Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTNews

Roana Rodrigues[1], Jackson Wilke da Cruz Souza[2], Paula Christina Figueira Cardoso[3]

[1]Programa de Pós-Graduação em Letras - Universidade Federal de Sergipe (UFS), São Cristóvão, SE - Brasil

[2]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia (UFBA), Salvador, BA - Brasil

[3]Departamento de Computação Aplicada - Universidade Federal de Lavras (UFLA), Lavras, MG - Brasil

roana@academico.ufs.br, jackcruzsouza@gmail.com, paula.cardoso@ufla.br

Abstract. *Rhetorical Structure Theory (RST) is a discourse theory in which the coherence of a text can be characterized by a tree structure, where the discourse units are the leaves and the nodes represent the rhetorical relations between them. Although it is known that the identification of connectives that indicate these relations plays an important role in text processing, the absence of a prototypical discourse marker does not eliminate the possibility of their interpretation. In this paper, we describe the analysis of a sample from a corpus already annotated with RST, aiming to identify how these relations are signaled in the discourse. The results highlight the importance of investigating other flags in addition to DMs.*

Resumo. *Rhetorical Structure Theory (RST) é uma teoria discursiva na qual a coerência de um texto pode ser caracterizada por uma estrutura de árvore, em que as unidades discursivas são as folhas e os nós representam as relações retóricas entre elas. Embora seja conhecido que a identificação de conectivos que indicam as relações desempenha um papel importante no processamento do texto, a ausência de um marcador discursivo (MD) prototípico não impede a* possibilidade de sua interpretação*. Nesta proposta preliminar, descreve-se a análise de um recorte de um córpus já anotado com RST, com o objetivo identificar como as relações são sinalizadas no discurso. Os resultados destacam a importância de investigar outros sinalizadores para além de MDs.*

1. Introdução

O modelo teórico RST (*Rhetorical Structure Theory*) é uma teoria linguístico-descritiva que tem o discurso como ponto de partida e visa analisar e descrever "fenômenos de ordem sintática, semântica e pragmática que se "gramaticalizam" nos textos" [Hirata-Vale e Oliveira 2014, p. 406]. Trata-se de um modelo muito utilizado no Processamento de Língua Natural (PLN), contribuindo sobretudo com a construção de *parsers* e de ferramentas automáticas de sumarização, tradução e avaliação de textos.

Tendo origem no trabalho de Mann e Thompson (1988), a RST se enquadra no chamado *Funcionalismo da Costa-Oeste Norte-Americana* e estabelece diretrizes para a anotação e descrição das relações retóricas (também nomeadas *discursivas* ou *de*

coerência) de um texto, ou seja, das relações que permitem a coesão e coerência textuais. Tem-se como objeto de estudo as relações estabelecidas entre *núcleo* e *satélite* - além da consideração de relações *multinucleares* -, explicitando-se a intenção (produção do falante/escritor) e o efeito (recepção do ouvinte/leitor) do que é dito. Na Figura 1, exemplificamos as relações retóricas de um fragmento de texto extraído do córpus CSTNews [Cardoso *et al*., 2011][1]:

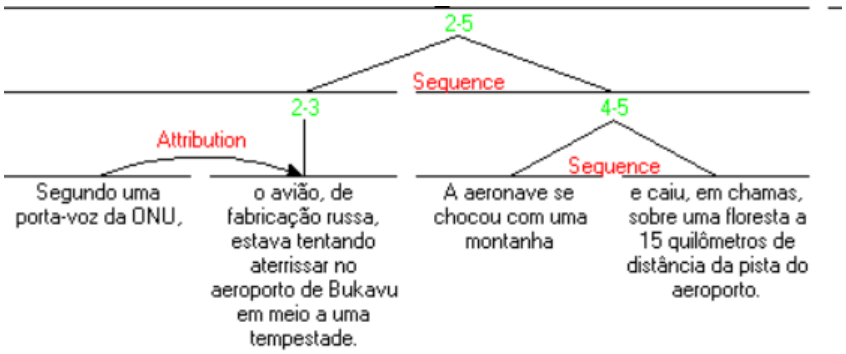

Figura 1. Exemplo de relações RST do córpus CSTNews

Na Figura 1, a árvore apresenta três relações retóricas: em 2-3, o *satélite* (*Segundo uma porta-voz da ONU*) caracteriza-se por apresentar a fonte de uma mensagem e o *núcleo*, a mensagem; tanto em 4-5, como em 2-5, têm-se relações multinucleares, pois apresentam núcleos em sequência, tendo como efeito o reconhecimento de uma sucessão temporal dos eventos.

Segundo Taboada e Mann (2006), a taxonomia adotada na RST não é fixa, isto é, não existe uma prescrição da teoria para o conjunto de relações retóricas possíveis para uma língua. No entanto, os autores salientam a necessidade de prudência na quantidade de relações, devido às dificuldades no processo manual de identificação e anotação de um texto. Para o português brasileiro (PB), Pardo (2005) propõe um conjunto de 32 relações retóricas possíveis[2], em que são estabelecidas informações relativas às restrições sobre o núcleo, o satélite, a relação núcleo-satélite, além de possíveis efeitos desencadeados no leitor.

As relações retóricas são comumente determinadas com base nos marcadores discursivos (MDs)[3] presentes em um texto. Da anotação do CorpusTCC, Pardo (2005, p. 64-67) criou um quadro com a distribuição dos MDs em função das relações que

[1] A árvore em (1) foi visualizada e extraída na RSTTool, ferramenta utilizada para elaborar e abrir diagramas, disponível para *download* em: http://www.wagsoft.com/RSTTool/section2.html. Acesso em junho de 2023.

[2]Relações retóricas do PB, segundo Pardo (2005): *antithesis, attribution, background, circumstance, comparison, concession, conclusion, condition, contrast, elaboration, enablement, evaluation, evidence, explanation, interpretation, join, justify, list, means, motivation, non-volitional cause, non-volitional result, otherwise, parenthetical, purpose, restatement, same-unit, sequence, solutionhood, summary, volitional cause, volitional result.*

[3] Assim como definido por Das e Taboada (2018), nesta pesquisa *marcadores discursivos,* também nomeados *conectivos*, são os elementos de um texto que estabelecem relações entre as proposições, incluem as conjunções, locuções conjuntivas, locuções preposicionais e expressões lexicalizadas.

sinalizam. Segundo o autor, embora nem todas as relações possuam MDs associados, os textos anotados com a relação *Sequence*, por exemplo, possuíam, majoritariamente, os marcadores *e, a partir de, em seguida*; já a relação *Explanation* foi marcada por *pois, isto é* e *porque*.

Taboada e Das (2013) destacam que a compreensão de textos parte da construção de uma representação das informações presentes nele, em que uma parcela desse processo compreende remontar as possibilidades de organização das proposições. Os autores apontam que a identificação de conectivos que indicam as relações possíveis facilita o processamento do texto, mas que a ausência de um marcador discursivo prototípico não furta a possibilidade de sua interpretação. Assim, argumentam que as relações de coerência são, na verdade, entidades cognitivas e, por conta disso, há possibilidade de interpretação do texto. Como resultado, trabalhos nessa perspectiva não consideram as relações não sinalizadas. Em (1), exemplo extraído de Das e Taboada (2018), as sentenças podem ser anotadas no modelo RST como *Contrast*; porém, seria uma relação implícita, já que não apresenta nenhum MD explícito. Apesar disso, é possível compreender o contraste quando considera-se as unidades lexicais *tall* e *short*.

(1) *John is tall. Mary is short*.[4]

Nesse sentido, em estudos mais recentes [Antonio 2017; Das e Taboada 2018] discute-se a necessidade de criação de tipologias dos sinalizadores discursivos para além dos marcadores, tais como *entonação, cadeia lexical, pontuação, tempo verbal,* entre outros, já que nem toda relação possui um marcador discursivo "explícito"/"prototípico" a ele relacionado.

Sendo assim, baseando-nos na proposta de Das e Taboada (2018), objetivamos analisar um recorte de um córpus já anotado em RST, para avaliar os sinalizadores presentes nos textos. Trata-se de uma caracterização preliminar da qual derivará uma tipologia de sinalizadores (simples e combinados) para as relações retóricas do PB. Para tanto, este artigo está organizado da seguinte maneira, além desta introdução: na seção 2, descrevemos os trabalhos de base para a presente investigação, ou seja, as contribuições de Antonio (2017) para o PB e Das e Taboada (2018) para o inglês; em seguida, apresentamos os processos e decisões metodológicas desta pesquisa; nossas primeiras percepções sobre o processo de anotação dos sinalizadores em córpus do PB e as considerações iniciais desta tarefa; e, então, as considerações finais e trabalhos futuros.

2. Sinalizadores discursivos

Nesta seção destacamos dois trabalhos relacionados aos objetivos traçados para esta pesquisa. É importante salientar que os trabalhos apresentados se baseiam em diferentes registros linguísticos: Antonio (2017) parte de um córpus oral, semi-formal e analisando o PB; já Das e Taboada (2018) analisam um córpus de textos jornalísticos do Inglês.

De acordo com Antonio (2017, p. 105), "as relações de coerência, por serem de sentido, e não de forma, podem ser estabelecidas e interpretadas independentemente de serem marcadas explicitamente por conectivos". Posto isso, o autor investiga a

[4] Tradução livre: "John é alto. Maria é baixa.".

percepção de professores universitários em relações retóricas de 10 excertos de textos orais, considerando-se elementos para além dos MDs.

Como resultado, Antonio (2017) elenca as seguintes pistas formais destacadas pelos informantes: sinais de pontuação (dois pontos); modo de oração (*pergunta-resposta*); aspectos fonológicos (*entonação*); aspectos morfossintáticos (*tempo verbal, expressões adverbiais*); aspectos semânticos (*interdependência entre os estados-de-coisas; o próprio sentido das porções textuais envolvidas ou de palavras-chave nessas porções textuais, como paralelismo nas construções; referência anafórica*); e aspectos cognitivos (*ativação de referentes a partir de um modelo cognitivo global*).

Para o inglês, Das e Taboada (2018), a partir de um córpus já anotado com as relações retóricas (*RST Discourse Treebank*), realizam uma anotação minuciosa dos sinalizadores discursivos dessas relações, construindo assim o *RST Signalling Corpus*. Para tanto, os autores consideraram elementos formais para além dos MDs, organizando a taxonomia conforme se apresenta na Figura 2, retirado de Das e Taboada (2018).

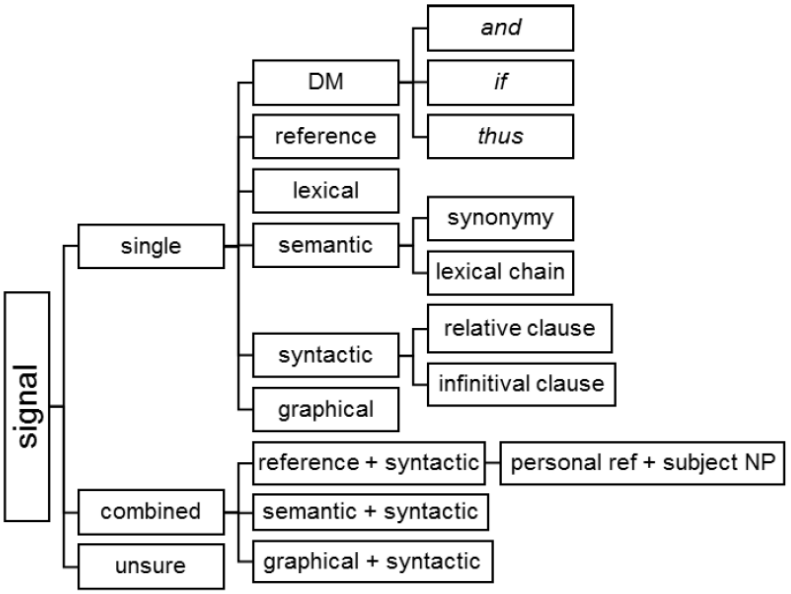

Figura 2. Fragmento da taxonomia hierárquica de sinalizadores discursivos

Na Figura 2, Das e Taboada (2018) pontuam que uma determinada relação pode ser anotada por um *sinalizador único* (tais como MDs, referência pessoal, oração relativa, dois pontos) ou um *sinalizador combinado* (vírgula + oração no particípio passado; construção sintática paralela + cadeia lexical, entre outros). Além disso, no processo de anotação, os autores relatam que houve casos anotados como *incertos*, nos quais não foi possível estabelecer com precisão o sinalizador que denota determinada relação.

3. Processos metodológicos

Para viabilizar esta investigação, foi selecionado o córpus CSTNews [Cardoso *et al.* 2011] que é anotado manualmente de diferentes maneiras quanto à organização do

discurso, sendo uma delas a RST. O córpus possui 50 conjuntos de textos (*clusters*), organizados por assunto, que foram coletados manualmente no ano de 2007. No total, são 140 textos jornalísticos, que juntos contabilizam 2.088 sentenças e 47.240 palavras.[5]

Os anotadores realizaram uma análise incremental, aproveitando a estrutura organizacional do texto fonte. Nessa abordagem, assume-se que as orações adjacentes dentro das sentenças devem ser relacionadas em primeiro lugar, seguido pelas sentenças adjacentes dentro dos parágrafos, e, por fim, os parágrafos adjacentes são relacionados. Na maioria das anotações, estabeleceram-se relacionamentos binários, ou seja, conectaram-se dois segmentos em uma relação. Ao final, observou-se a prevalência das relações *Elaboration, List, Attribution, Parenthetical* e *Same-unit* no córpus.

Nesta pesquisa partimos de um córpus anotado com RST, como proposto por Das e Taboada (2018) para a língua inglesa. Em nosso caso, utilizamos o CSTNews e selecionamos aleatoriamente, para a anotação manual dos sinalizadores, 9 *clusters*, que se constituem pelo conjunto de textos (de 2 a 3 textos jornalísticos) sobre a mesma notícia, totalizando 21 documentos anotados. Esse processo foi realizado por três anotadores em grupo, o que possibilitou discussões e tomadas de decisões conjuntas e imediatas. Para esta análise preliminar, decidimos identificar apenas sinalizadores intrasentenciais; os possíveis sinalizadores intersentenciais serão analisados numa fase posterior devido à dificuldade de haver consenso entre os anotadores, mesmo em um processo manual.

Com a anotação finalizada, passamos para a etapa de reflexão e análise dos sinalizadores apontados e as possíveis vinculações com relações retóricas específicas, conforme descrevemos na próxima seção.

4. Discussões e resultados

Antes da apresentação dos dados anotados, destacamos o fato de que, devido à decisão metodológica de anotação de documentos pertencentes a um mesmo *cluster*, muitas relações e sinalizadores se repetiram, por se tratar de proposições redundantes ou com pequenas e sutis variações. Além disso, indagamos a segmentação e anotação de algumas proposições do córpus, no entanto, neste trabalho preliminar, decidimos ignorar esses casos e investigá-los em trabalhos futuros.

Sendo assim, as discussões e resultados apresentados nesta seção são de cunho qualitativo, a partir de observações e discussões gerais dos dados anotados, com foco nas relações mais eminentes, a saber: *Attribution, Elaboration, Parenthetical, List, Sequence* e *Same-unit*. O Quadro 1 apresenta o tipo de sinalizador e as tags anotadas no córpus CSTNews para as relações destacadas.

[5] Detalhes sobre o processo de anotação RST do córpus CSTNews, tais como anotadores, regras de segmentação e concordância da tarefa, podem ser encontrados em Cardoso *et al.* (2011).

Quadro 1. Relações retóricas e sinalizadores discursivos

Relação	Sinalizadores	Exemplos
Attribution	pontuação (aspas, vírgula) + informação sintática (verbo de comunicação (acrescentar, dizer, informar) e fonte).	(2) **["**Estamos resistindo à tentação de chamar o par de planeta duplo porque ele provavelmente não se formou do jeito que os planetas no nosso sistema solar apareceram**",**] [**acrescenta Ivanov**.]
	MD (que) + informação sintática (verbo de comunicação (acrescentar, dizer, informar) e fonte).	(3) [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)].
Elaboration	pontuação (vírgula) + pronome relativo (que).	(4) [Inicialmente, Solange Vieira**,**] [**que** é assessora especial de Jobim**,**]
	pontuação + passiva	(5) [Segundo o jornal "Choson Sinbo"**,**] [**publicado pela Associação de Residentes Coreanos no Japão**]
List	paralelismo + MD (e)	(6) (...) uma dupla de planetas errantes (...) [**que giram** ao redor deles mesmos] [**e que vagam** livremente pelo espaço.]
Parenthetical	pontuação (parênteses, travessão, vírgula) + sigla	(7) [O presidente do Conselho de Ética do Senado, Leomar Quintanilha] [**(PMDB-TO)**]
	pontuação + mudança de tópico	(8) "(...) [publicado pela Associação de Residentes coreanos no Japão] [**(próxima ao regime comunista da Coréia do Norte),**]
Same-unit	pontuação (vírgula) + (concordância verbal (venceu)) + sucede uma relação encaixada (como *Elaboration* ou *Parenthetical*)	(9) [A seleção brasileira masculina de vôlei**, que é treinada por Bernardinho**,] **[venceu** a Finlândia por 3 sets a 0.]
Sequence	MD (e) + tempo verbal	(10) [Alvo de críticas incisivas da oposição desde o acidente com o Airbus da TAM, o atual presidente da Anac, Milton Zuanazzi, já **teria concordado** em renunciar] [**e deve** entregar o cargo nos próximos dias.]
	pontuação (vírgula, ponto final) + numeral	(11) [O time comandado pelo treinador Bernardinho só encontrou um pouco mais de dificuldades no **segundo** set**.**] [No **terceiro**, mesmo com várias reservas como o levantador Marcelinho e Samuel, os brasileiros conseguiram fechar a partida com tranqüilidade.]

Conforme se observa no Quadro 1, as relações que se sobressaíram na anotação possuem sinalizadores combinados, gramaticalizados por MDs prototípicos somados a outros sinais, sobretudo *pontuação*, aspectos morfológicos (*tempo verbal*) e informações sintáticas (*construções de comunicação, passiva lexical, concordância verbal, orações subordinadas*).

A relação *Attribution* apresentou duas possibilidades de combinações de sinalizadores, relacionadas ao discurso em estilo direto (2) e indireto (3). Para a anotação manual, foi possível identificar a introdução do estilo indireto pela conjunção

(MD) *que*. No entanto, é sabido que, para qualquer anotação morfossintática, a multifuncionalidade de *que* é uma questão complexa para o processamento automático da língua. É o caso, por exemplo, da diferença encontrada entre *que* nas relações de *Attribution* e *que* nas relações de *Elaboration*, em que o primeiro é anotado como MD, mas o segundo atua como pronome relativo. Essa complexidade inerente à anotação morfossintática ressalta a necessidade de sinalizadores combinados para a identificação adequada das relações em análise.

As relações *Elaboration* e *Parenthetical* aparecem encaixadas a proposições nucleares, unidas sobretudo por sinais de pontuação (vírgula, travessão e parênteses). A relação *Parenthetical* se distingue ao ser anotada sempre em que são inseridas siglas no texto, além dos casos de *mudança de tópico*, característicos do acréscimo de uma informação adicional (colocada entre parênteses ou travessão). Por sua vez, a relação *Elaboration*, mais abundante no córpus, se assemelha ao comportamento das orações relativas (restritivas e explicativas), tendendo a ser introduzidas pelo pronome relativo *que*, embora apresente outros comportamentos, como se verifica em (5). Referente à *tag* [pontuação + passiva], salientamos a dificuldade de anotação para a distinção entre *particípio passado* (daí a justificativa para anotação da relação pela existência de uma *passiva lexical*) e *adjetivo*, visto que algumas relações de *Elaboration* verificadas no córpus se caracterizam pelo encaixe introduzido por um adjetivo, como em: [***Invicto** na competição,*] [*o Brasil está tranqüilo na liderança do Grupo B*.]. Portanto, em trabalhos futuros a anotação [passiva] deverá ser estudada com mais detalhes.

As relações *List* e *Sequence* são multinucleares, mas a primeira é identificada por relacionar itens comparáveis apresentados nos núcleos, enquanto a segunda se caracteriza por desencadear no leitor o efeito de reconhecimento de sucessão temporal dos eventos apresentados. Essa diferença de *restrição* e *efeito* se materializa nas *tags* anotadas para cada relação: apesar de ambas serem marcadas pela conjunção *e*, a relação *List* enfatiza a igualdade e comparação entre as proposições, indicada, em muitos casos, pela simetria entre as estruturas sintáticas (paralelismo), como em (6); já a relação *Sequence* é marcada pela ideia de sucessão, que se faz evidente tanto pela distinção temporal a partir da conjugação verbal (10), quanto pela ordem numérica de um determinado processo/progresso (11).

Por fim, destacamos os sinalizadores combinados da relação *Same-unit*, em que as informações apresentadas constituem uma única proposição. Na maioria dos casos, essa relação foi identificada devido à concordância verbal. No entanto, uma característica comum aos casos de *Same-unit* foi o fato de essa relação ser precedida por alguma relação RST encaixada (*Parenthetical* ou *Elaboration*), o que pode ser uma informação útil para o estabelecimento de regras para identificação automática dessa e de outras relações associadas.

Evidentemente, outras relações foram anotadas com sinalizadores simples e combinados no recorte do córpus anotado, mas propusemos a descrição dessas seis relações devido à sua frequência elevada. Os exemplos do Quadro 1, de (2) a (11), ilustram a importância de se reconhecer sinalizadores para além dos MDs e indicam caminhos, ainda que preliminares, sobre as características e a possibilidade de

identificação e reconhecimento automático das relações retóricas de um texto - ao menos do gênero jornalístico como o trabalhado nesta pesquisa.

5. Considerações finais e trabalhos futuros

Neste trabalho preliminar, nosso objetivo foi investigar pistas que pudessem sinalizar as relações do modelo teórico RST, partindo de um córpus do PB pré-anotado. O tipo de estudo exploratório que realizamos aqui demonstra a importância e a dificuldade em classificar relações RST considerando apenas MDs.

A ampla utilização de MDs na identificação de relações RST pode ser justificada por conta da possível compreensão das relações do modelo como *unidades de coerência discursiva*; nesse caso, seriam necessários conectivos específicos entre as unidades. Ainda nesse sentido, outra possível justificativa para essa utilização é poder compreender a RST como um modelo gramatical e, por conta disso, parece pertinente partir de MDs para caracterizar as relações do modelo. Entretanto, como apresentado, estudos recentes que utilizaram córpus de outros gêneros textuais (como de redes sociais) salientam a necessidade de explorar outros sinalizadores para além dos marcadores prototípicos.

Ademais, os resultados aqui apresentados apontam para a importância de não considerar MDs de maneira unívoca e como características exclusivas de algumas relações RST. Antes, destacamos a necessidade de analisar combinações entre os MDs e outros sinalizadores. Os marcadores "caso" e "eventualmente", por exemplo, caracterizam a relação *Condition*, dado que não ocorrem em outras relações da teoria. Porém, identificamos que o marcador "mas", característico da relação *Contrast*, foi utilizado em EDUs anotados com a relação *Concession*. Observamos comportamento similar com a conjunção "e", que pode sinalizar tanto a relação *List* quanto a relação *Sequence*.

Quanto às limitações, destacamos o formato da anotação RST disponibilizado no córpus CSTNews. Iniciamos o trabalho identificando os possíveis sinalizadores com *tags* xml. Porém, há relações em que um único sinalizador pode estar entre informações que não foram consideradas na análise. A título de exemplo, tem-se a relação *Parenthetical*, em que a pontuação utilizada para identificar a relação, como travessões e parênteses. Decidir se anotamos com *tags* xml apenas os parênteses separadamente ou se anotamos incluindo o conteúdo dentre eles gera impactos diretos e substanciais na forma com que essa anotação será utilizada em classificadores automáticos, numa fase posterior a este estudo.

Quanto aos trabalhos futuros, pretendemos estender o estudo a outras porções textuais do córpus analisado, já que o estudo teve como ponto de partida a descrição de sinalizadores intrasentenciais. Outra tarefa a ser realizada, em estudo futuro, é ampliar a variabilidade do gênero textual, uma vez que o córpus utilizado é composto apenas por textos jornalísticos, garantindo que os sinalizadores que serão identificados possam ser provenientes de diferentes normas linguístico-gramaticais.

Agradecimentos

Este trabalho foi realizado no âmbito do Centro de Inteligência Artificial da Universidade de São Paulo (C4AI - http://c4ai.inova.usp.br/), com o apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (processo FAPESP #2019/07665-4) e da IBM. Este projeto também foi apoiado pelo Ministério da Ciência, Tecnologia e Inovações, com recursos da Lei N. 8.248, de 23 de outubro de 1991, no âmbito do PPI-Softex, coordenado pela Softex e publicado como Residência em TIC 13, DOU 01245.010222/2022-44.

6. Referências

Antonio, J. D. (2017) Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores. *Delta*, V. 33, pp. 79-108.

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. (2011) 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.

Das, D. e Taboada, M. (2018) RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, Vol 52, N. 1, pp. 149-184.

Hirata-Vale, F. B. M. e Oliveira, T. P. (2014) Modelos e Métodos de Análise Funcionalista. In: GONÇALVES, A. V.; GÓIS, M. L. S. (Org.). *Ciências da Linguagem:* O Fazer Científico - Volume 2. Campinas: Mercado de Letras.

Mann, W. C. e Thompson, S. A. (1988) Rhetorical structure theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, Vol. 8, N.3, pp. 243–281.

Pardo, T. A. S. (2015) Métodos para análise discursiva automática. Tese (Doutorado em Ciências da Computação e Matemática Computacional). São Carlos: Universidade de São Paulo, 211p.

Taboada, M. e Mann, W. C. (2006) Rhetorical Structure Theory: Looking back and moving ahead. *Discourse Studies*, Vol. 8, N. 3, pp. 423-459.

Taboada, M. e Das, D.. (2013) Annotation upon annotation: Adding signalling information to a corpus of discourse relations. *Dialogue & Discourse*, V. 4, N. 2, pp. 249-281.