Rev. Estud. Ling., Belo Horizonte, v. 29, n. 2, p. 1059-1087, 2021

O papel do corpus de estudo no aprimoramento descritivo da complementaridade informacional multidocumento

The role of the study corpus in the descriptive improvement of multi-document informational complementarity

Jackson Wilke da Cruz Souza
Universidade Federal de Alfenas (UNIFAL-MG), Varginha, Minas Gerais / Brasil
jackcruzsouza@gmail.com
http://orcid.org/0000-0003-1881-6780

Resumo: Em subáreas do Processamento Automático de Línguas Naturais (PLN), como a Sumarização Automática Multidocumento (SAM), é necessário compreender o comportamento linguístico de determinados fenômenos, especialmente os de natureza semântica. A *Cross-document Structure Theory* (CST) é bastante utilizada em estudos do PLN por proporcionar um conjunto de relações semânticas que organizam a informação entre unidades de análise (comumente, pares de sentenças), agrupadas entre *conteúdo* (a saber, redundância, complementaridade e contradição) e *apresentação* (a saber, fonte/autoria e estilo). Até então, a caracterização das relações CST baseava-se em atributos *genéricos* (como a quantidade de palavras em comum entre as sentenças de um par) e *específicos* (como a presença de advérbios temporais) para as relações de Redundância e Complementaridade. Entretanto, percebe-se que a delimitação de tais atributos ainda é incipiente, pois não inclui atributos semânticos e pragmáticos, níveis linguísticos que são possíveis de recuperar manualmente entre as unidades de análise da CST. Nesse sentido, objetiva-se, neste artigo, reconstruir o percurso metodológico de Souza (2019) ao que se refere ao estudo em *corpus* das relações CST em textos jornalísticos do Português, já que o conjunto de atributos disponíveis, até o momento, ainda produzia equívocos na identificação dos subtipos de complementaridade multidocumento, a saber temporal e atemporal. Partindo do *corpus* CSTNews, organizou-se um subconjunto de estudo com os 10 primeiros *clusters*, o que contabilizou 204 pares de sentenças. Como resultado, foram obtidas a descrição detalhada da complementaridade CST e a criação de uma tipologia de sinalizadores das relações que traduzem esse fenômeno, além da proposição de uma metodologia específica para o estudo de relações CST.

Palavras-chave: Complementaridade informacional multidocumento; Processamento Automático de Línguas Naturais; *Corpus* de estudo.

eISSN: 2237-2083
DOI: 10.17851/2237-2083.29.2.1059-1087

Abstract: In sub-areas of Natural Language Processing (NLP), such as Automatic Multidocument Summarization (AMS), it is necessary to understand the linguistic behavior of certain phenomena, especially those of a semantic nature. Cross-document Structure Theory (CST) is widely used in NLP studies because it provides a set of semantic relations that organize information between units of analysis (commonly, pairs of sentences) organized between *content* (namely, redundancy, complementarity and contradiction) and *presentation* (namely, source/authorship and style). Until then, the characterization of CST relationships was based on *generic attributes* (such as the number of words in common between sentences of a pair) and *specific attributes* (such as the presence of temporal adverbs) for the relationships of Redundancy and Complementarity. However, the delimitation of such attributes is still incipient, as they do not include semantic and pragmatic attributes, linguistic levels that are possible to recover between the CST units of analysis. In this sense, the aim of this paper is to reconstruct the methodological path of Souza (2019) with regard to the study in *corpus* of CST relations in Portuguese journalistic texts, since the set of available attributes, until then, still produced mistakes in the identification of multi-document complementarity subtypes, namely temporal and timeless. Based on the CSTNews *corpus*, a subset of studies was organized with the first 10 clusters, that are represented by 204 pairs of sentences. As a result, a detailed description of CST complementarity was obtained, as well as the creation of a typology of signaling relationships that translate this phenomenon, in addition to proposing a specific methodology for the study of CST relations.

Keywords: Multi-document informational Complementarity; Processing of Natural Languages; Study *corpus*.

Recebido em 10 de outubro de 2020

Aceito em 04 de janeiro de 2021

1 Introdução

As pesquisas na área de Linguística de *Corpus* (doravante, LC) têm se dedicado a estudar fenômenos linguísticos a partir de textos produzidos, em sua maioria, por humanos. Assim, derivam-se os estudos em Terminologia (TAGNIN; BEVILACQUA, 2015), em Linguística Aplicada (VIANA; TAGNIN, 2011), em Linguística Descritiva (RODRIGUES, 2019) ou em Processamento de Automático de Línguas Naturais (PLN) (CASELI, 2015).

Os *corpora*, de maneira geral, permitem identificar características que destacam a evidência de certos fenômenos em um ambiente linguisticamente natural. Ao tentar delimitar um candidato a termo de uma área médica, por exemplo, a análise do texto que está ao seu redor dará pistas ao pesquisador se aquele candidato é um hiperônimo ou hipônimo de outro termo. É nesse processo de observação dos fenômenos linguísticos nos *corpora* que é possível corroborar teorias, desenvolvê-las ou mesmo refutá-las.

De acordo com Sardinha (2000), um *corpus* pode ter finalidades de estudo, referência e treinamento (ou teste). O *corpus* de estudo subsidia análises preliminares do objeto e/ou do fenômeno em observação. O *corpus* de referência oferece suporte a uma análise contrastiva entre este e o *corpus* de estudo. Já o *corpus* de treinamento é submetido a testes que se pautarão no conhecimento levantado a partir das observações realizadas nos *subcorpora* de estudo e de referência.

Dentre as diversas contribuições que a Linguística de *corpus* pode promover à Computação, destacam-se, aqui, o estudo, a descrição e a caracterização de fenômenos linguísticos em *corpus*. Ao analisarem o comportamento linguístico da complementaridade entre textos jornalísticos do português, Souza (2015) e Souza e Di-Felippo (2018) basearam-se em características de maior acurácia recomendadas pela literatura, como a presença de *expressões temporais*. A partir da proposição de um conjunto de atributos, os autores submeteram os pares de sentenças ao *Waikato Environment for Knowledge Analysis* (Weka) (HALL *et al.,* 2009), resultando em algoritmos de Aprendizado de Máquina (AM) que discriminam as relações semânticas de complementaridade (a saber, *Historical Background*, *Follow-up* e *Elaboration*) da *Cross-document Structure Theory* (CST) (RADEV, 2000). Como resultado, os algoritmos de AM propostos obtiveram cerca de 75% de precisão em identificar as relações dos pares de sentença.

Entretanto, observou-se que entre as relações *Follow-up* e *Elaboration*, classificadas por Maziero (2012) e Maziero, Jorge e Pardo (2010), ainda havia equívocos devido à similaridade entre elas. De acordo com os autores, *Follow-up* ocorre quando, em um par de sentenças (S1 e S2), S2 apresenta acontecimentos que ocorrem após os acontecimentos em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si. Já a relação *Elaboration* ocorre quando, dado o par de sentenças, S2 detalha/refina/elabora algum

elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Para tanto, exemplificam-se esses apontamentos em (1).

(1)

S1: Confrontos entre o Exército e o grupo rebelde Tigres Tâmeis eclodiram na região de Muttur há duas semanas, após a guerrilha ter cortado o suprimento de água para alguns vilarejos.

S2: Os rebeldes afirmaram que consideram o novo bombardeio do Exército equivalente a "uma declaração de guerra".

De acordo com a definição proposta por Maziero (2012) e Maziero, Jorge e Pardo (2010), o par de sentenças em (1) pode ser identificado tanto como *Follow-up* (pois o evento narrado em S2 ocorre após S1), como *Elaboration* (pois S2 apresenta detalhes acerca da afirmação dos rebeldes indicados em S1). São em casos semelhantes a esse que os algoritmos desenvolvidos por Souza (2015) e Souza e Di-Felippo (2018) cometiam equívocos de classificação.

Além da similaridade entre as relações CST de complementaridade, até então, os estudos se baseavam em descrever as relações com base em atributos presentes (ou ausentes) na superfície textual. Assim, em um par de sentenças que ocorresse *advérbio ou locução adverbial de tempo*, potencialmente poderia ser classificado como complementaridade temporal. Isso muito se deve ao fato de a descrição estar bastante empenhada em promover algoritmos automáticos que pudessem identificar e classificar as relações CST. Nesse sentido, de certa maneira, foram deixados de lado atributos que não pudessem ser processados computacionalmente, sob a justificativa que tais atributos não deixam evidências de relações semânticas no texto.

Nesse contexto, Das e Taboada (2018) e Taboada e Das (2013) advogam que qualquer fenômeno semântico se expressa por meio do texto e sempre deixa marcas ou mesmo indícios para apontar informações adicionais para fora dele. Os autores reanotaram o *RST Signalling Corpus* (DAS; TABOADA; MCFETRIDGE, 2015), o qual contém textos em que as proposições estão estruturadas de acordo com o modelo *Rhetorical Structure Theory* (RST) (MANN; THOMPSON, 1987). O objetivo dos estudos foi identificar como os marcadores discursivos (como *conjunções ou locuções conjuntivas de adversidade*, por exemplo) contribuem para o sentido do discurso ao sinalizar relações no texto. Após esses estudos, os

autores organizaram tipologicamente os sinais em *genéricos* e *específicos*, que podem ocorrer sozinhos ou combinados.

Para estudar o comportamento linguístico-estrutural do modelo CST e identificar os sinalizadores de suas relações, é necessário conceber esse modelo como teoria semântica, ainda que suas relações não sejam intencionais. Assim, baseando-se no aprofundamento da análise da complementaridade proposta por Souza (2019), objetiva-se apresentar detidamente os procedimentos metodológicos acerca do estudo da complementaridade em *corpus*. Além do próprio estudo do fenômeno, objetiva-se salientar o lugar do *corpus* de estudo na LC como ferramenta de desenvolvimento de teorias linguísticas ou mesmo de suas possíveis revisões e/ou aprimoramentos.

Para tanto, este artigo está organizado em cinco seções. Na Seção 2, apresentam-se algumas reflexões sobre a LC, a fim de destacar os critérios de construção e anotação de *corpora* linguísticos. Na Seção 3, tem-se o panorama acerca dos sinalizadores linguísticos sob a perspectiva de duas teorias discursivas distintas, a RST e a CST. Na Seção 4, demonstra-se o estudo baseado em *corpus*, que resultou no levantamento de sinalizadores da complementaridade via modelo CST. Por fim, na Seção 5, tecem-se considerações finais, além de se delinearem trabalhos futuros.

2 Da construção à anotação de *corpus*

É notável como a LC passou por diversos aprimoramentos teóricos e metodológicos nos últimos anos, especialmente com a contribuição de abordagens computacionais. Tagnin (2018) aponta que, desde a publicação do *Brown University Standard Corpus of Present-Day American*, em 1964, o cenário na LC tem mudado. Essas mudanças são promovidas pela utilização de ferramentas e abordagens computacionais na área, o que resultou em transformações de perspectivas teóricas e técnicas sobre o conceito de *corpus* e como ele pode ser utilizado em pesquisas linguísticas.

Com relação ao conceito de *corpus*, Sardinha (2004) propõe que esse recurso linguístico pode ser definido como

> Um conjunto de dados linguísticos (pertencentes ao uso oral ou escrito da língua, ou a ambos), sistematizados segundo determinados critérios, suficientemente extensos em amplitude e

> profundidade, de maneira que sejam representativos da totalidade do uso linguístico ou de algum de seus âmbitos, dispostos de tal modo que possam ser processados por computador, com a finalidade de propiciar resultados vários e úteis para a descrição e análise. (SARDINHA, 2004, p.18.)

A partir dessa definição, é possível resgatar alguns critérios relevantes e necessários para que dado conjunto de textos não seja tido apenas como um arquivo, coletânea ou biblioteca de textos.

O primeiro critério que Sardinha (2004) chama à atenção é o *tamanho*, a fim de garantir que o fenômeno a ser observado possa ocorrer no conjunto de textos. Assim, o autor classifica o conjunto em *pequeno* (menos de 80 mil palavras), *pequeno-médio* (80 a 250 mil palavras), *médio* (250 mil a 1 milhão), *médio-grande* (1 milhão a 10 milhões) e *grande* (acima de 10 milhões de palavras). O *Brown Corpus*, por exemplo, tinha 2.000 palavras (frente a cerca de 1 milhão atualmente), enquanto o *Corpus* do Núcleo Interinstitucional de Linguística Computacional (NILC) possui mais de 40 milhões de palavras (KUHN; ABARCA; NUNES, 2000) e o iWeb, mais de 14 bilhões (DAVIES; KIM, 2019). O aumento no tamanho dos *corpora* deve-se, certamente, às facilidades que os sistemas e ferramentas computacionais proporcionaram às pesquisas em LC nos últimos anos.

Atrelado ao tamanho, Sardinha (2004) também propõe que a profundidade é um critério relevante na construção de um *corpus*. Se o conjunto de textos não apresentar profundidade, pouco relevante será sua grande extensão. Nesse sentido, a escolha dos textos que farão parte do conjunto deve seguir critérios rigorosos, quanto ao balanceamento dos textos (de gênero textual, por exemplo). Para tanto, o autor apresenta uma tipologia que tenta estabelecer certa profundidade aos *corpora* linguísticos, a saber: *modo* (falado ou escrito), *tempo* (sincrônico, diacrônico, contemporâneo ou histórico), *seleção* (de amostragem, monitor, dinâmico, estático ou equilibrado), *conteúdo* (especializado, regional/dialetal ou multilíngue), *autoria* (de aprendiz ou de língua nativa), *disposição interna* (paralelo ou alinhado) e *finalidade* (de estudo, de referência ou de treinamento/teste).

Outro critério importante é a representatividade. Biber (2012) aponta que o conjunto de textos deve demonstrar o fenômeno estudado em seu ambiente natural de ocorrência, dadas suas especificidades. O autor ressalta que esse critério deve ser anterior ao planejamento do

corpus, reconhecendo "os parâmetros situacionais que variam entre os textos de uma comunidade discursiva e também os tipos de características linguísticas que serão examinadas no *corpus*" (BIBER, 2012, p.11). Conceber a representatividade dessa maneira influencia diretamente a construção do *corpus*, que é organizada, segundo o autor, em duas etapas: "o planejamento original baseado em análises teóricas e de estudo-piloto de uma coleta de textos, por investigações empíricas detalhadas da variação linguística, e por uma revisão do planejamento" (BIBER, 2012, p.11).

Além dos critérios advindos da própria natureza e do conceito de *corpus*, também se derivam critérios específicos à pesquisa a que ele subsidia. Di-Felippo e Souza (2012) indicam que há decisões de projeto que devem fazer parte dos critérios de construção, pautando-se na utilização de recursos linguístico-computacionais em tarefas de estruturação do conjunto de textos e métodos semiautomáticos de extração de conhecimento. Ademais, os autores salientam que os critérios de (i) definição do objeto *corpus*, (ii) seleção do tipo de recurso linguístico a ser construído e (iii) decisões de projeto contribuem para um *design* do *corpus* adequado à pesquisa.

Um dos produtos que resulta da construção de um *corpus* é a anotação linguística. Essa tarefa pode ser definida como o processo de enriquecimento do *corpus*, adicionando (manual ou automaticamente) informações linguísticas, com objetivos teóricos (acolher uma teoria linguística, por exemplo) ou práticos (treinar um etiquetador morfológico, por exemplo). Embora o *corpus*, por si só, já seja um recurso bastante importante, quando anotado, torna-se caro à pesquisa que o desenvolveu, bem como a outras que posteriormente dele podem se beneficiar. Isso ocorre porque as anotações acrescentam valor ao *corpus*, permitindo que sejam realizados buscas e processamentos mais refinados (PEDRO; VALE, 2018).

Hovy e Lavid (2010) defendem que a anotação evidencia a perspectiva sobre a língua e a teoria linguística adotadas no estudo, como quais são as unidades de análise que são anotadas na RST (proposições) e na CST (sentenças, palavras ou porções textuais), por exemplo. Os autores equacionam metodologicamente a anotação em oito tarefas, a saber: (i) selecionar textos que sejam representativos para um *corpus* de treinamento, (ii) selecionar a teoria ou o conceito linguístico que subsidie um conjunto de etiquetas que será aplicado na tarefa, (iii)

anotar um pequeno fragmento do *corpus* de treinamento, (iv) medir comparativamente a concordância entre os anotadores, (v) decidir qual o nível de concordância será adotado no trabalho e, caso, não seja satisfatória, voltar a partir da etapa (ii) e fazer as adaptações necessárias, (vi) anotar uma maior parte do *corpus*, (vii) utilizar aprendizado de máquina a fim de, posteriormente, automatizar o processo de anotação e (viii) caso o desempenho dos algoritmos seja satisfatório, anotar automaticamente uma porção de textos ainda não anotados.

Mesmo após refinar o modelo teórico, é durante a anotação que se verificam certas incongruências no *corpus*, como a falta de representatividade do fenômeno estudado devido ao seu tamanho, por exemplo. Nesse contexto, Taboada e Das (2013) realizaram uma nova anotação sobre outra já feita: ao verificar que os marcadores discursivos usualmente utilizados para caracterizar e identificar as relações do modelo RST eram insuficientes, propuseram-se a estudar outras pistas linguístico-estruturais que pudessem ser consideradas como tal. A partir de então, os autores passaram a denominar tais pistas como *sinalizadores* das relações semânticas do modelo.

Construindo um paralelo aos pressupostos metodológicos de anotação de Hovy e Lavid (2010) em perspectiva à proposta de Taboada e Das (2013), adotou-se, neste trabalho, a estratégia de investigar quais os possíveis sinalizadores das relações de complementaridade informacional do modelo CST a partir da construção de um *corpus* de estudo. A seguir, têm-se as reflexões acerca dos sinalizadores que permitem remontar as relações de sentido entre unidades de análise dos modelos RST e CST.

3 Panorama dos sinalizadores de relações semânticas

O estudo de relações semânticas, em sua maioria, baseia-se no levantamento de conjuntos de pistas que possam caracterizar linguistica ou estruturalmente tais relações e, *a posteriori*, subsidiar a identificação (automática) de cada uma. Ao que se refere aos estudos que abordam direta ou indiretamente a complementaridade informacional, destacam-se Das e Taboada (2018), Taboada e Das (2013), para o modelo RST, e Maziero (2012), Souza (2015) e Souza e Di-Felippo (2018), para o modelo CST.

Comumente, a identificação (manual e automática) de relações RST é feita com base em marcadores discursivos. Em dado texto, ao

utilizar *se* como conjunção entre duas proposições, por exemplo, o autor planeja evidenciar uma *relação condicional* entre unidades discursivas. Dessa maneira, a conjunção marca a relação em questão.

Entretanto, Das e Taboada (2018) e Taboada e Das (2013) advogam que a ideia mais difundida na literatura sobre *marcador discursivo*, na verdade, limita a identificação das relações RST, pois se baseia apenas naquilo que está expresso no texto. Dado que as sentenças "Por conta de consumirem muita lactose, João e Pedro não podem ingerir leite" e "Os jovens continuam consumindo leite em sua versão 'sem lactose'" tenham sido extraídas do mesmo texto, percebe-se que o autor deseja evidenciar *contraste* entre as informações das duas proposições, porém, sem utilizar um marcador discursivo para tanto. Ademais, para que essa análise hipotética seja verdadeira, é preciso assumir que "os jovens" retomam anaforicamente "João e Pedro".

Assim, os autores propõem a hipótese de que, se a relação preterida pelo autor do texto é compreensível pelo leitor, a relação, então, é recuperável, ainda que o marcador não ocorra na superfície textual. Nesse sentido, a relação semântica deve apresentar algum sinalizador para que o leitor interprete o mais próximo possível a intenção do autor. No exemplo dado, um possível sinalizador seria a *anáfora lexical* entre as duas sentenças, além das proposições negativa ("não podem consumir leite") e afirmativa ("continuam consumindo leite") na primeira e na segunda sentenças, respectivamente

Taboada e Das (2013), revisando um *corpus* anotado com o modelo RST, propuseram um conjunto de sinalizadores que não foram previamente identificados como sinalizadores das relações previstas no modelo teórico, como é o caso das *anáforas lexicais*. Como resultado, apresentaram uma taxonomia de sinalizadores, incluindo os marcadores discursivos recorrentemente utilizados em estudos dessa natureza. Essa taxonomia organiza-se em nove categorias que superordenam outras subcategorias, a saber:

a) *Marcador discursivo:* sinalizadores que se caracterizam por serem marcas específicas de cada uma das relações RST; em geral, são tidos como expressões léxicas ou conjunções;

b) *Entidade*: sinalizadores que se caracterizam por estabelecerem similaridade ou dissimilaridade entre entidades nomeadas de unidades discursivas;

c) *Semântico:* sinalizadores que manifestam relações lexicais (hiperonímia, por exemplo) entre duas entidades de unidades discursivas distintas;

d) *Léxico:* sinalizadores que são traduzidos em palavras que indicam algum tipo de relação, como acrescentar uma informação;

e) *Morfológico:* sinalizadores que auxiliam a identificar fatores temporais por meio de desinências verbais;

f) *Sintático:* sinalizadores que indicam relações RST por meio de construções sintáticas específicas, como o discurso indireto;

g) *Gráfico:* sinalizadores que podem indicar relacionamento semântico por meio de pontuações, como as vírgulas em elipses;

h) *Numérico:* sinalizadores que evidenciam especificações entre unidades discursivas, detalhando ou ressaltando alguma informação apresentada genericamente (p.ex. "João, Pedro e Paulo foram acompanhar Maria no aeroporto" e "A garota estava acompanhada de seus três amigos no aeroporto");

i) *Gênero (textual):* sinalizadores que evidenciam marcas textuais específicas de cada gênero, como o informativo, em que as primeiras sentenças de um texto desse gênero terão informações genéricas, as quais serão elaboradas/detalhadas nas sentenças subsequentes.

Os autores concluíram que (i) há sinalizadores que ocorrem mais recorrentemente em determinadas relações (como é o caso de *gênero textual*, que ocorre mais em *Elaboration*); (ii) há aqueles que caracterizam certas relações somente sob combinações com outros (como é o caso de *construção frasal* que ocorre juntamente com *sintático* para caracterizar a relação *Background*); e (iii) há sinalizadores que, até então, não tinham sido explorados (como é o caso de *pontuação*).

Em um estudo mais recente, Das e Taboada (2018) refinaram a análise prévia, e organizam os sinalizadores em *singulares* e *combinados*. Os singulares são os mesmos apresentados no trabalho anterior, e os combinados são *referenciais* (ou anafóricos), *semânticos* e *gráficos* em conjunto com algum do tipo sintático, cada um deles.

Já acerca da identificação das relações do modelo CST, há os trabalhos de Maziero (2012), Souza (2015) e Souza e Di-Felippo (2018).

Visando à identificação automática das relações CST no contexto da Sumarização Automática Multidocumento, Maziero (2012) propõe uma série de métodos que se baseiam em sinalizadores explícitos na sentença. O autor parte do princípio de que as relações desse modelo semântico sempre compartilham informações entre duas unidades de análise, manifestando, portanto, redundância em maior ou menor grau (RADEV, 2000). Especificamente, ao analisar duas sentenças (S1 e S2) advindas de textos distintos, mas que versam sobre o mesmo assunto, o autor identifica as relações CST com base em (i) Diferença de tamanho em palavras (S1-S2), (ii) Porcentagem de palavras em comum em S1, (iii) Porcentagem de palavras em comum em S2, (iv) Posição de S1 no texto (início, meio ou fim), (v) Número de palavras na maior *substring* entre S1 e S2, (vi) Diferença no número de substantivos entre S1 e S2, (vii) Diferença no número de advérbios entre S1 e S2, (viii) Diferença no número de adjetivos entre S1 e S2, (ix) Diferença no número de verbos entre S1 e S2, (x) Diferença no número de nomes próprios entre S1 e S2, (xi) Diferença no número de numerais entre S1 e S2 e (xii) Sobreposição de sinônimos entre S1 e S2.

Além desses métodos, Maziero (2012) utilizou alguns específicos para a identificação das relações *Identity*, *Contradiction*, *Attribution*, *Indirect Speech* e *Translation*. O método formulado para a identificação da relação *Contradiction*, por exemplo, prevê apenas os casos de contradição do tipo explícita, isto é, resultantes de diferenças numéricas entre as sentenças de um par.

Para avaliar os métodos propostos, o autor utilizou o *corpus* CSTNews (CARDOSO *et al.,* 2011; DIAS *et al.,* 2014). Esse *corpus* se caracteriza por ser um conjunto multidocumento de textos jornalísticos em português, e está anotado com as relações do modelo CST. Trata-se de 50 *clusters* de notícias (em média, 3 textos) que possuem um mesmo assunto, sendo provenientes de fontes jornalísticas *online* distintas. No total, o CSTNews possui 140 textos, somando 2.088 sentenças e 47.240 palavras.

Com base nos métodos descritos, o Maziero (2012) desenvolveu algoritmos de AM, cuja precisão geral foi de 68,13%. Essa precisão é a média ponderada da precisão dos métodos para a identificação das relações *Overlap*, *Subsumption*, *Elaboration*, *Equivalence*, *Historical Background* e *Follow-up*, *Identity*, *Contradiction*, *Attribution*, *Indirect Speech* e *Translation*. Segundo o autor, essa precisão é considerada boa,

devido à subjetividade inerente à tarefa de identificação das relações multidocumento.

Especificamente sobre a identificação da complementaridade informacional via CST, há os trabalhos de Souza (2015) e Souza e Di-Felippo (2018). De acordo com os autores, tal fenômeno pode ser identificado com base em informações linguístico-estruturais, capturadas por pistas que evidenciam a complementação temporal entre as sentenças de um par. A análise foi traduzida em métodos de identificação (automática) dos tipos de complementaridade (temporal e atemporal) e das relações CST que a codificam, a saber: (i) distância entre as sentenças, (ii) sobreposição de nome/substantivo, (iii) ocorrência de advérbios temporais em S1, (iv) ocorrência de advérbios temporais em S2, (v) ocorrência de expressões temporais em S1, (vi) ocorrência de expressões temporais em S2, (vii) sobreposição de subtópicos, (viii) ocorrência de marcador discursivo em S1 e (ix) ocorrência de marcador discursivo em S2. Entretanto, esses estudos ainda se baseiam nos sinalizadores presentes na superfície textual das sentenças, além de, na maioria, se restringirem à presença e à ausência de informação temporal, como demonstrado em (2).

(2)

S1: A seleção brasileira de vôlei voltou a fazer bonito, desta vez na final da Liga Mundial, disputada contra a Rússia neste domingo no ginásio de Spodekna, em Katowice, na Polônia.

S2: Sua última derrota em finais da Liga Mundial, aliás, ocorreu em 2002, coincidentemente para a Rússia.

Em (2), narra-se sobre a participação da seleção brasileira na Liga Mundial de Vôlei. Na primeira sentença do par, aborda-se o desempenho do time da edição do evento daquele ano, disputado contra a Rússia, na Polônia; já a segunda informa a derrota do Brasil sobre a seleção russa, mas disputando a edição de 2002 do mesmo campeonato. Assim, S2 em relação a S1 apresenta uma informação complementar do tipo histórica sobre a participação do Brasil em uma edição anterior do evento esportivo.

Tendo em vista que os estudos realizados por Souza (2015) e Souza e Di-Felippo (2018) baseiam-se apenas em sinalizadores que

auxiliam a recuperação da informação complementar somente em traços presentes na superfície textual, propõe-se o refinamento da análise do fenômeno linguístico. Para tanto, baseando-se em Taboada e Das (2013) e Das e Taboada (2018) a fim de identificar sinalizadores que possam recuperar a complementaridade informacional, analisou-se o fenômeno em um *corpus* de estudo, construído a partir do CSTNews, considerando os princípios da LC. Ademais, como produto dessa análise, realizou-se a revisão manual da anotação da complementaridade nos pares de sentenças disponíveis no *corpus*, bem como a anotação manual dos sinalizadores de cada uma das relações que traduzem o fenômeno.

4 Análise da complementaridade no *corpus* cstnews

Para a realização deste estudo, selecionou-se o CSTNews (CARDOSO *et al.,* 2011; DIAS *et al.,* 2014). Como dito, o *corpus* está organizado em *clusters*, os quais representam as seções dos jornais *online* de onde os textos foram coletados, a saber "mundo", "política", "cotidiano", "ciência" e "esporte". Além dos textos-fonte (dois ou três), o *corpus* também contém sumários monodocumento e multidocumento de referência (manuais) e automáticos, alinhamento manual das sentenças dos sumários multidocumento às respectivas sentenças dos textos-fonte e uma série de camadas de anotações linguísticas. Dentre elas, estão: (i) anotação de relações semânticas multidocumento via CST; (ii) anotação de expressões temporais dos textos-fonte; (iii) etiquetação morfossintática (ou *tagging*); (iv) anotação dos sentidos dos substantivos e verbos; (v) anotação de aspectos informacionais nos sumários multidocumento (o quê, onde, quando, por exemplo), (vi) anotação automática dos textos-fonte via RST e (vii) anotação manual de subtópicos informativos em cada texto-fonte do *corpus*.

A anotação CST, em especial, foi realizada semiautomaticamente. Aleixo e Pardo (2008) revisaram o conjunto de rótulos das relações CST proposto para o inglês (ZHANG; GOLDENSHON; RADEV, 2002) e, a partir dessa revisão, decidiram aglutinar em um mesmo rótulo relações que apresentaram redundância entre si, e excluíram aquelas que não ocorreram nos textos do *corpus*.

Neste estudo, foram selecionados somente os pares de sentenças anotados com as relações que traduzem a complementaridade, a saber, *Historical Background*, *Follow-up* e *Elaboration*, representado por 73,

260 e 319 pares, respectivamente. Tendo em vista que no CSTNews há 713 pares de sentenças anotadas com as relações de complementaridade, construiu-se um *subcorpus* de estudo, que abrangeu os 10 primeiros *clusters* do *corpus*, resultando em 204 pares de sentença, sendo: (i) 12 pares anotados com a relação *Historical Background*, (ii) 94 com a relação *Follow-up* e (iii) 98 com *Elaboration*.

A respeito da complementaridade, Maziero (2012) e Maziero, Jorge e Pardo (2010) definem que o fenômeno ocorre pela relação que é estabelecida entre duas sentenças, S1 e S2, sendo cada uma delas provenientes de textos distintos; S2 deve apresentar a informação complementar em relação a algum elemento presente em S1. Admite-se ainda que as sentenças do par podem compartilhar conteúdo informacional, mas uma delas deve ter alguma informação aditiva que não esteja presente na outra.

Os autores compreendem a complementaridade em dois tipos. A do tipo temporal envolve sobreposição de conteúdo entre as sentenças de um par, sendo que S2 apresenta informação adicional baseada na informação temporal, a qual trata de um acontecimento anterior ou posterior ao evento principal descrito em S1. As relações *Historical-Background* e *Follow-up* traduzem esse tipo de complementaridade

Ainda segundo os autores, a complementaridade atemporal também se caracteriza pela sobreposição de conteúdo entre as sentenças de um par, sendo que uma das sentenças fornece informação adicional sobre o tópico principal. No entanto, o que a diferencia da complementaridade temporal é o fato de a informação adicional não ser de natureza temporal (MAZIERO, 2012; MAZIERO; JORGE; PARDO, 2010), e nem sempre ser marcada linguisticamente na superfície textual (SOUZA, 2015; SOUZA; DI-FELIPPO, 2018). A relação *Elaboration* compreende esse tipo de complementaridade.

4.1 Procedimentos metodológicos para levantamento de sinalizadores da complementaridade

Para levantar os sinais que evidenciam a complementaridade entre as sentenças de um par, optou-se por dois procedimentos metodológicos. Inicialmente, partiu-se do conjunto de sinais levantados por Souza (2015) e Souza e Di-Felippo (2018) e Taboada e Das (2013), assumindo que ambos os trabalhos, embora se ancorem em teorias linguísticas diferentes

(CST e RST), compartilham o pressuposto de que as relações, ora entre sentenças, ora entre proposições, são de natureza semântica.

O outro procedimento baseou-se na análise manual das sentenças já anotadas para identificar possíveis sinalizadores não previstos pelos trabalhos prévios, ou que não eram usualmente utilizados na descrição dos fenômenos multidocumento. Tal estudo consistiu em, dado um par de sentenças, (i) delimitar o trecho da sentença em que se manifestava a complementaridade e (ii) registrar os sinalizadores que auxiliam na recuperação da relação CST de complementaridade. Em (3) exemplifica-se tal procedimento.

(3)

S1: O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que 57 pessoas morreram e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo.

S2: <HB>A maior tragédia ferroviária da história do Egito ocorreu em fevereiro de 2002, após o incêndio de um trem que cobria o trajeto entre Cairo e Luxor, lotado de passageiros, e que deixou 376 mortos</HB>, segundo números oficiais.

Em (3) tem-se um par de sentenças que narra um acidente ferroviário que causou a morte de 57 pessoas, no Egito. A segunda sentença apresenta a informação histórica em relação à primeira, no trecho delimitado por "<HB>" e "</HB>"[1]. Após a identificação da informação complementar, observaram-se os possíveis traços que marcavam essa relação, tal como (a) *construção superlativa* (no caso, "a maior tragédia ferroviária da história do Egito"), (b) informação temporal marcada por uma *expressão temporal* (no caso, "em fevereiro de 2002"), (c) *discurso reportado* como estratégia sintática que compreende a informação complementar (d) *similaridade entre os eventos* narrados nas sentenças

[1] Durante a anotação foram utilizados delimitadores para identificar a informação complementar nos pares de sentenças e, posteriormente, em análises computacionais, dinamizar a recuperação automática dos trechos, já que as marcações foram feitas com base em XML. Assim, foram utilizadas as siglas HB, para *Historical Background*, FU, para *Follow-up* e ELAB, para *Elaboration*.

e (e) o *aspecto pontual* em S2, já que não se trata de um evento que se repete.

Dos traços levantados a partir da análise do exemplo em (3), (b) já havia sido identificado por Souza (2015) e Souza e Di-Felippo (2018) e (d), por Taboada e Das (2013), como marcas de fenômenos semânticos em suas respectivas teorias.

Ao se deparar com um novo sinalizador, verificava-se se ele já tinha ocorrido nos pares de sentenças anteriormente analisados, em quaisquer das relações de complementaridade no *corpus* de estudo. Especialmente para a relação *Historical Background*, esse procedimento estendeu-se a todos os pares anotados com esse rótulo devido à baixa ocorrência de pares de sentenças no CSTNews. Esse estudo durou cerca de oito meses.

Um aspecto relevante nesse procedimento metodológico é a identificação dos trechos das sentenças que continham a informação complementar. Até então, o CSTNews não apresentava essa delimitação. Tendo em vista que a identificação dos sinais das relações semânticas deve ser realizada a partir do processamento cognitivo, ou seja, mapeando as intenções do autor (DAS; TABOADA, 2018), ter os trechos delimitados auxilia esse tipo de análise, pois busca-se perceber o que motivou os anotadores a atribuírem dada relação às sentenças. Por conta disso, todos os trechos de complementaridade foram identificados previamente ao estudo.

4.2 Sinalizadores da complementaridade informacional multidocumento

Os sinalizadores apresentados nesta seção estão organizados em dois subgrupos: aqueles que discriminam as relações CST de complementaridade entre si, e aqueles que são capazes de auxiliar na recuperação da complementaridade, porém ocorrem em pelo menos duas das três relações.

4.2.1 Sinalizadores de Historical Background

a) *Expressões superlativas de comparação* – Esse tipo de construção frástica ocorre sempre em que há eventos relacionados por sucessivas repetições que "se superam". Em (4), a primeira sentença narra sobre uma indenização financeira que a Igreja

Católica americana pagou a vítimas de abuso sexual, enquanto a segunda narra sobre "*o maior pagamento* já feito pela Igreja Católica".

(4)

S1: A Igreja Católica chegou a um acordo financeiro estimado em US$ 660 milhões (aproximadamente R$ 1,2 bilhão) com mais de 500 pessoas que alegam ter sido vítimas de abuso sexual por padres em Los Angeles, nos Estados Unidos.

S2: <HB>Este seria o maior pagamento já feito pela Igreja desde que surgiu o escândalo de abuso sexual envolvendo religiosos em 2002 e elevaria o total de indenizações pago pela Igreja desde 1950, nos Estados Unidos, a US$ 2 bilhões (R$ 3,7 bilhões).</HB>

b) *Relação entre aspectos semânticos* – As sentenças de um par podem apresentar eventos de diferentes aspectos semânticos (pontuais ou habituais), como a queda de um avião em certa localidade e o fato desse evento estar relacionado à ocorrência habitual desse tipo de acidente naquele mesmo local. Em (5), em S1 há um aspecto pontual, já que o evento narrado ocorreu apenas uma vez, enquanto a informação veiculada em S2 é de aspecto habitual, recuperado pela expressão "*são frequentes*".

(5)

S1: Um acidente aéreo na localidade de Bukavu, no leste da República Democrática do Congo (RDC), matou 17 pessoas na quinta-feira à tarde, informou nesta sexta-feira um portavoz das Nações Unidas.

S2: <HB>Acidentes aéreos são frequentes no Congo,</HB> onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.

c) *Relação entre eventos similares não idênticos* – As sentenças de um par podem apresentar eventos semelhantes, mas jamais idênticos, tendo um intervalo temporal grande entre eles. Em (5), por exemplo, em que se narra sobre o acidente aéreo em Bukavu, em S1, e sobre a frequência de acidentes aéreos similares na região,

em S2, a distância temporal entre os eventos é considerável, tendo em vista que a informação veiculada em S2 ocorre após S1.

4.2.2 Sinalizadores de Follow-up

a) *Posterioridade entre eventos* – Há eventos que ocorrem em sucessão e são separados por um pequeno intervalo temporal. Em (6), S1 narra sobre uma jogada futebolística que ocorre "aos 27 minutos" do jogo, enquanto S2 apresenta outra jogada que acontece em seguida, "aos 31" minutos.

(6)

S1: Aos 27min, Kaká arrancou e chutou de fora da área.

S2: <FU>Kaká acertou um belíssimo chute de longe no ângulo aos 31 e fez 3 a 0.</FU>

b) *Previsão de eventos* – Nos pares de sentenças em que se notou este sinalizador, apresentaram-se eventos que ocorreram sequencialmente por meio da relação que se estabelece entre os tempos verbais das sentenças, de maneira que em S1 têm-se verbos no presente ("Lula *tem*...") e em S2, verbos flexionados no futuro do pretérito ("O presidente *teria...* "), como demonstrado em (7). Percebeu-se que esse tipo de sinalizador ocorre em *clusters* cujos assuntos principais são política, desastres naturais e esporte, já que é relevante informar sobre a possibilidade de um evento futuro.

(7)

S1: De acordo com a pesquisa, Lula (PT) tem 44% das intenções de voto, contra 25% de Geraldo Alckmin (PSDB) e 11% de Heloísa Helena (PSOL).

S2: <FU>O presidente teria 53% das intenções de voto contra 30% de Heloísa.</FU>

c) *Efetivação de evento projetado* – Há casos em que se identificou a relação entre as sentenças do par por meio de previsões/possibilidades que se realizam/concretizam, como a possível

indicação de Solange Vieira a um cargo comissionado, em S1, e a efetivação dessa indicação, em S2, como demonstra-se em (8). Tal como a *Relação hipotética entre eventos*, este tipo de sinalizador é comum em *clusters* que abordam textos sobre política e esporte.

(8)

S1: O ministro da Defesa, Nelson Jobim, deve encaminhar o nome da economista Solange Vieira para assumir uma das diretorias da Agência Nacional de Aviação Civil (Anac).

S2: <FU>O ministro da Defesa, Nelson Jobim, informou no fim da noite desta terça-feira que a economista Solange Vieira, de 38 anos, será a nova presidente da Agência Nacional de Aviação Civil (Anac).</FU>

d) *Prolongamento do mesmo evento* – Há casos em que o evento descrito na segunda sentença é apenas a extensão do mesmo narrado na primeira. Em (9), narra-se sobre o acidente aéreo ocorrido no aeroporto de Congonhas; na primeira sentença, apresentam-se o plano de voo e alguns detalhes sobre o acidente, enquanto na segunda, a informação complementar à primeira centra-se em continuar narrando sobre o plano de voo e a possível causa do acidente.

(9)

S1: Um dia antes do acidente, na segunda-feira, 16, o avião também teria apresentado problemas ao aterrissar em Congonhas, durante o voo 3215, procedente de Belo Horizonte (Confins), só conseguindo parar muito próximo do final da pista.

S2: O problema teria sido detectado pelo sistema eletrônico de checagem do próprio avião, <FU>e ainda assim a aeronave da TAM, um Airbus A320, continuou voando, com o reverso direito desligado.</FU>

4.2.3 Sinalizadores de Elaboration

a) *Foco argumentativo distinto* – Há casos no *corpus* em que as sentenças do par apresentam focos argumentativos diferentes. O par de sentenças em (10) narra sobre uma reforma ocorrida em uma das pistas no aeroporto de Congonhas; a primeira sentença aborda o fato de não haver atraso nos voos internacionais, enquanto a segunda aponta que ocorreram atrasos entre partidas e chegadas de voos. Nesses casos, as informações não foram tidas como contraditórias, já que a relação não se constrói sob o questionamento das informações propositivas, mas sob apontamentos narrados por algum agente na primeira sentença.

(10)

S1: Nenhuma partida ou chegada internacional, segundo os painéis da Infraero, estavam fora do horário, o que não ocorria com os voos domésticos.

S2: <ELAB>As informações da Infraero não batem com as do painel das companhias aéreas, são 20 partidas atrasadas e 24 pousos atrasados.</ELAB>

b) *Informação adicional* – Há casos em que a complementaridade ocorre sob a narração de uma informação adicional na segunda sentença não prevista na primeira. O par de sentenças em (11) narra sobre a indicação nominal para ocupar o cargo de diretor da Agência Nacional de Aviação Civil; na primeira sentença, apresenta-se o fato de indicar uma economista (no caso, Solange Vieira – informação recuperável apenas a partir da leitura dos textos-fonte do *cluster*), enquanto a complementaridade na segunda se dá pela adição da informação da duração do mandato do cargo.

(11)

S1: Mas, diante da dificuldade para encontrar pessoas que aceitassem assumir uma das diretorias da agência reguladora, após a renúncia de três diretores, Jobim decidiu indicar a economista para o cargo.

S2: <ELAB>Como os diretores de agências têm mandato de cinco anos</ELAB>, só podem sair por renúncia, decisão judicial ou acusação de improbidade administrativa.

4.2.4 Sinalizadores não discriminativos de relações CST de complementaridade

Como dito, há sinalizadores que auxiliam na recuperação da informação complementar entre as sentenças de um par, mas que pela ocorrência em mais de uma relação CST não foi possível classificá-los como discriminativos. Por conta do espaço dispensado neste texto, escolheu-se explanar a *categoria* dos sinalizadores presentes na Tabela 1. No entanto, ressalta-se que os sinalizadores *genéricos* e *específicos* estão detalhados em Souza (2019).

a) *Sinalizador do tipo Estrutural* – Este tipo de sinalizador permite que se possa recuperar a informação complementar somente a partir da leitura prévia dos textos que compõem o *cluster*, pois somente após esse procedimento, identificando as sentenças do par em seus respectivos textos-fonte que é possível constatar a complementaridade.

b) *Sinalizadores do tipo Referenciação* – Este tipo de sinalizador auxilia a identificação da informação complementar por meio da recuperação de algum referente na primeira sentença, em relação à segunda do par. Identificou-se que essa referenciação se deu por meio de Anáforas nominal e lexical.

c) *Sinalizadores do tipo Morfológico* – Estes sinalizadores recuperam a informação complementar por meio de marcas de tempo nos verbos (como "estavam na rua", em S1, e "está fazendo a vistoria", em S2), expressões nominais (como "novo bombardeio"), verbos de elocução (como "disseram" e "comunicaram") e diferenças numéricas que estabelecem adição de informação.

d) *Sinalizadores do tipo Sintático* – Estes sinalizadores identificam a complementaridade com base em adjuntos adverbiais ("como também", por exemplo), discurso reportado em que se marca a fonte da informação (como "os rebeldes afirmaram que..."), deslocamento de tema-rema entre as sentenças do par e orações aditivas, explicativas, objetivas direta e orações reduzidas.

e) *Sinalizadores do tipo Semântico* – Este tipo de sinalizador recupera a informação complementar por meio de palavras do mesmo campo semântico (como a relação que há entre "ataques", em S1, e "ameaça", "bombardeio" e "guerra", em S2), relações semânticas de causa-efeito, hiponímia e parte-todo, expressões temporais (como "Olimpíadas de Pequim") e itens lexicais que denotam acréscimo (como "acrescentou").

f) *Sinalizadores do tipo Pragmático* – Por fim, estes sinalizadores auxiliam na recuperação da complementaridade por meio de detalhamento ou conhecimento de mundo (como a relação que há entre "Companhia de Engenharia de Tráfego" e "São Paulo").

4.3 Organização tipológica dos sinalizadores de complementaridade

Com base no estudo realizado, foi possível organizar tipologicamente os sinalizadores descritos. A categorização foi feita posteriormente à análise, após a contabilização da ocorrência de cada um. Observou-se que havia regularidade entre os sinalizadores, permitindo propor categorias, que os organizassem em *genéricos* e *específicos*, resultando na organização demonstrada na Tabela 1.

Como já demonstrado, há sinalizadores específicos de cada relação CST de complementaridade, os quais desempenham papel essencial na caracterização das relações. Ademais, há sinalizadores que auxiliam na recuperação da complementaridade informacional, mas não são capazes de discriminar as relações entre si.

Com relação aos sinalizadores não específicos, observou-se que eles ocorreram ora entre dois tipos de relação CST (temporal e atemporal), ora entre duas relações do mesmo tipo. No primeiro caso, é possível cogitar que as fronteiras entre as relações não estavam bem claras para os anotadores do *corpus*, o que pode ser resultado da junção de rótulos que aconteceu para a adaptação do modelo CST para o português. O segundo caso indica que o sinalizador diferencia o tipo, mas não a relação, dando indícios que a descrição deve ser aprimorada, observando a correlação entre os sinalizadores para caracterização do tipo e da relação CST.

TABELA 1 – Tipologia de sinalizadores de complementaridade no *corpus* de estudo no CSTNews

Tipologia			Relação CST de complementaridade			Total
			Elaboration	*Follow-up*	*Historical Background*	
Categoria	Sinal genérico	Sinal Específico				
Referenciação	**Anáfora**	Anáfora Associativa	50	31	0	81
		Anáfora Nominal	132	89	20	241
--------	**Estrutural**	Leitura do Cluster	48	60	14	122
Morfológico	**Classe de palavras**	Numeral	11	35	2	48
		Expressão Nominal	2	15	0	17
		Expressão Preposicional	7	0	17	24
	Temporal	Tempo verbal	12	134	3	149
	Verbos de elocução	Verbos de elocução	26	51	0	77
Sintático	**Período simples**	Adjunto adverbial	31	40	2	73
		Expressão superlativa	0	0	26	26
	Período composto	Discurso reportado	67	52	0	119
		Oração aditiva	26	2	0	28
		Oração explicativa	37	5	7	49
		Oração objetiva direta	22	7	0	29
		Oração reduzida	12	3	0	15
	Deslocamento	Tema-Rema	108	1	2	111
Semântico	**Campo semântico**	Campo semântico	29	34	0	63
	Relações semânticas	Causa-Efeito	12	23	0	35
		Hiponímia	16	4	0	20
		Parte-Todo	42	15	0	57
	Temporal	Expressão temporal	4	109	57	170
	Sentido de acréscimo	Semântica Lexical	27	42	8	77
Pragmático	**Sobre o evento**	Detalhe	103	59	0	162
		Posterior	0	92	0	92
		Previsão	0	17	0	17
		Prolongamento	0	57	0	57
		Projeção	0	18	0	18
		Similaridade	0	0	39	39
	Argumentação	Foco argumentativo	17	0	0	17
	Suplementação	Informação adicional	52	0	0	52
	Aspectualidade	Fato pontual	0	0	38	38
		Frequência	0	0	38	38
	Conhecimento adicional	Conhecimento de Mundo	5	28	14	47

Fonte: Elaboração própria.

Com relação à ocorrência dos sinalizadores *genéricos* percebe-se que os sinalizadores do tipo *pragmático* são mais frequentes no *corpus* (577 ocorrências), seguido dos tipos *sintático* (450), *semântico* (422), *anafórico* (322), *morfológico* (315) e *estrutural* (122). Até então, os sinalizadores identificados por Souza (2015) estavam restritos à distinção do tipo de complementaridade, a saber, temporal e atemporal, compreendidos apenas nos tipos *morfológico* (tempo verbal) e *semântico* (expressão temporal).

Entretanto, após a análise do *corpus* de estudo, concluiu-se, como previsto, que a complementaridade informacional via modelo CST é mais bem compreendida por meio de sinalizadores do tipo *pragmático*. Alguns desses sinalizadores, como demonstrado, são capazes de caracterizar cada uma das relações CST de complementaridade, bem como auxiliar na recuperação da informação adicional entre as duas sentenças de um par. A não consideração desse tipo de sinalizador, é uma possível causa para que houvesse equívocos quanto à distinção das relações *Follow-up* e *Elaboration* em Souza (2015).

Outra conclusão possível de se apontar é uma possível reconsideração sobre a classificação da complementaridade proposta por Maziero (2012) e Maziero, Jorge e Pardo (2010). Os autores distinguem os tipos de complementaridade entre temporal e atemporal, considerando a presença ou a ausência de informação adicional baseada em aspectos temporais. No entanto, como observado na Tabela 1, a informação temporal é capturada somente por dois de sinalizadores específicos. Nesse sentido; todos os outros sinalizadores deveriam recuperar a complementaridade atemporal (logo, a relação *Elaboration*). No entanto, isso é inverdade, como é possível concluir, pois os do tipo *pragmático*, por exemplo, ocorrem mais frequentemente nas relações do tipo temporal. Assim, é possível que, futuramente, um novo estudo sobre a complementaridade informacional resulte em uma nova classificação do fenômeno via modelo CST, a qual não seja baseada no aspecto temporal presente ou ausente nas sentenças de um par, mas na informação pragmática veiculada pelas sentenças.

5 Considerações finais

Neste trabalho, aprofundou-se a descrição do fenômeno da complementaridade que ocorre em conjuntos de textos jornalísticos

que abordam um mesmo assunto. Especificamente, com base em estudo de *corpus*, estendeu-se a descrição já realizada por Souza (2015) que, ao se basear apenas em atributos restritos a informações temporais nas sentenças de um par, já havia obtido resultados bastantes satisfatórios.

Diferentemente das relações RST que são propositais, no modelo CST a informação complementar não intencional por parte dos autores dos textos, mas ocasionada a partir da anotação semântica das relações previstas no modelo. Nesse sentido, mais que determinar quais sinalizadores auxiliam na recuperação da complementaridade, enquanto fenômeno linguístico, eles delimitam o ponto de vista dos anotadores do *corpus* CSTNews. Embora essa consideração seja irrefutável, cabe destacar que a concordância medida entre os anotadores proporciona margem de confiança nos dados, já que seguiram uma metodologia de anotação semelhante à proposta por Hovy e Lavid (2010). Essa questão pode ser confirmada quanto à regularidade da ocorrência dos sinalizadores nas relações CST de complementaridade, o que determina que são características ora do fenômeno, ora das relações que o traduzem.

Com relação à frequência, percebem-se sinalizadores que tiveram baixa ocorrência nos pares de sentença. Em termos de descrição linguística, eles podem ser essenciais para a compreensão da manifestação da complementaridade em determinados contextos, como é o caso de *expressões nominais*, em *Follow-up*. Em termos de aplicação computacional, é possível que sinalizadores como esse sejam descartados das análises, por não serem considerados bastante robustos em relação aos outros.

Ao que se refere ao gênero textual, não é possível ser assertivo quanto à caracterização da complementaridade em relação a esse fator. Como descrito, o *corpus* de estudo é composto apenas por textos jornalísticos e, preliminarmente, tende-se a apontar que os sinalizadores são típicos desse gênero. Entretanto, será necessário, em trabalhos futuros, construir-se um *corpus* multidocumento de treinamento com variação de gênero textual, a fim de verificar a ocorrência dos sinalizadores propostos neste trabalho anotando novos textos. Isso permitirá determinar se os sinalizadores são restritos ao gênero textual, como também corroborar se são característicos do fenômeno da complementaridade.

Como proposto por Hovy e Lavid (2010), um dos próximos passos que deve ser seguido na anotação de *corpus* é a proposição de métodos automáticos para essa tarefa. Entretanto, é necessário destacar que o

estado da arte ainda não permite que os mesmos resultados apresentados aqui sejam alcançados, por serem derivados da análise manual do *corpus* de estudo. Além disso, os *parsers* disponíveis atualmente em PLN podem ser capazes de identificar os sinalizadores dos tipos *sintático*, *morfológico*, *referenciação*; entretanto, ainda não processam com alto desempenho os sinalizadores específicos dos tipos *pragmático* e *semântico*, os quais são mais proeminentes na complementaridade, além de serem discriminativos das relações CST entre si.

Por fim, como parte dos trabalhos futuros, pretende-se observar a ocorrência combinada dos sinalizadores de complementaridade com abordagens de AM. Além disso, pretende-se verificar se os sinalizadores aqui delimitados também têm potencial para subsidiar o aprimoramento da descrição de outras relações CST, como aquelas que traduzem a redundância e a contradição.

Agradecimentos

Em tempos em que a ciência é atacada, seus investimentos são cada vez mais limitados e professores são desvalorizados, é importante destacar o auxílio financeiro empenhado nesta pesquisa pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e a orientação atenciosa da Profa. Dra. Ariani Di Felippo ao projeto de doutoramento, do qual se deriva este artigo: certamente o destino deste estudo poderia ter sido outro sem o financiamento e a dedicação de minha orientadora. Muito obrigado!

Referências

ALEIXO, P.; PARDO, T. A. S. CSTNews: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (Cross-document Structure Theory). São Carlos: USP; UFSCar; UNESP, 2008. (Série Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional - NILC)

BIBER, D. Representatividade em planejamento de *corpus*. Tradução de Paula Marcolin. *Cadernos de Tradução*, Porto Alegre, v. 1, n. 30, p. 11-45, 2012.

CARDOSO, P. C. F.; MAZIERO, E. G.; JORGE, M. L. C.; SENO, E. M. R.; DI-FELIPPO, A.; RINO, L. H. M.; NUNES, M. G. V.; PARDO, T. A. S. CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. *In:* RST BRAZILIAN MEETING, 3rd., 2011, Cuiabá. *Proceedings* […]. Cuiabá: SBC, 2011. p. 88-105.

CASELI, H. M. O uso de corpora paralelos para a criação de um tradutor automático estatístico. *In:* VIANA, V.; TAGNIN, S. E. O. *Corpora na Tradução*. São Paulo: HUB Editorial, 2015. p. 243-267.

DAS, D.; TABOADA, M. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, [*S.l.*], v. 52, n. 1, p. 149-184, 2018. DOI: https://doi.org/10.1007/s10579-017-9383-x

DAS, D.; TABOADA, M.; MCFETRIDGE, P. *RST Signalling Corpus*. Philadelphia: Linguistic Data Consortium, 2015.

DAVIES, M.; KIM, J. The Advantages and Challenges of 'Big Data': Insights from the 14 Billion Word iWeb Corpus. *Linguistic Research*, [*S.l.*], v. 36, p. 1-34, 2019. DOI: https://doi.org/10.17250/khisli.36.1.201903.001

DIAS, M. S.; GARAY, A. Y. B.; CHUMAN, C.; BARROS, C. D.; MAZIERO, E. G.; NOBREGA, F. A. A.; SOUZA, J. W. C.; CABEZUDO, M. A. S.; DELEGE, M.; JORGE, M. L. R. C.; SILVA, N. L.; CARDOSO, P. C. F.; BALAGE FILHO, P. P.; CONDORI, R. E. L.; MARCASSO, V.; DI-FELIPPO, A.; NUNES, M. D. G. V.; PARDO, T. A. S. Enriquecendo o *corpus* CSTNews: a criação de novos sumários multidocumento. *In*: INTERNATIONAL CONFERENCE ON COMPUTATIONAL PROCESSING OF THE PORTUGUESE LANGUAGE – PROPOR, 2014, São Carlos. *Proceedings*… São Carlos: SBC, 2014. p. 239-243.

DI-FELIPPO, A.; SOUZA, J. W. C. O projeto do *corpus* para a construção de uma wordnet terminológica. In: PINTO, M. V.; SHEPHERD, T. M. G.; SARDINHA, T. B. (org.). *Caminhos da Linguística de Corpus*. Campinas: Mercado de Letras, 2012. p. 225-245.

HALL, M. *et al*. The WEKA Data Mining Software: An Update. *ACM SIGKDD Explorations Newsletter*, [*S.l.*], v. 11, n. 1, p. 10-18, 2009. doi: https://doi.org/10.1145/1656274.1656278

HOVY, E.; LAVID, J. Towards a 'Science' of Corpus Annotation: A New Methodological Challenge for Corpus Linguistics. *International Journal of Translation*, [*S.l.*], v. 22, n. 1, p. 13-36, 2010.

KUHN, D.; ABARCA, E.; NUNES, M. G. Corpus Nilc de português escrito no Brasil. São Carlos: São Carlos: USP; UFSCar; UNESP, 2000. (Série Relatórios Técnicos do Núcleo Interinstitucional de Linguística Computacional - NILC)

MANN, W. C.; THOMPSON, S. A. *Rhetorical Structure Theory*: A theory of Text Organization. Marina del Rey, CA: University of Southern California, Information Sciences Institute, 1987.

MAZIERO, E. G. *Identificação automática de relações multidocumento*. 2012. 118f. Tese (Doutorado em Ciências da Computação) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Paulo, 2012.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying Multi-Document Relations. *In:* INTERNATIONAL WORKSHOP ON NATURAL LANGUAGE PROCESSING AND COGNITIVE SCIENCE, 2010, Funchal. *Proceedings* [...]. Funchal: Polytechnic Institute of Setúbal, 2010. p. 60-69.

PEDRO, W. G.; VALE, O. A. ComentCorpus: o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um corpus opinativo. *In:* FINATTO, M. J. B.; REBECHI, T.; SARMENTO, S; BOCORNY, A.E.P. (org.). *Linguística de corpus*: perspectivas. Porto Alegre: Instituto de Letras da Universidade Federal do Rio Grande do Sul, 2018. p. 19-40.

RADEV, D. R. A Common Theory of Information Fusion from Multiple Text Sources Step One: Cross-Document Structure. *In:* SIGDIAL WORKSHOP ON DISCOURSE AND DIALOGUE, 1ST., 2000, Hong Kong. *Proceedings*... Hong Kong: Association for Computational Linguistics, 2000. p. 74-83. DOI: https://doi.org/10.3115/1117736.1117745

RODRIGUES, R. *Contribuições para um léxico-gramática das construções locativas do espanhol*. 2019. 174f. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2019.

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

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

SOUZA, J. W. C. *Aprofundamento da caracterização linguístico-computacional da complementaridade em um corpus jornalístico multidocumento*. 2019. 117f. Tese (Doutorado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2019.

SOUZA, J. W. C. *Descrição linguística da complementaridade para a sumarização automática multidocumento*. 2015. 105f. Dissertação (Mestrado em Linguística) – Programa de Pós-Graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2015.

SOUZA, J. W. C.; DI FELIPPO, A. Caracterização linguística da complementaridade: subsídios para Sumarização Automática Multidocumento. *ALFA: Revista de Linguística*, São Paulo, v. 62, n.1, p. 125-150, 2018. DOI: https://doi.org/10.1590/1981-5794-1804-6

TABOADA, M.; DAS, D. Annotation upon Annotation: Adding Signalling Information to a *Corpus* of Discourse Relations. *Dialogue Discourse*, [*S.l.*], v. 4, n. 2, p. 249-281, 2013. DOI: https://doi.org/10.5087/dad.2013.211

TAGNIN, S. E a Linguística de Corpus vai desbravando novos horizontes. *In*: FINATTO, M. J. B.; REBECHI, T.; SARMENTO, S; BOCORNY, A. E. P. (org.). *Linguística de corpus*: perspectivas. Porto Alegre: Instituto de Letras da Universidade Federal do Rio Grande do Sul, 2018. p. 11-15.

TAGNIN, S. E. O.; BEVILACQUA, C. Corpora *na Terminologia*. São Paulo: HUB Editorial, 2015.

VIANA, V.; TAGNIN, S. E. O. Corpora *no ensino de línguas estrangeiras*. São Paulo: Hub Editorial, 2011.

ZHANG, Z.; GOLDENSHON, S. B.; RADEV, D. R. Towards CST-Enhanced Sumarization. *In:* NATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE (AAAI-2002), 18th., 2002, Edmonton. *Proceedings* […]. Edmonton: AAAI, 2002. p. 439-445.