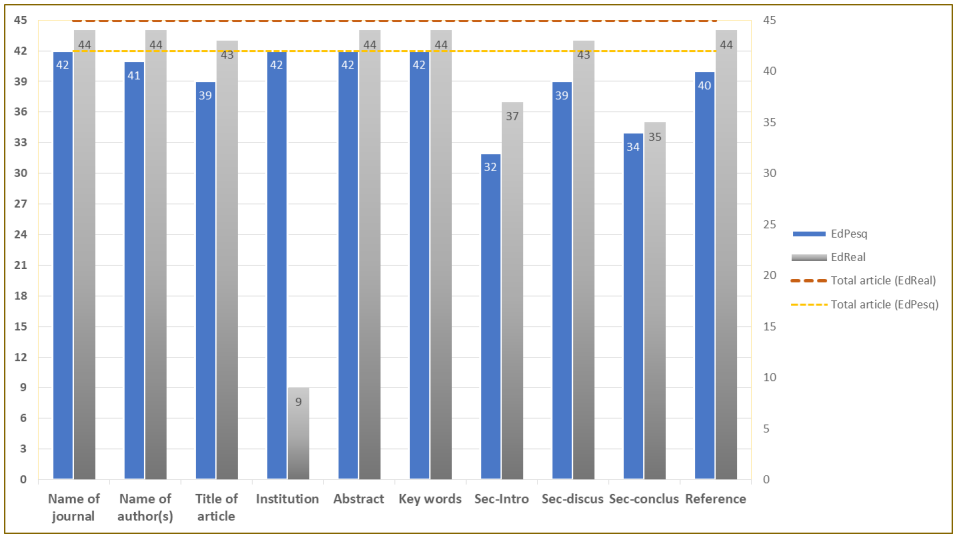

Figura 6. Elementos filtrados por *corpus*.
Fonte: Elaborado pela autora.

Na Figura 6, são mostradas as quantidades de elementos que foram filtradas dos artigos científicos de cada *corpus*. Os resultados apontam que a ferramenta AnoTex conseguiu filtrar os elementos constitutivos do gênero dos 42 artigos selecionados, da Revista EdPesq, nas seguintes proporções:

- *Elementos do <fornt>*: Nome da Revista 100%; Nome Autor(es): 97,6%; Título do Artigo: 92,9%; Instituição: 100%; Resumo: 100%; Palavras-Chave: 10%;
- *Elementos do <body>*: Seção Introdução: 76,2%; Seção Discussão: 92,9%; Seção Conclusão: 81,0%;
- *Elementos do <back>*: Referência: 95,2%. A filtragem dos elementos constitutivos do gênero textual, dos 45 artigos selecionados da Revista EdReal, gerou os seguintes resultados:
- *Elementos do <fornt>*: Nome da Revista 100%; Nome Autor(es): 100; Título do Artigo: 97,8%; Instituição: 20,0%; Resumo: 97,8%; Palavras-Chave: 97,8%;
- *Elementos do <body>*: Seção Introdução: 82,2%; Seção Discussão: 95,6%; Seção Conclusão: 77,8%;
- *Elementos do <back>*: Referência: 97,8%.

O resultado da filtragem do elemento Instituição <institution content-type=""> </institution> na Revista EdReal chamou a atenção em função da discrepância em relação à filtragem do mesmo elemento na Revista EdPesq. Essa alteração se deu em função de um problema na marcação da *tag* <aff id="aff1">, por parte da revista, em que foi deixado um "I" solto no meio do código, confundindo o *parsing* do AnoTex. Para resolver esse problema, o "I" foi envolvido com a marcação <label> I</label>, igualmente marcado nos outros arquivos que funcionaram, o que solucionou o problema e possibilitou a filtragem do dado (Figura 7 e Figura 8).

```
    <xref rid="aff1" ref-type="aff">I</xref>
</contrib>
- <aff id="aff1">
    I
    <institution content-type="original">Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre/RS - Brasil</institution>
    <institution content-type="normalized">Universidade Federal do Rio Grande do Sul</institution>
    <institution content-type="orgname">Universidade Federal do Rio Grande do Sul</institution>
  - <addr-line>
      <named-content content-type="city">Porto Alegre</named-content>
      <named-content content-type="state">RS</named-content>
    </addr-line>
    <country country="BR">Brazil</country>
</aff>
```

Figura 7. Representação do código com problema na marcação.
Fonte: Adaptado pela autora do *CorpACE*.

Embora tenha ocorrido essa discrepância na filtragem do elemento <institution>, em função de

```
<aff id="aff1">
    <label>I</label>
    <institution content-type="original">Universidade Federal do Rio Grande do Sul (UFRGS),
    <institution content-type="normalized">Universidade Federal do Rio Grande do Sul</institut
    <institution content-type="orgname">Universidade Federal do Rio Grande do Sul</institutio
    <addr-line>
        <named-content content-type="city">Porto Alegre</named-content>
        <named-content content-type="state">RS</named-content>
    </addr-line>
    <country country="BR">Brazil</country>
</aff>
```

Figura 8. Representação do código sem problema e após correção.
Fonte: Adaptado pela autora do *CorpACE*.

um problema na produção do XML, o resultado foi considerado, de um modo geral, satisfatório conforme demonstrado na Figura 9.

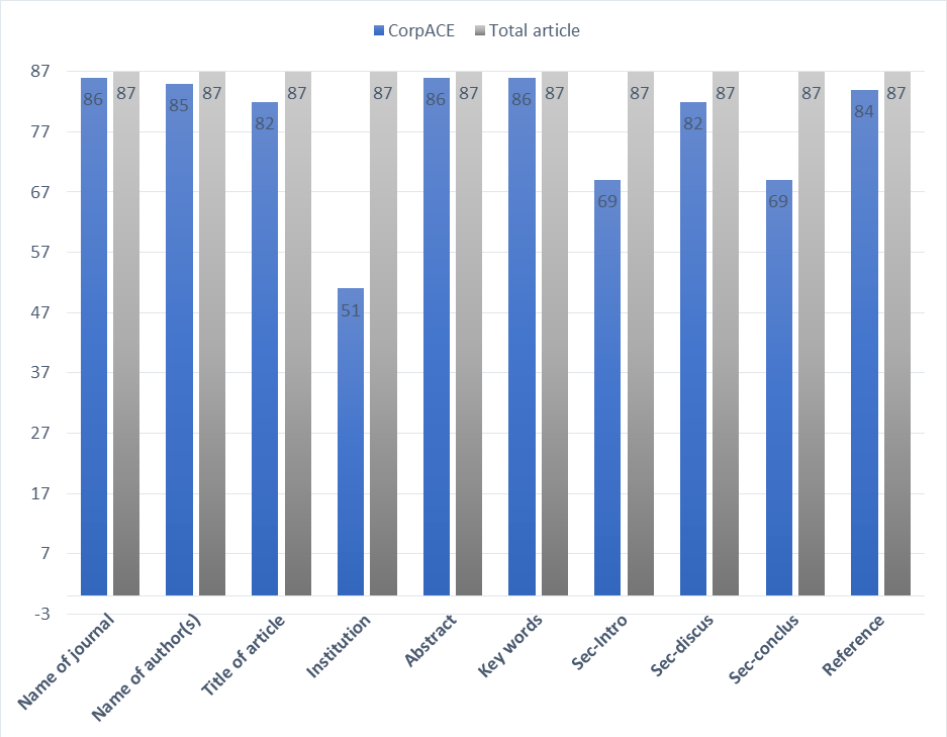

Figura 9. Total de elementos filtrados para o *CorpACE*.
Fonte: Elaborado pela autora.

Na Figura 9, são mostradas as quantidades de elementos que foram filtradas dos artigos científicos no total. Os resultados apontam que a ferramenta AnoTex conseguiu filtrar, pelo menos 4 e no máximo os 10 elementos representativos do gênero em todos os 87 artigos selecionados, nas seguintes proporções:

- *Elementos do <fornt>*: Nome da Revista: 100%; Nome Autor(es): 98,9%; Título do Artigo: 95,4%; Instituição: 58,6%; Resumo: 98,9%; Palavras-Chave: 98,9%;
- *Elementos do <body>*: Seção Introdução: 79,3%; Seção Discussão: 94,3%; Seção Conclusão: 79,3%;
- *Elementos do <back>*: Referência: 96,6%.

Ainda quanto à análise das distorções, outro elemento que chamou a atenção foi a marcação do título do artigo <article-title> </article-title>, embora todas as marcações com essa denominação tenham sido filtradas, foi considerado para o resultado apenas as que trouxeram a *tag* e o conteúdo do texto (o título propriamente dito). Dos 87 arquivos analisados, em 04 deles a *tag* </article-title> foi filtrada, mas o texto do título não. O que veio no lugar foi a expressão (null) em vez da informação sobre o título. Foi observado, nessas quatro amostras, que esses títulos estavam representados com caracteres diferentes do tipo parte em negrito ou nota de rodapé, por exemplo. O arquivo incluía uma *tag* que confundia a lógica da filtragem. Segundo a lógica do AnoTex, como foi encontrada mais de

uma *tag*, uma nova indireção foi criada. Com isso, o título ficou situado num nível mais abaixo do qual deveria estar localizado.

Normalmente, seguindo as recomendações do guia de publicações da SciELO, o dado relevante do título, propriamente dito, fica no nível etiquetado como "article-title", conforme demonstrado na Figura 10.

```
<title-group>
    <article-title>Ensino de História, Diálogo Intercultural e Relações Étnico-Raciais</article-title>
</title-group>
```

Figura 10. Representação do Título.
Fonte: Adaptado pela autora do *CorpACE*.

Com a variação da marcação "bold" o título ficou em "article-title->bold->"o título"". A Figura 11 ilustra a referida variação.

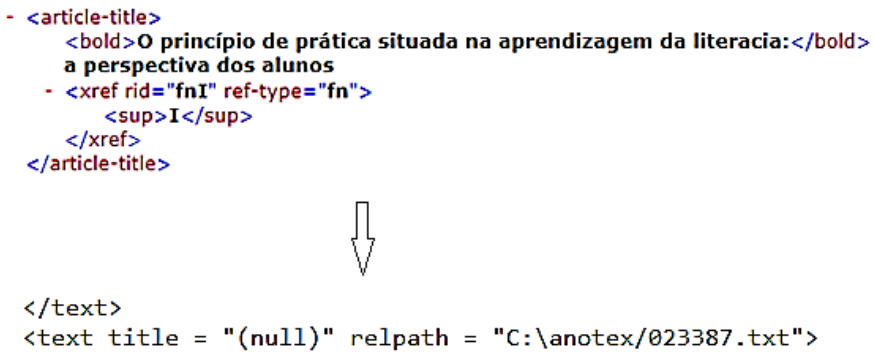

Figura 11. Título situado num nível mais abaixo do qual deveria ser.
Fonte: Adaptado pela autora do *CorpACE*.

Esse arquivo mostra uma clara variação na anotação, por parte da revista, em relação às recomendações do GUIA... (2016). Nesse caso, foi concentrada muita informação onde deveria haver apenas uma, mais direta e valiosa, no caso o título na forma como nós humanos usaríamos para referenciar o artigo. Uma vez que essas marcações são feitas pelos profissionais da Revista, isso demonstra que o título deveria passar por uma normalização antes de ser *tagueado*, pois permitiria maior aproveitamento do recurso de linguagem que está sendo disponibilizado. Para resolver esse problema, em específico, foi feita uma alteração para essa variação e caso o AnoTex achasse a *tag* "bold" deveria ignorá-la. No caso ilustrado foi retirada a *tag bold* (<bold> e </bold>) e o título voltou ser filtrado.

Outro recurso que demonstrou um pouco de variação em sua anotação foram os dados filtrados referentes aos elementos Sec-intro (seção introdução), Sec-discus (seção discussão), Sec-conclu (seção conclusão). Embora exista uma recomendação com as especificações no Guia de uso de elementos e atributos XML para documentos que seguem a implementação *SciELO Publishing Schema*, Versão 1.5.1 (GUIA..., 2016), conforme foi demonstrado na Tabela 2 - Tipo de seções, os artigos não seguem um padrão estrito e homogêneo de formato para a marcação do corpo do texto <body>. Sendo assim, foi considerado para compor o resultado apenas os artigos que possuíssem seções com o atributo <sec-type> (Figura 12).

Seguindo essa especificação, os resultados demonstraram que, em uma das amostras analisadas, um dos artigos estava sem marcação para o atributo seção, e em outras amostras a marcação para esse atributo não conferia com as seções padrão que são caracterizadas conforme recomendação do Guia. Na filtragem desses elementos, foi observado que às vezes as seções de primeiro nível do artigo apresentavam o atributo <sec-type> definidas pelo guia SciELO como: <sec sec-type="intro"> com 69 ocorrências; <sec sec-type="discussion"> com 82 ocorrências; e, <sec sec-type="conclusions"> com 69 ocorrências; às vezes vinham apenas com o atributo <sec> sem a especificação do tipo, com 17 ocorrências; e, às vezes não possuíam nenhuma marcação, todo o texto do artigo estava marcado apenas por parágrafos (isso foi observado em apenas uma amostra). Essa constatação possibilitou o

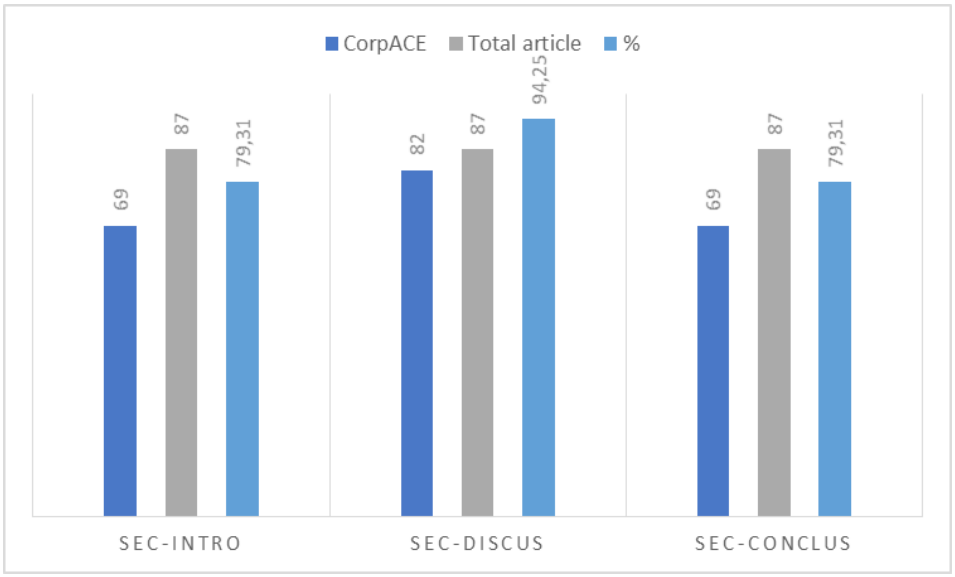

Figura 12. Filtragem dos elementos tipo de seção.
Fonte: Elaborado pela autora.

entendimento de que o sistema de marcação de dados não obriga a existência do atributo *sec/@sec-type*, ou seja, não é uma recomendação obrigatória para as revistas.

À medida que o uso das marcações em XML das seções dos artigos, foram analisadas para no processamento do texto, foram observadas algumas questões que confundiram a lógica do *parsing* do AnoTex. Em uma amostra analisada, por exemplo, uma seção de primeiro nível que condiz com uma lista de valores próximos, obrigatoriamente apresentava um atributo *@sec-type* e continha aninhadas dentro da seção de primeiro nível "introdução" as seções "discussão" e "conclusão". Por serem, também, de primeiro nível as seções "discussão" e "conclusão", pela lógica, não deveriam vir aninhadas dentro da seção "introdução". Nesse caso, como o guia não recomenda esse tipo de marcação não foi possível para o AnoTex prever a variação na marcação e, portanto, a ferramenta não conseguiu filtrar as seções que estavam aninhadas dentro da <sec sec-type="intro">. Para explicar esse caso, foram levantadas duas hipóteses para variação: poderia ter ocorrido um erro na produção do XML, ou as seções "discussão" e "conclusão" seriam dependentes diretas da primeira seção (introdução), por isso não houve outras seções no texto. Entretanto, o Guia não aborda esse tipo de marcação.

Essas variações, embora em pequeno número, revelaram ser imprescindível a aplicação de um padrão mais estrito e homogêneo na representação do XML pelas revistas para que elas sejam mais bem "rastreadas" *on-line* e processadas por ferramentas computacionais. As marcações, além de se tornarem mais úteis e mais semânticas, seriam facilitadas devido à padronização. Entretanto, essa realidade ainda está longe de ser aplicada em função da extensa quantidade de variação de formatos e dados a serem representados nos artigos. Além disso só para o padrão SciELO, existem 8 versões diferentes que vão adicionando *tag's* desde a versão 1.1 até 1.8. Os esquemas de codificação precisam evoluir para atender novas demandas de maneira a não quebrar compatibilidade com versões antigas dos programas, e os programas precisam estar preparados para lidar com versões antigas dos dados possibilitando a compatibilidade retroativa.

É um grande desafio atender às novas demandas e manter a compatibilidade retroativa com versões antigas, pois às linguagens artificiais seguem uma padronização que faz parte do desenvolvimento e da implementação tecnológica, são extremamente sensíveis ao contexto. Por outro lado, a língua natural, antes de ser representada/transformada pelas marcações em XML, por exemplo, deve ser entendida e normalizada. Entretanto, existem formas variadas de representação para um mesmo recurso da linguagem natural o que, em alguns casos, pode não favorecer a compatibilidade de interação com a linguagem artificial.

Embora constatada a ocorrência de algumas variações que confundiram a lógica do *parsing*, a

análise da maior parte dos dados coletados revelou que, dentre os elementos que podem ser filtrados pelo AnoTex, e, que compuseram o *CorpACE*, pode-se destacar uma estrutura básica do gênero artigo científico, constituída de elementos pré-textuais (<front>), textuais (<body>) e pós-textuais (<back>). Essa rigidez observada na estrutura da maioria dos 87 artigos coletados decorre exatamente da função do texto-base, que deve ser elaborado conforme normas preestabelecidas pelos periódicos, e, com fins e propósitos a que se destina, a fim de garantir a comunicação da cientificidade do assunto abordado. Isso ratifica o pressuposto de que a estrutura do gênero analisado é condicionada à sua função social.

As marcações em XML enriquecidas com informações linguísticas sobre a construção composicional do gênero, na forma de representações arbóreas em que se indicam as relações entre elementos do contexto de produção (constantes dos nós <front> <article-meta> </article-meta> </front>), ou sentenças ou fragmentos de sentenças da infraestrutura geral do texto (compreendidos em <body> <sec> </sec> </body> e <back> <ref-list> </ref-list> </back>), possibilitam a transformação do texto puro em texto estruturado. Por meio dessa representação, a partir de um mesmo XML, é possível a reutilização, apresentação da informação de formas distintas e indicação das principais características do gênero que podem ser usadas em aplicações da linguística computacional.

Essas marcações que foram filtradas, compiladas e exportadas pela ferramenta computacional AnoTex, geraram um arquivo de saída em formato XML. Uma das vantagens reveladas por esse tipo de configuração de arquivo de saída foi permitir a criação de um *corpus* que poderá ser usado por um número maior de ferramentas e de forma que atenda a propósitos variados, seguindo, via de regra, os dados que são disponibilizados pelo arquivo XML SciELO. Além disso, a separação entre a estrutura e apresentação dos elementos coletados dará maior maleabilidade ao corpus, pois, por meio dessa configuração, é possível a apresentação dos dados de formas distintas, além de permitir o carregamento dos metadados em uma base de dados.

Outro ponto importante a ser destacado é que cada elemento que compõe o *corpus* inclui um caminho relativo (*relpath*), nesse caso o *corpus* é composto por um arquivo XML principal ('CorpACE.xml') e de uma subpasta ('texts') contendo todos os textos na íntegra em txt. Esse caminho é importante, pois no processamento automático de texto quando uma ferramenta ler uma seção <text> ... </text>, saberá exatamente onde encontrar o texto correspondente na íntegra, utilizando esse procedimento. Essa funcionalidade é demonstrada pelas setas na Figura 13:

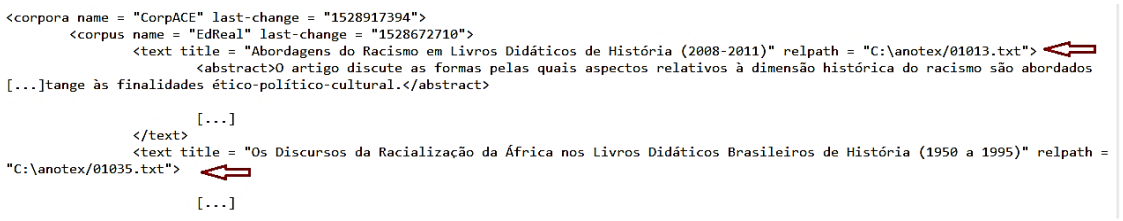

Figura 13. Caminho relativo para localização dos textos.
Fonte: Adaptado pela autora do AnoTex v0.1b.

A análise dos dados coletados ratificou o entendimento inicial da pesquisa que a estrutura dos artigos científicos está condicionada à sua função social, uma vez que são eventos linguísticos caracterizados por um conjunto de propósitos comunicativos e por serem tipos relativamente estáveis de enunciados. As marcações em XML das representações de aspectos relacionados com o *contexto de produção* do artigo científico podem influenciar a forma como o texto se organiza. Os elementos filtrados constitutivos dessas marcações revelam: emissor/enunciador, receptor/destinatário, lugar/instituição, suporte e conteúdo temático. Já da *arquitetura geral* do gênero artigo que está relacionada ao gerenciamento do conteúdo do texto - a construção composicional característica do gênero - a princípio, são filtradas pelo AnoTex as capacidades discursivas do gênero compreendidas em três subdivisões, caracterizadas nas Figuras: 14 denominada Elementos *Pré-textuais*, 15 denominada Elementos *Textuais* e 16 denominada Elementos *Pós-textuais*, a seguir:

1. Elementos *Pré-textuais* constantes do <front> como: periódico, título, autor, instituição, resumo

e palavras-chave (Figura 14).

```
<front>
        <journal-meta>
                <journal-title>Educação &amp; Realidade</journal-title>
                <abbrev-journal-title abbrev-type="publisher">Educ. Real.</abbrev-journal-title>
            </journal-title-group>
        </journal-meta>
        <article-meta>
            <title-group>
                <article-title>Abordagens do Racismo em Livros Didáticos de História (2008-2011)</article-title>
            </title-group>
            <contrib-group>
                <contrib contrib-type="author">
                    <name>
                        <surname>Roza</surname>
                        <given-names>Luciano Magela</given-names>
                    </name>
                <institution content-type="original">Universidade Federal dos Vales do Jequitinhonha e Mucuri (UFVJM),
Diamantina/MG - Brasil</institution>
                <abstract>
                <title>Resumo:</title>
                <p>O artigo discute as formas pelas quais aspectos relativos à dimensão histórica do racismo são abordados em
atividades propostas em um conjunto de Livros Didáticos de História aprovado nas edições do PNLD 2008 e 2011. Nas propostas de atividades
analisadas se buscou perceber como o tema do racismo foi abordado e quais diálogos foram construídos com a história do pós-abolição e o
ensino de História. Os Resultados obtidos demonstram que uma diversidade de abordagens tem caracterizado o tratamento do tema,
especialmente, no que tange às finalidades ético-político-cultural.</p>
                </abstract>
                <kwd-group xml:lang="pt">
                    <title>Palavras-chave:</title>
                    <kwd>Ensino de História</kwd>
                    <kwd>Lei 10.639/03</kwd>
                    <kwd>Atividades</kwd>
                    <kwd>Livros Didáticos</kwd>
                    <kwd>Pós-Abolição.</kwd>
                </kwd-group>
        </article-meta>
</front>
```

Figura 14. Elementos Pré-textuais.
Fonte: Adaptado pela autora do *CorpACE*.

2. Elementos *Textuais* constantes do <body> como: introdução, discussão e considerações finais (principais seções do artigo) (Figura 15).

```
body>
        <sec sec-type="intro">
            <title>Introdução</title>
        </sec>
        <sec sec-type="discussion">
            <title>Apontamentos acerca da Historiografia sobre o Pós-abolição e suas Repercussões Didáticas</title>
        </sec>
        <sec sec-type="discussion">
            <title>Reflexões sobre Atividades na História Escolar e em Livros Didáticos</title>
        </sec>
        <sec sec-type="discussion">
            <title>A Utilização da Trajetória Histórica do Racismo como Tema em Atividades Pedagógicas em Livros Didáticos de
História</title>
        </sec>
        <sec sec-type="conclusions">
            <title>Considerações Finais</title>
        </sec>
</body>
```

Figura 15. Elementos Textuais.
Fonte: Adaptado pela autora do *CorpACE*.

3. Elementos Pós-textuais constantes do <back> como: referências bibliográficas (Figura 16).

Essa configuração do modelo computacional representativo para o gênero artigo científico enfatizado no uso de etiquetas XML permitiu destacar as características e a visualização das dimensões constitutivas do gênero, e delimitar os objetivos a serem atingidos em relação aos diferentes propósitos de seu uso. Além disso, essa representação arbórea dos elementos constitutivos do corpus pode dar pistas das características do gênero que podem ser mineradas e valoradas para o processamento do texto.

A indicação da localização e quantidade de ocorrências da filtragem dos elementos em negrito e itálico podem se tornar relevantes uma vez que os elementos mais conclusivos e significativos podem aparecer em páginas mais próximas da seção "introdução" (<sec sec-type="intro">) ou da seção "conclusão" (<sec sec-type="conclusions">) nos artigos científicos (LIN; HOVY, 1997). Nesse caso, por meio da estrutura básica do gênero, poderia ser estabelecido um peso para o termo, observando-se a página onde ele ocorresse, além do número de vezes que ele fosse encontrado e marcado ao longo de todo artigo. Do ponto de vista linguístico, essa é uma informação que ainda precisa ser explorada na valoração do ranqueamento das sentenças.

```
<back>
        <ref-list>
                <title>Referências</title>
                <ref>
                </ref>
        </ref-list>
</back>
```

Figura 16. Elementos Pós-textuais.
Fonte: Adaptado pela autora do *CorpACE*.

5 Conclusão

Na medida em que foram analisadas, na prática, as marcações em XML (enriquecidas com informações linguísticas sobre a construção composicional do gênero, que indicam as relações entre elementos do contexto de produção e/ou da infraestrutura geral do texto-base), para o processamento do texto por ferramenta de PLN, foram verificadas algumas questões que confundiram a lógica do *parsing* do AnoTex. Entretanto, a análise da maior parte dos dados coletados revelou que dentre os elementos que podem ser filtrados pelo AnoTex, e, que compuseram o *CorpACE*, pode-se destacar uma estrutura básica do gênero artigo científico, constituída de elementos pré-textuais (<front>), textuais (<body>) e pós-textuais (<back>). Essa rigidez observada na estrutura da maioria dos 87 artigos coletados decorre exatamente da função do texto-base, que deve ser elaborado conforme normas preestabelecidas pelos periódicos, e, com fins e propósitos a que se destina, com o objetivo de garantir a comunicação da cientificidade do assunto abordado. O que ratifica o pressuposto de que a estrutura do gênero analisado é condicionada à sua função social.

Diante da análise sobre as ferramentas para anotação em *corpus*, atualmente disponíveis, argumenta-se que essas não descrevem suficientemente as características do gênero textual base em combinação com um nível adequado de facilidade de uso. Uma vez que a principal finalidade da abordagem dessas ferramentas é a anotação morfossintática e semântica, ficou demonstrada uma lacuna na anotação da estrutura do gênero base. Sendo assim, é possível afirmar que há necessidade de uma estreita relação entre o sistema de anotação automática de texto com a análise do gênero do texto-base. E, por meio dessa abordagem, o AnoTex preenche uma importante lacuna no desenvolvimento das aplicações para a identificação e análise de dados estruturados representativos do gênero, de modo a proporcionar novas diretrizes na área de processamento de grandes quantidades de texto. Uma vez que os textos são eventos comunicativos carregados de propósitos, as decisões sobre as suas propriedades relevantes implicam reconhecer as propriedades da fonte, uma vez que estas são alimentadas por um conjunto de requisitos de propósitos comunicativos.

Outro ponto importante que merece ser destacado é que os esquemas de codificação precisam evoluir para atender novas demandas de maneira a não quebrar compatibilidade com versões antigas dos programas, e os programas precisam estar preparados para lidar com versões antigas dos dados possibilitando a compatibilidade retroativa. Entretanto, a língua natural, antes de ser representada/-transformada em linguagem artificial, pelas marcações em XML por exemplo, deve ser entendida e normalizada, sob a pena de em alguns casos não favorecer a compatibilidade de interação.

A tecnologia deve ser abordada explorando os múltiplos conhecimentos necessários para o aperfeiçoamento das características linguísticas dos aplicativos com ênfase em PLN, ao invés da superespecialização. Só assim novos estudos serão abordados e, dentre eles, poderá existir uma possibilidade para a união de conhecimentos distintos, de forma a tornar mais eficientes a comunicação entre homens e máquinas.

Referências

ALENCAR, A. F. de. *About Aelius Brazilian Portuguese POS-Tagger*. [S.l.: s.n.], 2013. Disponível em: <http://aelius.sourceforge.net/>. Acesso em: 27 dez. 2021.

ALENCAR, A. F. de. *Aelius User's Manual*. [S.l.: s.n.], 2013. Disponível em: <http://aelius.sourceforge.net/manual.html>. Acesso em: 27 dez. 2021.

ALENCAR, L. F de. Aelius: uma ferramenta para anotação automática de corpora usando o NLTK. In: IX Encontro de Linguística de Corpus. Porto Alegre: PUCRS, 2010. Disponível em: <http://corpuslg.org/gelc/media/blogs/elc2010/slides/Figueiredo_de_Alencar.pdf>. Acesso em: 27 dez. 2021.

BAKHTIN, M. M. *Estética da criação verbal*. São Paulo: Martins Fontes, 1997.

BHARTI, S. K.; BABU, K. S. Automatic Keyword Extraction for Text Summarization: a survey. *European Journal of Advances in Engineering and Technology*, v. 4, n. 6, p. 410–427, 2017.

BHATIA, N.; JAISWAL, A. Literature Review on Automatic Text Summarization: Single and Multiple Summarizations. *International Journal of Computer Applications*, v. 117, n. 6, p. 25–29, mai. 2015. DOI: 10.5120/20560-2948. Disponível em: <http://research.ijcaonline.org/volume117/number6/pxc3902948.pdf>. Acesso em: 27 dez. 2021.

BRONCKART, J.-P. *Atividade de linguagem, textos e discursos. Por um interacionismo sócio-discursivo*. São Paulo: EDUC, 1999.

CAMBRIA, E.; WHITE, B. Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, v. 9, n. 2, p. 48–57, mai. 2014. DOI: 10.1109/MCI.2014.2307227. Disponível em: <http://ieeexplore.ieee.org/document/6786458/>. Acesso em: 27 dez. 2021.

COHEN, J. D. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science*, v. 46, n. 3, p. 162–174, abr. 1995. DOI: 10.1002/(SICI)1097-4571(199504)46:3<162::AID-ASI2>3.0.CO;2-6. Disponível em: <https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-4571(199504)46:3%3C162::AID-ASI2%3E3.0.CO;2-6>. Acesso em: 27 dez. 2021.

DE OLIVEIRA JÚNIOR, R. L.; ESMIN, A. A. A. Monitoramento Automático de Mensagens de Fóruns de Discussão de Texto Semi-Supervisionado. In: SBIE - Simpósio Brasileiro de Informática na Educação. Rio de Janeiro: SBIE, 2012.

DIMA, E. et al. A Metadata Editor to Support the Description of Linguistic Resources. In: PROCEEDINGS of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey: European Language Resources Association (ELRA), mai. 2012. p. 1061–1066. Disponível em: <http://www.lrec-conf.org/proceedings/lrec2012/pdf/468_Paper.pdf>. Acesso em: 27 dez. 2021.

DOMINGUES, M. L.; FAVERO, E. L.; DE MEDEIROS, I. P. O desenvolvimento de um etiquetador morfossintático com alta acurácia para o português. In: VALE, O. A. (Ed.). *Avanços da Linguística de Corpus no Brasil*. São Paulo: Humanistas, 2008. p. 267–286.

FIALHO, P. et al. INESC-ID@ASSIN: Medição de Similaridade Semântica e Reconhecimento de Inferência Textual. *Linguamática*, v. 8, n. 2, p. 33–42, dez. 2016. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/v8n2-4>. Acesso em: 27 dez. 2021.

FONSECA, C. A. *AnoTex: anotador de artigo científico para retextualização automática*. 2018. Dissertação (Mestrado Profissional em Educação) – Universidade Federal dos Vales do Jequitinhonha e Mucuri, Diamantina. Disponível em: <http://acervo.ufvjm.edu.br/jspui/handle/1/2114>. Acesso em: 27 dez. 2021.

FONSECA, C. A. et al. AnoTex: rotina de filtragem de dados estruturados do gênero artigo científico como contribuição para o PLN. *Texto Livre: Linguagem e Tecnologia*, v. 11, n. 3, p. 40–64, dez. 2018. DOI: 10.17851/1983-3652.11.3.40-64. Disponível em: <https://periodicos.ufmg.br/index.php/textolivre/article/view/16811>. Acesso em: 27 dez. 2021.

FRAKES, W. B.; BAEZA-YATES, R. (Ed.). *Information retrieval: data structures & algorithms*. Englewood Cliffs, N.J: Prentice Hall, 1992.

GUIA de uso de elementos e atributos XML para documentos que seguem a implementação SciELO Publishing Schema. — SciELO Publishing Schema 1.5.1 documentation. [S.l.: s.n.], 2016. Disponível em: <http://docs.scielo.org/projects/scielo-publishing-schema/pt_BR/1.5-branch/>. Acesso em: 27 dez. 2021.

JONES, K. S. Automatic summarising: The state of the art. *Information Processing & Management*, v. 43, n. 6, p. 1449–1481, nov. 2007. DOI: 10.1016/j.ipm.2007.03.009. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0306457307000878>. Acesso em: 27 dez. 2021.

JONES, K. S. Some thesauric history. *Aslib Proceedings*, v. 24, n. 7, p. 400–411, jul. 1972. DOI: 10.1108/eb050353. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/eb050353/full/html>. Acesso em: 27 dez. 2021.

JONES, K. S.; WALKER, S.; ROBERTSON, S.E. A probabilistic model of information retrieval: development and comparative experiments. *Information Processing & Management*, v. 36, n. 6, p. 779–808, nov. 2000. DOI: 10.1016/S0306-4573(00)00015-7. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0306457300000157>. Acesso em: 27 dez. 2021.

KUCUK, M. E.; OLGUN, B.; SEVER, H. *Application of Metadata Concepts to Discovery of Internet Resources*. [S.l.: s.n.], 2000. DOI: 10.1007/3-540-40888-6_29. Disponível em: <https://www.infona.pl//resource/bwmeta1.element.springer-f35534a0-afa3-3605-a53e-d291fae9c131>. Acesso em: 27 dez. 2021.

LANDAUER, T. K; FOLTZ, P. W.; LAHAM, D. An introduction to latent semantic analysis. *Discourse Processes*, v. 25, n. 2-3, p. 259–284, jan. 1998. DOI: 10.1080/01638539809545028. Disponível em: <http://www.tandfonline.com/doi/abs/10.1080/01638539809545028>. Acesso em: 27 dez. 2021.

LAU, R. Y. K. et al. Towards Fuzzy Domain Ontology Based Concept Map Generation for E-Learning. In: LEUNG, H. et al. (Ed.). *Advances in Web Based Learning – ICWL 2007*. Berlin, Heidelberg: Springer, 2008. (Lecture Notes in Computer Science), p. 90–101. DOI: 10.1007/978-3-540-78139-4_9.

LIN, C.-Y.; HOVY, E. Identifying topics by position. In: PROCEEDINGS of the fifth conference on Applied natural language processing -. Washington, DC: Association for Computational Linguistics, 1997. p. 283–290. DOI: 10.3115/974557.974599. Disponível em: <http://portal.acm.org/citation.cfm?doid=974557.974599>. Acesso em: 27 dez. 2021.

LIN, F.-R.; HSIEH, L.-S.; CHUANG, F.-T. Discovering genres of online discussion threads via text mining. *Computers & Education*, v. 52, n. 2, p. 481–495, fev. 2009. DOI: 10.1016/j.compedu.2008.10.005. Disponível em: <https://linkinghub.elsevier.com/retrieve/pii/S0360131508001528>. Acesso em: 27 dez. 2021.

LIU, X.; LI, C.; FENG, Z. Analyze of Subject Research Hot Spots Based on An Improved Algorithm of TF*IDF——Taking Information Science for Example– Information Science 2017 07 . *Information Science*, v. 7, n. 35, p. 015, 2017. Disponível em: <http://en.cnki.com.cn/Article_en/CJFDTotal-QBKX201707015.htm>. Acesso em: 27 dez. 2021.

LIU, X.; WEBSTER, J. J.; KIT, C. An Extractive Text Summarizer Based on Significant Words. In: LI, W.; MOLLÁ-ALIOD, D. (Ed.). *Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*. Berlin, Heidelberg: Springer, 2009. (Lecture Notes in Computer Science), p. 168–178. DOI: 10.1007/978-3-642-00831-3_16.

LOVINS, J. B. Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, v. 11, n. 1, p. 22–31, 1968.

LUHN, H. P. The Automatic Creation of Literature Abstracts. *IBM Journal of Research and Development*, v. 2, n. 2, p. 159–165, abr. 1958. DOI: 10.1147/rd.22.0159. Disponível em: <http://ieeexplore.ieee.org/document/5392672/>. Acesso em: 27 dez. 2021.

LUI, A. K.-F.; LI, S. C.; CHOY, S. O. An Evaluation of Automatic Text Categorization in Online Discussion Analysis. In: SEVENTH IEEE International Conference on Advanced Learning Technologies (ICALT 2007). [S.l.: s.n.], jul. 2007. p. 205–209. DOI: 10.1109/ICALT.2007.59.

LYSE, G. I.; MEURER, P.; DE SMEDT, K. COMEDI: A component metadata editor. In: SELECTED papers of the CLARIN Annual Conference 2014. Bergen, Norway: Linköping University Electronic Press, 2015. Disponível em: <https://ep.liu.se/en/conference-article.aspx?series=ecp&issue=116&Article_No=8>. Acesso em: 27 dez. 2021.

MANARIS, B. Natural Language Processing: A Human-Computer Interaction Perspective. *Advances in Computers*, v. 47, n. 100, p. 1–66, 1998.

MARCANTONIO, A. T.; SANTOS, M. dos; LEHFELD, N. A. de S. *Elaboração e divulgação do trabalho científico*. [S.l.]: Atlas, 1993.

MARCONI, M. de A.; LAKATOS, E. M. *Fundamentos de metodologia científica*. São Paulo: Atlas, 2010.

MARCUSCHI, L. A. Gêneros textuais emergentes no contexto da tecnologia digital. In: MARCUSCHI, L. A.; XAVIER, A. C. (Ed.). *Hipertexto e gêneros digitais: novas formas de construção de sentido*. Rio de Janeiro: Lucerna, 2004. p. 13–67.

MARCUSCHI, L. A. Gêneros textuais: definição e funcionalidade. In: DIONÍSIO, A. P.; MACHADO, A. R.; BEZERRA, M. A. (Ed.). *Gêneros textuais e ensino*. Rio de Janeiro: Lucerna, 2002. p. 19–36.

MATENCIO, M. de L. M. Atividade de (Re)textualização em práticas acadêmicas: um estudo do resum. *Scripta*, v. 6, n. 11, p. 109–122, out. 2002. Disponível em: <http://periodicos.pucminas.br/index.php/scripta/article/view/12453>. Acesso em: 27 dez. 2021.

REITER, E. A Structured Review of the Validity of BLEU. *Computational Linguistics*, v. 44, n. 3, p. 393–401, set. 2018. DOI: 10.1162/coli_a_00322. Disponível em: <https://direct.mit.edu/coli/article/44/3/393-401/1598>. Acesso em: 27 dez. 2021.

ROCHA, V. J. C.; GUELPELI, M. V. C. PragmaSUM: automatic tex summarizer based on user profile. *International Journal of Current Research*, v. 9, n. 7, p. 53935–53942, 2017.

ROLIM, V.; FERREIRA, R.; COSTA, E. Identificação Automática de Dúvidas em Fóruns Educacionais. In: p. 936. DOI: 10.5753/cbie.sbie.2016.936. Disponível em: <http://www.br-ie.org/pub/index.php/sbie/article/view/6779>. Acesso em: 27 dez. 2021.

SALTON, G.; MCGILL, M. J. *Introduction to modern information retrieval*. New York: McGraw-Hill, 1983. (McGraw-Hill computer science series).

SANTOS, C. N. dos; ZADROZNY, B. Learning Character-level Representations for Part-of-Speech Tagging. In: PROCEEDINGS of the 31st International Conference on Machine Learning (ICML-14). [S.l.: s.n.], 2014. p. 1818–1826.

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

SCARTON, C. E.; ALUÍSIO, S. M. Análise da Inteligibilidade de textos via ferramentas de Processamento de Língua Natural: adaptando as métricas do Coh-Metrix para o Português. *Linguamática*, v. 2, n. 1, p. 45–61, abr. 2010. Disponível em: <https://linguamatica.com/index.php/linguamatica/article/view/44>. Acesso em: 27 dez. 2021.

SCHNEUWLY, B.; DOLZ, J. *Gêneros orais e escritos na escola*. Campinas, SP: Mercado de Letras, 2004.

SILVA, B. C. D. da. O estudo Lingüístico-Computacional da Linguagem. *Letras de Hoje*, v. 41, n. 2, set. 2006. Disponível em: <https://revistaseletronicas.pucrs.br/ojs/index.php/fale/article/view/597>.

SILVA, J. G. B.; PEREIRA, M. T. B. F.; BUENO, L. A elaboração de um artigo científico: subsídios à apropriação desse gênero textual. *Horizontes*, v. 32, n. 1, jun. 2014. DOI: 10.24933/horizontes.v32i1.88. Disponível em: <https://revistahorizontes.usf.edu.br/horizontes/article/view/88>. Acesso em: 27 dez. 2021.

SILVA, L. A.; TRINDADE, D. et al. Mineração de Dados em publicações de Fóruns de Discussões do Moodle como geração de Indicadores para aprimoramento da Gestão Educacional. *Anais dos Workshops do Congresso Brasileiro de Informática na Educação*, v. 4, n. 1, p. 1084, out. 2015. DOI: 10.5753/cbie.wcbie.2015.1084. Disponível em: <http://br-ie.org/pub/index.php/wcbie/article/view/6220>. Acesso em: 27 dez. 2021.

SOUSA, M. C. P. de. O Corpus Tycho Brahe: contribuições para as humanidades digitais no Brasil. *Filologia e Linguística Portuguesa*, v. 16, spe, p. 53, dez. 2014. DOI: 10.11606/issn.2176-9419.v16ispep53-93. Disponível em: <http://revistas.usp.br/flp/article/view/88404>. Acesso em: 27 dez. 2021.

SOUSA, M. C. P. de; KEPLER, F. N.; FARIA, P. P. F. de. E-Dictor: novas perspectivas na codificação e edição de corpora de textos históricos. In: CAMINHOS da Linguística de Corpus. São Paulo: Mercado de Letras, 2010. p. 225–246.

SOUSA, M. C. P. de; KEPLER, F. N.; FARIA, P. P. F. de. Uma proposta de automatização das edições XML do e-Dictor. In: ANAIS. [S.l.: s.n.], 2016. Disponível em: <https://sefuefs2015.wordpress.com/>. Acesso em: 27 dez. 2021.

SWALES, J. *Genre analysis: English in academic and research settings*. Cambridge: Cambridge Univ. Pr, 1990. (Cambridge applied linguistics series).

TONELLI, S.; PIANTA, E. Matching documents and summaries using key-concepts Sara. In: PROCEEDINGS of the Seventh DEFT Workshop. Montpellier, France: [s.n.], 2011. p. 73–83.

VIEIRA, R.; LIMA, V. L. S. de. Lingüística computacional: princípios e aplicações. In: ANAIS do XXI Congresso da SBC. I Jornada de Atualização em Inteligência Artificial. [S.l.: s.n.], 2001. Disponível em: <http://www.inf.unioeste.br/~jorge/MESTRADOS/LETRAS%20-%20MECANISMOS%20DO%20FUNCIONAMENTO%20DA%20LINGUAGEM%20-%20PROCESSAMENTO%20DA%20LINGUAGEM%20NATURAL/ARTIGOS%20INTERESSANTES/lingu%EDstica%20computacional.pdf>. Acesso em: 27 dez. 2021.

WEBSTER, J. J.; KIT, C. Tokenization as the initial phase in NLP. en. In: PROCEEDINGS of the 14th conference on Computational linguistics -. Nantes, France: Association for Computational Linguistics, 1992. v. 4, p. 1106. DOI: 10.3115/992424.992434. Disponível em: <http://portal.acm.org/citation.cfm?doid=992424.992434>. Acesso em: 27 dez. 2021.

Contribuições dos autores

Claudia Aparecida Fonseca: Conceituação, Análise Formal, Investigação, Metodologia, Escrita – rascunho original, Escrita – revisão e edição; **Marcus Vinícius Carvalho Guelpeli**: Curadoria de dados, Análise Formal, Programas, Administração de projetos, Supervisão, Visualização; **Rafael Santiago de Souza Netto**: Análise Formal, Programas, Visualização, Escrita – rascunho original .