Revista Trem de Letras Alfenas, MG V. 5 n.1 1-16 e018009 2018

ANÁLISE DE CONJUNÇÕES E LOCUÇÕES CONJUNTIVAS DE ADIÇÃO COMO SINALIZADORES DE COMPLEMENTARIDADE INFORMACIONAL

Jackson Wilke da Cruz Souza (Unifal-MG) [1]

Resumo: Com o objetivo de fomentar subsídios teóricos à identificação da complementaridade informacional segundo o modelo teórico CST, neste trabalho, verificou-se a ocorrência de conjunções e locuções conjuntivas de adição. Para tanto, a partir do corpus CSTNews, foram selecionados pares de sentenças previamente anotadas com as relações Follow-up, Historical background e Elaboration, as quais traduzem o fenômeno da complementaridade. Em seguida, os pares de sentenças foram submetidos ao software AntConc, o qual permitiu a construção de listas de colocados. Os resultados demonstram que as conjunções e locuções conjuntivas são sinalizadores em potencial de complementaridade informacional e, mais especificamente, das relações Follow-up e Elaboration.

Palavras-chave: Complementaridade. Processamento de Línguas Naturais. Linguística de corpus.

Abstract: With the aim of promoting theoretical subsidies for the identification of informational complementarity according to the CST theoretical model, in this paper, we verified the occurrence of conjunctions and conjunctive phrases of addition. For that, from the CSTNews corpus, we selected pairs of sentences previously annotated with the Follow-up, Historical background, and Elaboration relations, which translate the phenomenon of complementarity. Then, we submitted the pairs of sentences to the AntConc software, which allowed the construction of placed lists. The results demonstrate that conjunctions and conjunctive phrases are potential indicators of informational complementarity and, more specifically, of Follow-up and Elaboration relationships.

Keywords: Complementarity. Natural Language Processing. Corpus linguistics.

Resumen: Con el objetivo de promover subsidios teóricos para la identificación de complementariedad informacional según el modelo teórico CST, en este trabajo se verificó la ocurrencia de conjunciones y frases conjuntivas de suma. Para eso, del corpus de CSTNews, se seleccionaron pares de oraciones previamente anotadas con las relaciones de Seguimiento, Antecedentes Históricos y Elaboración, que traducen el fenómeno de la complementariedad. Luego, los pares de oraciones se sometieron al software AntConc, lo que permitió la construcción de listas colocadas. Los resultados demuestran que las conjunciones y frases conjuntivas son indicadores potenciales de complementariedad informativa y, más específicamente, de relaciones de Seguimiento y Elaboración.

Palabras clave: Complementariedad. Procesamiento natural del lenguaje. Lenguaje del cuerpo.

[1] e-mail: jackcruzsouza@gmail.com

1. Introdução

O acesso e disponibilização da informação estão cada vez mais fáceis e abundantes. De acordo com as projeções de Taufer (2013) para o ano de 2020, chegaria a ser produzido 40 *zetabytes* de informação e disponibilizado na *Web*. Isso faz com que o usuário tenha a sua disposição uma fonte quase que inesgotável de conhecimento.

Atrelado a isso, muitas fontes informativas, como jornais *online*, podem publicar acerca dos mesmos eventos numa velocidade e alcance muito grandes. Isso faz com que os textos, por vezes, possam ser redundantes (similares), contraditórios (no caso de atualizações, por exemplo) ou complementares uns aos outros. Assim, caracterizam-se os fenômenos multidocumentos.

Algumas áreas do Processamento de Línguas Naturais (PLN) têm interesse de processar esse tipo de informação e identificar (semi)automaticamente tais fenômenos, como sistemas de pergunta-resposta, tradução e sumarização. Para o Português do Brasil (PB) há pesquisas que descrevem linguisticamente a contradição (SILVA; DI-FELIPPO, 2014), a redundância (SOUZA *et al.*, 2012) e a complementaridade (SOUZA, 2015).

Com relação à complementaridade, SOUZA (2015) levantou um conjunto de atributos linguístico-estruturais para identificar automaticamente a complementaridade. De acordo com o autor, tal fenômeno pode ser identificado com base em informações linguístico-estruturais, sobretudo na presença ou ausência de atributos que evidenciem a complementação temporal entre as sentenças de um par. O objetivo do autor foi desenvolver métodos específicos de identificação automática dos tipos de complementaridade (temporal e atemporal) e as relações que os codificam (*Historical background*, *Follow-up* e *Elaboration*). Para tanto, foram desenvolvidos algoritmos de Aprendizado de Máquina (AM) com base nos atributos linguístico-

estruturais identificados no estudo. A análise dos resultados das medidas de avaliação dos classificadores (a saber, precisão, cobertura e medida-f) permitiu concluir que os classificadores apresentam desempenho superior em distinguir a relação *Historical background* de *Elaboration*. A identificação da relação *Follow-up* ainda apresentava equívocos, a ponto de os classificadores confundirem-na com a relação *Elaboration*.

Entretanto, algo ainda não evidenciado no trabalho foi a investigação de *conjunções aditivas* como sinalizador de caracterização da informação complementar. Dessa forma, neste trabalho investiga-se o comportamento sintático das conjunções aditivas quanto característica da informação complementar entre pares de sentenças que foram extraídas de textos-fonte distintos, mas que versam sobre o mesmo assunto. Para tanto, parte-se do princípio de que o valor semântico das conjunções pode sinalizar a complementaridade entre sentenças de um par.

Propõe-se, então, um estudo preliminar e exploratório de *corpus*, observando listas de colocados às conjunções aditivas em pares de sentenças anotados com as relações CST de complementaridade. Para tanto, foi utilizado o *software* de análise de *corpora* textuais AntConc (ANTHONY, 2014). Nele foi possível levantar listas de colocados e realizar observações quanto à ocorrência da informação complementar.

Este artigo está organizado em cinco seções, além desta introdução. Na Seção 2, apresentam-se o fenômeno da complementaridade segundo o modelo teórico *Cross-document Structure Theory* (CST) (RADEV, 2000). Na Seção, 3 apresenta-se a metodologia deste estudo, que consistiu na construção do *subcorpus* de análise e informações sobre o procedimento de análise com a ferramenta AntConc. Na Seção 4, demonstram-se os resultados deste trabalho a partir da análise pertinente à Linguística de *corpus*, bem como as discussões cabíveis a partir dela. Por fim, na Seção 5, tecem-se as considerações finais deste estudo, além de salientar as limitações e trabalhos futuros.

2. A complementaridade via modelo teórico CST

Em subáreas do PLN, em especial a Sumarização Automática, é bastante comum a utilização do modelo teórico CST (RADEV, 2000). A CST compreende um conjunto de relações semânticas que capturam fenômenos linguísticos que provêm da análise multidocumento, como a redundância, contradição e complementaridade. Como resultado, esse modelo visa rotular com relações específicas desses três fenômenos, em pares, unidades intertextuais (como as sentenças, por exemplo).

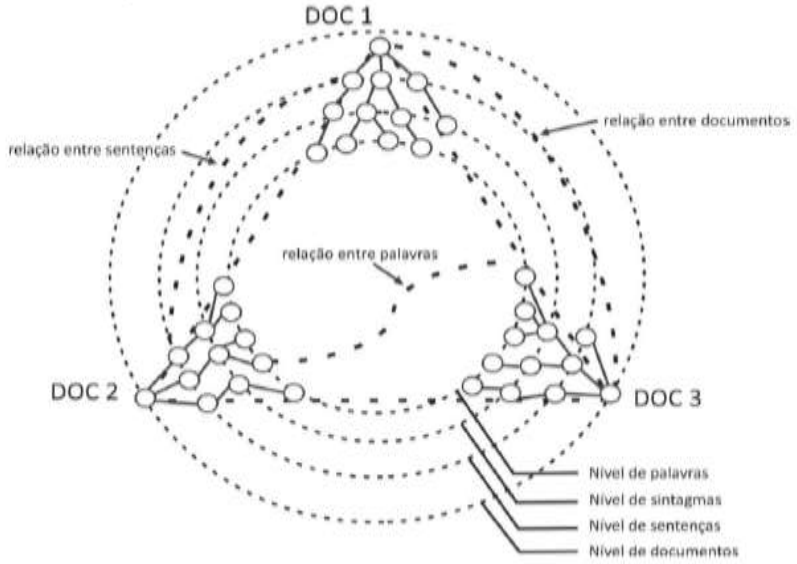
É necessário considerar que, apesar de a redundância, a complementaridade e a contradição serem fenômenos linguísticos legítimos, a partir da ótica aqui proposta, trata-se de fenômenos não propositais. Em outros modelos semânticos, como a *Rhetorical Structure Theory* (RST)[2] (MANN; THOMPSON, 1987), as relações semânticas são construídas propositalmente pelos autores dos textos., em que é possível inferir intenções retóricas a partir da materialidade textual. Já no modelo CST, as relações semânticas são identificadas pelos anotadores, não pelos autores. Nesse sentido, ao analisar textos distintos que falam sobre o mesmo assunto, o anotador identifica traços que caracterizam tais fenômenos, como, por exemplo, o fato de haver informações divergentes entre pares de sentenças, revelando a *contradição* entre elas.

De acordo com Radev (2000), ao estudar um conjunto de textos que possuem o mesmo assunto, é possível que as informações sejam compartilhadas lexicalmente, sintagmaticamente, sentencialmente ou intertextualmente. Assim, as relações propostas no

[2] A proposta do modelo teórico-discursivo da RST é analisar os textos por meio da geração de árvores sintáticas, em que as unidades de análise (no caso, as sentenças) estejam interconectadas por relações semânticas, como *Elaboration* e *Contraste*. Nesse sentido, quando a árvore sintática de um texto apresenta interconexão entre todas as unidades de análise, há um texto coeso e coerente.

modelo CST podem rotular relações estabelecidas entre unidades informacionais desses diferentes níveis linguísticos, como ilustrado na Figura 1.

Figura 1: ***Esquema de relacionamento CST***



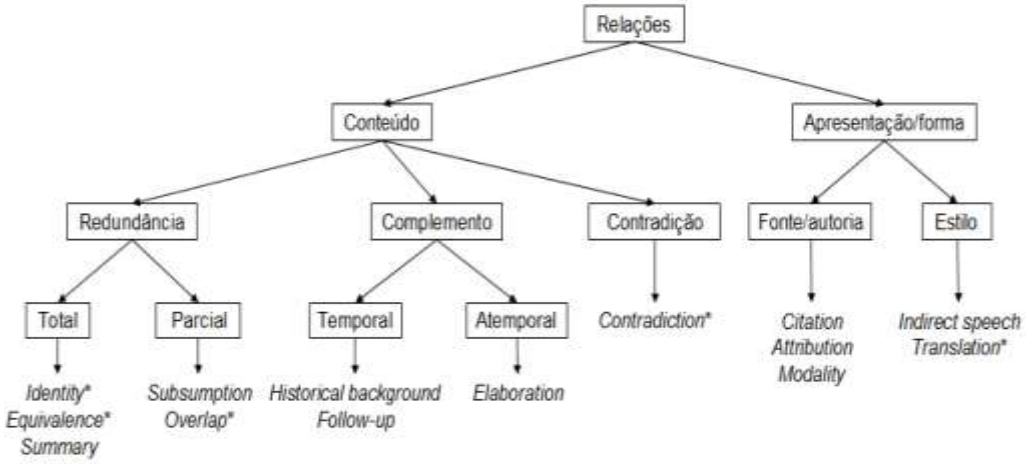
Fonte: Radev (2000).

Na Figura 1, vê-se que os níveis de análise nos quais as relações CST são identificadas compõem uma espécie de hierarquia (por palavras, sintagmas, sentenças ou textos), representados por linhas pontilhadas, ainda que usualmente a análise seja realizada no nível sentencial. Na ilustração, cada um dos 3 documentos (DOC 1, DOC 2 e DOC 3) está representada por um subgrafo, que codifica relações intertextuais. As relações CST que podem ser estabelecidas nos diferentes níveis estão representadas por linhas pontilhadas mais grossas.

Ao realizar a análise de textos jornalísticos em PB, Aleixo e Pardo (2008) observaram que algumas relações identificadas em trabalhos para o inglês (RADEV, 2000) não ocorriam no *corpus* em PB. Assim, ao final da análise, os autores decidiram que os rótulos que não ocorressem nos textos deveriam ser excluídos do conjunto de relações, ou deveriam ser

unificados, quando apresentassem alto grau de similaridade. A versão desse conjunto resultou em 14 relações, organizadas posteriormente em uma tipologia proposta por Maziero *et al.* (2010). Tal tipologia está representada na Figura 2.

Figura 2: ***Tipologia de relações CST para o PB.***



Fonte: Maziero* et al. *(2010).

Na tipologia ilustrada na Figura 2, as relações CST foram organizadas em dois grandes grupos: *relações de conteúdo* e *forma*, com suas respectivas subdivisões. As relações de conteúdo podem ser organizadas em função dos fenômenos linguísticos "redundância" (subdividindo-se em *total* e *parcial*), "complementaridade" (apresentando os tipos *temporal* e *atemporal*) e "contradição". As relações de forma, por sua vez, podem ser do tipo "fonte/autoria" ou "estilo". Na figura o símbolo "*" indica que a relação não tem direcionalidade.

Segundo Maziero *et al.* (2010), a complementaridade, de modo geral, ocorre entre um par de sentenças (S1 e S2), em que S2 apresenta informação complementar em relação a algum elemento presente em S1.

As relações CST de complementaridade temporal podem ser de 2 tipos. O par de sentenças será classificado como complementaridade temporal quando: (i) S2 apresenta informações históricas/passadas sobre algum elemento presente em S1 e (ii) S2 apresenta acontecimentos/eventos que sucederam os acontecimentos/ eventos presentes em S1; os acontecimentos em S1 e em S2 devem ser relacionados e ter um espaço de tempo relativamente curto entre si.

Quadro 1: Par de sentenças anotadas com complementaridade temporal.

Complementaridade temporal	Sentenças
(i) S2 apresenta informações históricas/ passadas sobre algum elemento presente em S1 (S1→S2)	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 porta-voz das Nações Unidas. S2: Acidentes aéreos são frequentes no Congo, onde 51 companhias privadas operam com aviões antigos principalmente fabricados na antiga União Soviética.
(ii) S2 apresenta acontecimentos/ eventos que sucederam os acontecimentos/ eventos presentes em S1 (S1→S2)	S1: A pista auxiliar de Congonhas abriu às 6h, apenas para decolagens. S2: Congonhas só abriu para pousos, às 8h50.

Fonte: Elaboração própria.

No Quadro 1, a complementaridade do tipo (i) é ilustrada por um par de sentenças provenientes de textos que relatam "um acidente aéreo Congo". As sentenças do par estabelecem relação de complementaridade temporal porque S1 e S2 apresentam conteúdo em comum ("acidente aéreo no Congo"), sendo que S2 apresenta uma informação adicional (histórica) sobre esse conteúdo que, nesse caso, diz respeito à "ocorrência frequente de acidentes aéreos no Congo (por causa do uso de aviões velhos)". De acordo com a tipologia apresentada por Maziero *et al.* (2010), esse tipo de complementaridade temporal é capturado pela relação CST *Historical background*.

Ainda com relação ao Quadro 1, a complementaridade temporal do tipo (ii) é ilustrada por um par de sentenças que possui como assunto principal os "atrasos e cancelamentos no

aeroporto de Congonhas devido ao mau tempo". As sentenças estão em complementaridade temporal porque S1 e S2 apresentam informação comum ("abertura das pistas do aeroporto de Congonhas"), sendo que S2 demonstra um acontecimento que sucedeu ao evento descrito em S1 após um intervalo curto de tempo: "o horário de abertura da pista (principal) para pouso", que ocorreu após a "abertura da pista auxiliar para decolagem". Segundo a tipologia apresentada por Maziero *et al.* (2010), esse tipo de complementaridade temporal é explicitado pela relação CST *Follow-up*. A relação de sequência temporal entre o evento focalizado em S2 e o evento descrito em S1 envolve a ocorrência de *expressões temporais* que, segundo Baptista *et al.* (2008), são do tipo "tempo_calendário" e subtipo "data" ("6h" e "8h50").

A relação de complementaridade atemporal, ao contrário das exemplificadas anteriormente, não foca em conteúdo que indica a localização no tempo (anterior ou posterior) de um acontecimento/fato em relação a outro. Essa complementaridade estabelece-se quando, dado um par de sentenças, S2 detalha algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1. Além disso, o elemento elaborado em S2 deve ser o foco de S1.

Quadro 2: Par de sentenças anotadas com complementaridade atemporal.

Complementaridade atemporal	**Sentenças**
S2 detalha/refina/elabora algum elemento presente em S1, sendo que S2 não deve repetir informações presentes em S1 (S1 ←S2)	S1: Apesar da definição, o cronograma da obra não foi divulgado. S2: O cronograma da obra depende de estudos finais que estão sendo realizados pela Infraero.
	S1: As vítimas do acidente foram 14 passageiros e três membros da tripulação. S2: Segundo fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.

Fonte: Elaboração própria.

No Quadro 2, o primeiro par é formado por sentenças provenientes de textos que informam sobre a "reforma da pista principal do aeroporto de Congonhas". Nele, observa-se que S1 e S2 possuem conteúdo comum ("cronograma da obra"), sendo que S2 fornece uma informação adicional sobre esse conteúdo. No caso, a informação adicional em relação a S1 é o foco de S2 e consiste em "a razão pela qual o cronograma da obra não foi divulgado" ("dependente de estudos finais que estão sendo realizados pela Infraero"). Assim, esse par de sentenças foi anotado com a relação *Elaboration*, pois a informação adicional não está pautada no aspecto temporal, como é característico das relações CST demonstradas no Quadro 1.

Baseando-se no fato de que as relações de complementaridade sempre apresentam conteúdo informacional aditivo entre as sentenças de um par, e o fato de conjunções aditivas evidenciarem semanticamente o sentido de adição, parece pertinente investigar se é possível utilizar a ocorrência dessas conjunções como traço sinalizador desse fenômeno linguístico. Na próxima seção, apresenta-se a metodologia utilizada neste estudo a fim de corroborar esta hipótese.

3. Metodologia

Para analisar especificamente a complementaridade, fez-se um recorte no CSTNews (CARDOSO *et al.,* 2011) dos pares de sentenças anotados com as relações de complementaridade. O CSTNews trata-se de um *corpus* composto por 50 coleções de textos jornalísticos, organizados pelas seções dos jornais (como mundo, dinheiro e esporte). Cada coleção é composta por dois ou três textos que versam sobre o mesmo assunto, além de haver diversas anotações linguísticas, a saber: (i) relacionamentos semânticos multidocumento via CST; (ii) anotação de expressões temporais dos textos-fonte; (iii) etiquetação morfossintática; (iv) anotação dos sentidos dos substantivos e verbos; (v) anotação de aspectos informacionais

nos sumários multidocumento (*o quê*, *onde* e *quando,* por exemplo), (vi) anotação semiautomática dos textos-fonte via RST e (vii) anotação manual de subtópicos informativos em cada texto-fonte do *corpus*.

O recorte realizado consistiu em selecionar, por meio da interface *online*[3] de consulta ao *corpus*, apenas os pares de sentenças anotadas com as relações *Follow-Up, Historical Background* e *Elaboration*. Esse recorte resultou em um *subcorpus* do CSTNews, cujos dados quantitativos estão descritos na Tabela 1.

Tabela 1: Elaboração própria.

Complementaridade	Relação CST	Qt. de par	Total
Atemporal	*Elaboration*	343	343
Temporal	*Follow-up*	293	370
	Historical background	77	
--	--	--	**713**

Fonte: Elaboração própria.

Após a construção do *subcorpus*, os pares de sentença foram armazenados em um arquivo .txt e, em seguida, submetidos ao AntConc (ANTHONY, 2005).

O AntConc é um *software* livre e de acesso gratuito para análise de *corpora* linguísticos. Nele, há diversas ferramentas para análises dos textos. Na ferramenta *Concordancer* é possível observar qual o contexto frasal em torno de determinado item lexical; para tanto, é necessário indicar a quantidade de caracteres à direita e à esquerda do item a ser analisado. Na ferramenta *(Key)Word list* podem ser elaboradas as listas de palavras mais recorrentes ou específicas no conjunto de textos (*keywords*) ou ainda a ocorrência de todas as palavras do

[3]Disponível em: http://nilc.icmc.usp.br/CSTNews/

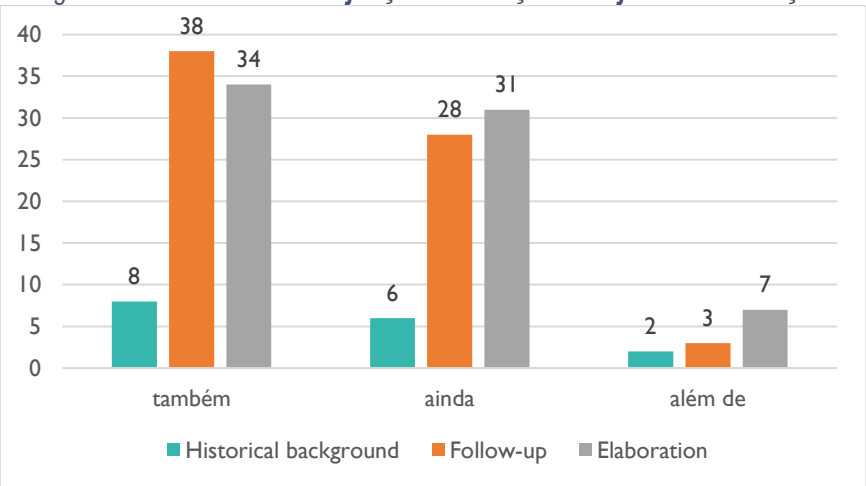
texto (wordlist). Na ferramenta *Concordance Search Term Plot* é possível plotar gráficos para observação da ocorrência do termo de busca ao longo do *corpus*, já que o Antconc permite a análise de múltiplos textos simultaneamente.

Neste estudo, a ferramenta utilizada foi *Concordancer* para analisar conjunções e locuções conjuntivas de valor aditivo. Partindo da definição da complementaridade proposta por Maziero (2012), foi necessário restringir a observação às primeiras sentenças dos pares, já que a informação complementar se apresenta nessas sentenças. Com isso, apenas as conjunções "*também*", "*ainda*" e "*além de*" foram analisadas, já que somente elas ocorreram segundo a referida restrição.

4. Resultados e discussão

Nesta seção apresentam-se os resultados das análises com base nas listas de colocados[4] gerados a partir do AntConc. A ocorrência das conjunções e locuções conjuntivas de valor semântico aditivo está demonstrada na Figura 3.

Figura 3: ***Ocorrência de conjunções e locuções conjuntivas de adição.***



[4] Entende-se como Colocados "uma relação lexical entre duas ou mais palavras em que há uma tendência à coocorrência entre algumas palavras que ocorrem no texto" (STUBBS, 2002, *apud* GEERAERTS, 2010).

Fonte: Elaboração própria.

Observa-se, a partir da Figura 3, que a ocorrência da conjunção "também" é superior que a conjunção "ainda" e a locução "além de" nos pares de sentença anotados com as relações CST de complementaridade. A relação *Follow-up* apresenta maior ocorrência entre as conjunções "também" e "ainda"; a relação *Elaboration* apresenta maior ocorrência para a locução "além de"; por fim, a relação *Historical background* obteve a menor ocorrência entre as conjunções analisadas.

Além dessa análise quantitativa, foi observado o comportamento sintático das conjunções nos pares de sentença do *subcorpus*. A conjunção "também" ocorreu colocada a um verbo 37 vezes; a colocação "ainda + verbo" ocorreu 26 vezes e "ainda + assim", 9 vezes. Não foram obtidos resultados desta análise para a locução "além de". A seguir, têm-se exemplos recuperados do *subcorpus* em que essas construções são explicitadas.

(1)

S1: O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congolesa, a Trasept Congo, *também levava* uma carga de minerais.

S2: Ao menos 17 pessoas morreram após a queda de um avião de passageiros na República Democrática do Congo.

Em (1), a conjunção "também" juntamente com o verbo "levava", em S1, evidencia a informação complementar ("uma carga de minerais") em relação à S2. Nas ocorrências do *corpus*, sempre que a conjunção "também" ocorre ela está atrelada à informação

complementar. Isso acontece porque a complementaridade, nesse caso, está vinculada a um referente recuperado em S1, o qual, por sua vez, retoma a informação principal em S2.

(2)

S1: O Brasil *ainda derrubou* o antigo recorde panamericano, que era de 3m17s18, com a nova marca de 3m15s90.

S2: Em uma disputa emocionante, o Brasil conquistou nesta sexta-feira a medalha de ouro no revezamento 4x100 metros livres, uma das provas mais charmosas da natação, ao cravar o tempo de 3min15s90 (novo recorde pan-americano e sul-americano).

Em manuais de gramática, a conjunção “ainda”, em geral, é apresentada como valor concessivo. Em (2), a conjunção e o verbo complementam a informação do desempenho dos nadadores do Brasil, admitindo que eles derrubaram “o antigo recorde panamericano”. Além da conjunção “ainda” estar acompanhando um verbo e, nesse caso, ter valor aditivo, sempre que ela ocorre na forma de locução conjuntiva de “ainda assim”, também possui valor complementar. Aqui, ambas as formas de ocorrência estão atreladas à recuperação da informação principal de S2, em S1.

(3)

S1: Lula disse que *além de* melhorar a qualidade de vida dos brasileiros, as obras vão gerar empregos.

S2: O dado concreto é que nós vamos fazer deste país um verdadeiro canteiro de obras em se tratando de infra-estrutura", disse.

Por fim, em (3), a locução conjuntiva “além de”, em S1, sinaliza a informação complementar (“melhorar a qualidade de vida dos brasileiros”) com relação à informação principal, em S2 (“pais um verdadeiro canteiro de obras”). Aqui, a locução está vinculada à informação principal de S1, mas de forma remissiva em S2.

5. Considerações finais

O estudo aqui realizado é de caráter preliminar e exploratório. Assim, os resultados obtidos, apesar de não terem sido verticalizados, podem ser tomados como indícios do comportamento linguístico da complementaridade informacional segundo o modelo CST.

Estudos anteriores a este já demonstraram a dificuldade de se caracterizar a complementaridade informacional, tendo em vista como o fenômeno é analisado de acordo com o modelo teórico supracitado. Nesse sentido, ter dois pontos norteadores nesta área de estudo: a descrição linguística e a implementação computacional.

Na área de Linguística (e subáreas envolvidas, como Linguística descritiva e Linguística de *corpus*) é necessário descrever o fenômeno com suas devidas complexidades de ocorrência no conjunto de textos analisados. Entretanto, é preciso compreender que a totalidade da descrição não seja passível de implementação computacional. É por conta deste último apontamento que métodos utilizam pouco conhecimento linguístico (categorizados como superficiais) de identificação da complementaridade são amplamente utilizados em PLN e, por vezes, apresentam resultados superiores quando comparados aos métodos profundos, ou seja, que utilizam muito conhecimento linguístico.

Compreendendo o PLN como uma área de interseção entre Linguística e Computação, é primordial tomar a metodologia deste trabalho e os resultados obtidos como sinalizadores da complementaridade. As conjunções e locuções conjuntivas de adição não são características exclusivas do fenômeno linguístico aqui explorado, muito menos característica definitória das relações CST que o traduzem. Porém, a ocorrência desse tipo de conjunção pode ser tida como traços que sinalizam a presença de complementaridade entre pares de sentenças extraídas de textos distintos e que versam sobre o mesmo assunto. Algo bastante semelhante vem sendo discutido por Taboada e Das (2013) no âmbito da RST.

Assim, desafios a serem trabalhados em trabalhos futuros são (i) propor descrições linguísticas mais profundas e que considerem outros níveis de análise, como o semântico e o pragmático; (ii) observar as conjunções aditivas em coocorência a outros sinalizadores (superficiais e/ou profundos) da complementaridade; e (iii) analisar o impacto do uso de conjunções aditivas na identificação automática da complementaridade em pares de sentenças.

Referências bibliográficas

ALEIXO, P.; PARDO, T.A.S. CSTNews: um corpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). **Série de relatórios técnicos do NILC (NILC-TR-08/05)**. São Carlos/SP, p. 15, 2008.

ANTHONY, L. AntConc: design and development of a freeware corpus analysis toolkit for the technical writing classroom. In: **IPCC 2005. Proceedings. International Professional Communication Conference**, 2005. IEEE, 2005. p. 729-737.

BAPTISTA, J. HAGÈGE, C. MAMEDE, N. Proposta de anotação e normalização de expressões temporais da categoria TEMPO para o HAREM II. In: **Actes de Encontros do Segundo HAREM**. 2008.

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: **Proceedings of the 3rd RST Brazilian Meeting**, pp. 88-105. Cuiabá/MT, Brasil. 2011.

GEERAERTS, D. **Theories of Lexical Semantics**. New York: Oxford University Press, 2010

MANN, W.C.; THOMPSON, S.A. **Rhetorical structure theory**: A theory of text organization. University of Southern California, Information Sciences Institute, 1987.

MAZIERO, E. G.; JORGE, M. L. C.; PARDO, T. A. S. Identifying multi-document relations. In **Proceedings of International Workshop on Natural Language Processing and Cognitive Science**. Funchal/Madeira. p. 60-9. 2010.

MAZIERO, E.G. **Identificação automática de relações multidocumento**. 2012. 118 f. Dissertação (Mestrado em Ciências de computação e Matemática computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2012.

RADEV, D. R. A common theory of information fusion from multiple text sources step one: cross-document structure**.** In **Proceedings of the 1st SIGdial workshop on Discourse and dialogue.** Vol 10. p. 74-83. 2000.

SILVA, N.; DI-FELIPPO, A. **Descrição e análise do fenômeno da contradição para a Sumarização Automática Multidcoumento.** Série de relatórios técnicos do NILC. NILC-TR-14-03. São Carlos/SP. 2014.

SOUZA, J. W. C.; DI-FELIPPO, A.; PARDO, T. A. S. **Investigação de métodos de identificação de redundância para Sumarização Automática Multidocumento.** Série de Relatórios do NILC. NILC-TR-12. São Carlos-SP. 2012.

SOUZA, J.W.C. **Descrição linguística da complementaridade para a sumarização automática multidocumento**. 2015. 102 f. Dissertação (Mestrado em Linguística) - Programa de Pós-graduação em Linguística, Universidade Federal de São Carlos, São Carlos, 2015.

STUBBS, Michael. Conrad, concordance, collocation: heart of darkness or light at the end of the tunnel?. **The Third Sinclair Open Lecture, University of Birmingham**, 2004.

TABOADA, M.; DAS, D. Annotation upon Annotation: Adding Signalling Information to a Corpus of Discourse Relations. In: Dipper, S.; Zinsmeister, H.; Webber, B. (orgs). **Dialogue and Discourse**., v.4, n. 2, p. 249-281. 2013.

TAUFER, P. Massa de informações digitais pode ser usada em benefício da população. **Jornal da Globo, 26 dez. 2013**. Disponível em: <http://g1.globo.com/jornal-da-globo/noticia/2013/12/massa-de-informacoes-digitais-pode-ser-usada-em-beneficio-da-populacao.html> Acesso em: 02 fev. 2015.