A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos

Paula Cardoso[1], Jackson Souza[2], Roana Rodrigues[3], Ewerson Dantas[3], Larissa Santa Bárbara[3], Mateus Araújo[2], Naira Gama[2], Tobias Almeida[4], Gabriel Cruz[2]

[1]Universidade Federal do Pará - Belém/PA

[2]Universidade Federal da Bahia - Salvador/BA

[3]Universidade Federal de Sergipe - Aracajú/SE

[4]Universidade Federal de Lavras - Lavras/MG

pcardoso@ufpa.br,{jacksoncruz,pereiramateus,gabrielsizinio,nairagama}@ufba.br,

{roana,larissajesus}@academico.ufs.br, tobias.almeida@estudante.ufla.br

Abstract. *Due to their capability to enable the observation of linguistic and social behaviors, annotated corpora have become of interest to various fields of study. In the context of Rhetorical Structure Theory (RST), this paper presents the methodological and practical processes involved in annotating discourse markers within a Brazilian Portuguese journalistic corpus. Additionally, we provide initial quantitative and qualitative assessments of the decisions made by the annotation team.*

Resumo. *Por serem recursos que permitem a observação de comportamentos e usos linguísticos e sociais, os corpora anotados passaram a ser de interesse de diferentes áreas do conhecimento. No contexto da Rhetorical Structure Theory (RST) apresentamos neste trabalho os processos metodológicos e práticos de anotação de sinalizadores discursivos em um corpus jornalístico do português do Brasil. Ainda, apresentamos as primeiras avaliações (quanti e qualitativa) sobre as decisões tomadas pelo grupo de anotadores.*

1. Introdução

A Linguística de *corpus* (LC), enquanto área, instiga a utilização de técnicas e metodologias que nos levam a reunir grandes conjuntos de dados textuais (escritos, orais ou multimodais), a fim de descrever fenômenos linguísticos. Em interface ao Processamento de Linguagem Natural (PLN), uma das tarefas que a LC se propõe a realizar é a anotação desse conjunto de dados, tida como "o processo de enriquecer um *corpus*, adicionando informações linguísticas inseridas por humanos ou máquinas com um objetivo teórico ou prático" [Pedro e Vale 2018].

Por serem recursos que permitem a observação de comportamentos e usos linguísticos e sociais, os *corpora* anotados passaram a ser de interesse de diversas áreas do conhecimento, como Humanidades digitais, Linguística e Computação.

[Pustejovsky e Stubbs 2012] apontam que a análise dos *corpora* permite desvendar a natureza da linguagem e, consequentemente, capturar possíveis propriedades que possam ser modeladas computacionalmente.

Porém, esse processo de anotação tende a ser custoso, já que grande parte é realizada de forma semiautomática e requer intervenção humana. [Hovy e Lavid 2010] apresentam uma metodologia genérica sobre esse processo, que engloba etapas como preparação do conjunto de dados, instanciação da base teórica, anotação de fragmentos do *corpus*, medição das decisões de anotação e escalabilidade do processo de maneira automática. No entanto, essa tarefa pode ser ajustada conforme o tipo de anotação a ser realizada, o que pode levar à omissão de algumas das etapas sugeridas pelos autores.

[Taboada e Das 2013] e [Liu e Zeldes 2019], a partir de *corpora* pré-anotados com o modelo *Rhetorical Structure Theory* (RST) [Mann e Thompson 1987] identificaram uma série de pistas linguísticas e estruturais que serviam de sinalizadores para as relações discursivas previamente identificadas. Ambos os trabalhos organizaram os Sinalizadores Discursivos (SD) em função de suas características (semânticas ou sintáticas, por exemplo), pondo em xeque a ideia de que as relações RST deveriam ser identificadas majoritariamente por meio de Marcadores Discursivos (MD), tidos como preposições e conjunções.

Com base nessa metodologia, [Rodrigues et al. 2023] descreveram SDs para além dos MDs a partir do *corpus* CSTNews [Cardoso et al. 2011]. Tal recurso linguístico-computacional consiste em um conjunto de textos jornalísticos em Português que já havia sido anotado segundo o modelo RST. A RST propõe que um texto coerente é formado por unidades mínimas de discurso (*Elementary Discourse Units* - EDU ou proposições) que desempenham funções retóricas para que o objetivo comunicacional do autor seja atingido. Partindo dessa anotação prévia, os anotadores deste trabalho, por sua vez, identificaram apenas os sinalizadores que consideraram relevantes para caracterizar e/ou indicar determinadas relações, como em (1) - extraído do *corpus* CSTNews.

(1) [A seleção brasileira masculina de vôlei,]^{A} [que é treinada por Bernardinho,]^{B} [venceu a Finlândia por 3 sets a 0.]^{C}

As porções (1a) e (1c) foram conectadas por meio da relação RST *Same-Unit*, indicando que se trata da mesma unidade, já que estão separadas por haver detalhamento informacional em 1B em relação à (1a) por meio da relação *Elaboration*. [Rodrigues et al. 2023] indicaram que a pontuação (no caso, vírgula), a concordância verbal e o encaixamento de outra relação RST poderiam ser utilizadas como pistas para a identificação da relação *Same-unit*.

Esse estudo preliminar resultou em um manual de anotação de SDs em textos jornalísticos [Dantas et al. 2024], em que há, além de instruções, a proposta da primeira taxonomia de sinalizadores de relações RST para o PB. Destaca-se que esse tipo de recurso com explicações, exemplos e instruções objetivas subsidia a decisão dos anotadores diante de fatos novos e/ou já conhecidos [Duran et al. 2022].

Assim, objetivamos neste trabalho relatar as etapas metodológicas e práticas de anotação de SDs no *corpus* CSTNews, além de apontar as primeiras avaliações sobre as decisões tomadas pelo grupo de anotadores. Para tanto, este trabalho está organizado em 5 seções, além desta Introdução. Na Seção 2, destacamos trabalhos relacionados ao

processo de anotação e análise em RST, sobretudo para o PB. Na Seção 3, detalhamos a metodologia de anotação empreendida neste estudo. Na Seção 4 apresentamos os resultados e as discussões correspondentes. Por fim, na Seção 5, tecemos algumas considerações finais.

2. Trabalhos Relacionados

Identificar relações RST por meio de marcas explícitas no texto não é uma tarefa nova, especialmente em PLN para análise de discurso. Os MDs são tidos como conectivos entre porções textuais, sinalizando determinadas relações discursivas, como o "mas" para oposição, por exemplo. A análise das relações discursivas (ou de coerência) está intimamente ligada a descobrir a intenção do autor ao apresentar partes do texto em uma ordem e combinação específicas. Portanto, trata-se de uma tarefa que vai além de identificar os MDs.

A literatura [Marcu 2000, Pardo 2005, Taboada e Das 2013] indica que a identificação de MD, em função da relação RST a que ocorrem, facilita o processamento do texto. Estudos recentes [Das e Taboada 2018, Liu e Zeldes 2019] afirmam que os MDs sinalizam apenas um número restrito de relações dentro de um texto, e sugerem que as relações RST podem ser identificadas por sinais que vão além deles. Como os MDs não marcam explicitamente as relações e não são exclusivos, a noção de SD parece ser mais apropriada do que a de MD nesse contexto.

[Das e Taboada 2018] argumentam que para uma comunicação ser eficaz, é fundamental que as relações sejam interpretadas de maneira relativamente clara, o que requer sinalizadores precisos. Os autores acreditam que as relações de coerência são entidades cognitivas, e, portanto, é possível descobrir como ouvintes e leitores as identificam usando indicadores que auxiliem o processo interpretativo. Utilizando o *RST Discourse Treebank*, os autores realizaram uma anotação detalhada dos SD, resultando no *RST Signalling Corpus* (RST-SC). Eles observaram que pode haver relações sinalizadas por um único sinalizador (como MD, referências pessoais, orações relativas ou dois pontos) ou por combinações de SD (como vírgula + oração no particípio passado, ou construção sintática paralela + cadeia lexical). Quando surgia uma nova instância de um tipo específico de relação, os anotadores consultavam a taxonomia para encontrar o(s) sinalizador(es) mais adequado para aquela instância. Durante o processo de anotação, os autores observaram casos em que não foi possível determinar com precisão o SD que representava uma determinada relação.

Em [Liu e Zeldes 2019] descreve-se um esforço de anotação para ancorar SD a partir de diversas categorias tais como sintática, semântica, gráfico and morfológica. Seus resultados mostraram que, com 11 documentos e 4.732 *tokens*, 923 foram instâncias de SD, o que representou mais de 92% dos sinais discursivos. O tipo semântico representou a maioria dos casos, enquanto as relações discursivas ancoradas por DM corresponderam a apenas cerca de 8,5% dos tokens ancorados.

Quanto à língua portuguesa, [Pardo 2005] foi o precursor em investigar a construção de analisadores discursivos. A partir de um *corpus* de textos científicos e anotado com RST, o autor identificou diversos padrões de análise que especificam os relacionamentos entre as relações retóricas e seus marcadores textuais. Apesar de muitos padrões serem baseados em MD, o autor ressalta que não existe uma relação *sine qua non*

entre MD e as relações que sinalizam, pois uma mesma relação pode ser sinalizada por vários marcadores (por exemplo, a relação *Concession* pode ser sinalizada pelos marcadores "entretanto", "no entanto", entre outros) e um mesmo marcador pode sinalizar várias outras relações (por exemplo, o marcador "porque" pode sinalizar as relações *Cause* e *Result* (volitivas ou não), *Justify*, *Explanation*, entre outras).

Ainda com relação ao português, [Maziero 2016] investigou atributos de organização textual, da morfossintaxe, da sintaxe, da semântica e discurso para construir um analisador discursivo baseado na RST. A partir da análise de *corpora* anotados com RST, o autor aponta que: a) existem relações que apresentam grande subjetividade, tais como as relações *Evidence*, *Justify* e *Explanation*; b) a relação *Same-unit* ocorre apenas no nível intrassentencial, o que é esperado, pois é responsável por ligar proposições quebradas por uma relação de *Parenthetical* ou *Elaboration*, por exemplo; c) algumas relações são mais frequentes no nível intrassentencial do que no inter-sentencial. Após vários experimentos com aprendizado de máquina para identificação das relações discursivas no nível intrassentencial, o autor concluiu que atributos morfossintáticos proporcionaram melhores resultados do que os atributos semânticos e discursivos.

3. Metodologia

A anotação de SDs foi feita a partir do *corpus* CSTNews[1]. Os textos do *corpus* estão organizados em 50 conjuntos, com dois ou três documentos que noticiam o mesmo evento. Por essa característica multidocumento e de redundância, a anotação foi feita apenas no maior texto do conjunto, pois acreditamos que quanto maior for o texto, maior é a chance de encontrarmos mais relações RST e, possivelmente, essas relações ocorram nos outros outros textos da mesma coleção. Nesse caso, foram separados 50 textos para esta tarefa de anotação.

A anotação de SDs foi realizada por meio da ferramenta rstWeb [Zeldes 2016], que é uma plataforma desenvolvida para facilitar a análise e a anotação de textos com base na RST. Essa ferramenta permite aos usuários realizar análises estruturais detalhadas dos textos, identificando proposições e suas relações de coerência conforme proposto pela teoria. Neste trabalho, a taxonomia de SDs, na Figura 1, foi implementada.

Os textos escolhidos foram pré-processados e distribuídos a um grupo de oito pesquisadores. A anotação aconteceu de maneira assíncrona, em que cada anotador recebia semanalmente 3 ou 4 textos. Cada texto foi anotado por três anotadores para que pudéssemos ter uma versão do *corpus* com a decisão sobre a indicação dos SDs por maioria simples.

Para promover discussão e resolução de dúvidas, especialmente sobre casos não previstos pelo manual, foram conduzidas reuniões semanais com o grupo. Além disso, dois dos anotadores, por terem mais experiência com tarefas nesse sentido, nunca ficavam juntos no trio, para que pudessem auxiliar na resolução de dúvidas de maneira assíncrona. Ressalta-se que os anotadores possuíam diferentes formações acadêmicas (linguistas ou cientistas da computação) e com experiências distintas em tarefas de anotação de *corpus*. Por conta disso, foi necessária uma etapa de treinamento para que o grupo se familiarizasse com a taxonomia de SDs e com o modelo RST, além de ter acesso ao manual de

[1]Disponível em: `http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/`

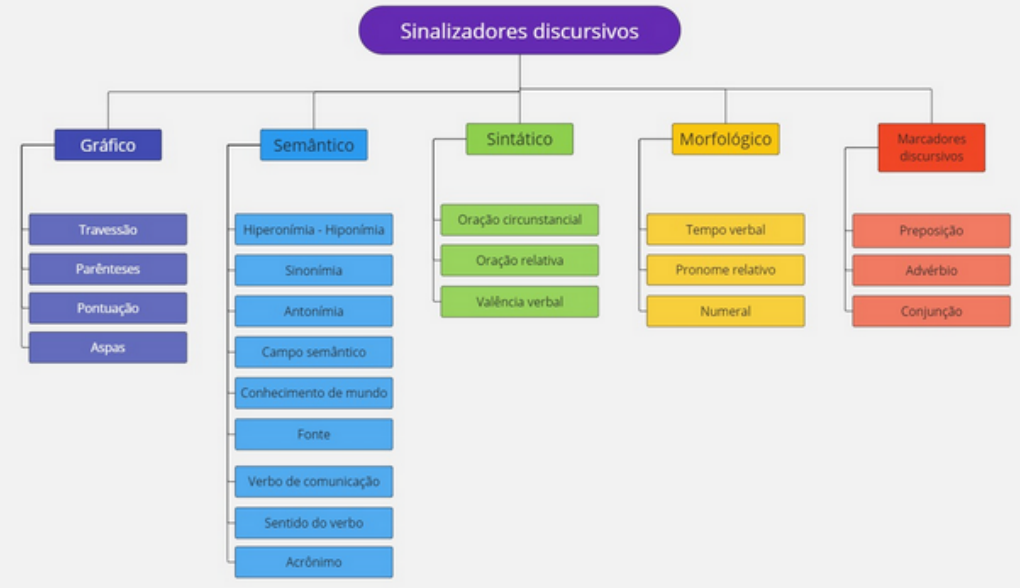

Figura 1. Taxonomia de sinalizadores discursivos proposta por Autores (2023).

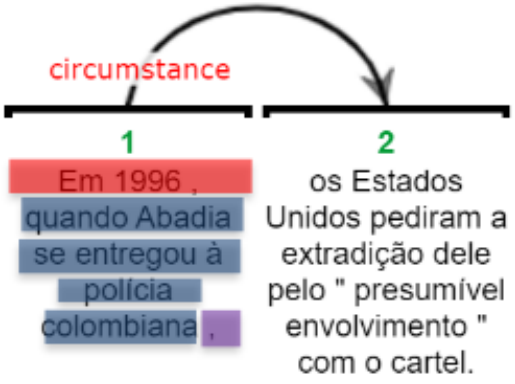

Figura 2. Taxonomia de sinalizadores discursivos proposta por Autores (2023).

anotação para auxiliar em suas decisões.

Na fase de treinamento, foi realizada a anotação de três textos do *corpus*. Em reuniões síncronas, os anotadores puderam corrigir possíveis equívocos e identificar quais unidades do discurso deveriam ser devidamente anotadas, ou seja EDUs que estivessem presentes em um mesmo período sintático.

Na Figura 2, têm-se um exemplo de anotação para a relação *Circumstance*. Essa relação RST deve apresentar uma situação realizável, em que o satélite (EDU 1, provê a situação que é apresentada no núcleo - EDU 2). No exemplo, tem-se que essa relação está sendo sinalizada por meio de Advérbio (vermelho), Oração circunstancial (azul) e Pontuação (lilás).

Três textos de diferentes tamanhos foram anotados por todos os anotadores em distintas fases do processo, com o objetivo de medir periodicamente a concordância do grupo. Esse processo se repetiu a cada 10 conjuntos de textos anotados. Ao final de 2 meses, 47 textos foram anotados.

Trecho anotado com a relação RST Concession	Anotadores	Sinalizadoes indicados
“Tal capacidade de mutação fez escola,	A	“mas” + “,”
mas dificilmente as criaturas saberão	B	“mas” + “,”
superar o criador.”	C	“mas”

Tabela 1. Comparação de anotações.

A concordância foi medida automaticamente a partir de duas abordagens. Na abordagem *gold* observa-se estritamente o que o grupo de anotadores apontaram como sinalizador, sendo, portanto, mais restrita. Já na abordagem *silver* definiu-se um intervalo de cinco janelas (à esquerda e à direita) em relação ao sinalizador-alvo, como demonstrado na Tabela 1.

Na Tabela 1, tem-se um exemplo de como as duas abordagens da concordância foram aplicadas. Os anotadores A e B indicaram os mesmos sinalizadores, ao passo que o sinalizador C indicou apenas um em comum com o grupo. Caso fosse considerada apenas uma análise mais restritiva sobre a concordância, a decisão do anotador C prejudicaria o cálculo, ao passo que numa abordagem mais ampla, sua decisão não traria tantos prejuízos.

Apesar de todos os esforços metodológicos de distribuição de textos, e de todos os anotadores estarem alinhados junto ao modelo teórico e à ferramenta utilizados, é possível que fatores externos à tarefa influenciam na disposição dos anotadores, fazendo-os eventualmente não apenas discordarem sobre um sinalizador, mas também não se atentarem a realizar a indicação adequada. Por conta disso, escolheu-se neste trabalho não apenas realizar uma análise mais restrita sobre a concordância, mas também mais ampla, admitindo-se nesta, sobretudo, a dimensão mais subjetiva da tarefa.

Em ambas as abordagens utiliza-se a medida *Krippendorff Alpha* [Krippendorff 2011]. Trata-se de uma medida que avalia a concordância entre dois ou mais anotadores, o que se encaixa melhor no contexto deste trabalho, já que cada texto foi anotado por três pessoas. O resultado da concordância é medido num intervalo que varia entre -1 e 1, em que valores mais próximos a 1 indicam alta concordância; valores próximos a 0 indicam baixa concordância; e valores próximos a -1 indicam discordância total.

4. Resultados e Discussão

Na Tabela 2, tem-se a média dos resultados das concordâncias *gold* e *silver* da anotação em diferentes etapas do processo. Como dito, na fase de treinamento (*clusters* 1, 2 e 3), os anotadores realizaram uma primeira anotação e, após reunião de alinhamento, fizeram correções. O cálculo da concordância geral (*clusters* 16, 31 e 39) foi feito sobre o mesmo texto anotado por todo o grupo.

Dado que o processo de anotação pode ser longo e complexo, fatores externos aos aspectos linguísticos (como cansaço e diminuição da atenção, por exemplo) pode ter influenciado os anotadores. É possível perceber isso ao comparar as fases de treinamento com as demais, em que as demais sofreram decréscimos discretos. Além disso, outro possível aspecto que pode ter influenciado nesse resultado é a distribuição das relações

	Concordância	
Fase do trabalho	**Gold**	**Silver**
Treinamento	0,581	0,693
Concordânca geral	0,460	0,596
Rodadas de anotação	0,496	0,691

Tabela 2. Resultado da concordância.

RST no *corpus* CSTNews. [Cardoso et al. 2011] apontam que há relações RST que ocorrem apenas uma vez, como *Otherwise*, por exemplo, e outras que aconteceram de maneira predominante, como *Elaboration*, que ocorreu 1,514 vezes. Nesse caso, é possível que, ao se deparar com uma relação RST não prevista na fase de treinamento e, portanto, ausente no manual de anotação, os anotadores enfrentaram dificuldades em indicar possíveis sinalizadores das relações em questão.

Além de uma análise quantitativa, foram feitas observações qualitativas preliminares. Para tanto, durante a anotação, os anotadores realizaram indicações de dúvidas, inconsistências e/ou outras questões em um formulário eletrônico. Ao final de cada semana, todos os apontamentos eram compilados e discutidos entre o grupo para aprimorar o processo. A partir disso, é possível destacar alguns pontos:

a) Considerações sobre o processo de anotação

Em caso de não encontrar uma etiqueta para representar o fenômeno observado, o anotador poderia registrar os *tokens* envolvidos e marcar como CPD (Casos Para Discutir depois). Em discussões e análises preliminares, os anotadores destacaram a intenção de marcar as entidades mencionadas no segmento textual. [Das e Taboada 2018], por sua vez, descrevem que os anotadores discordavam bastante entre entidade e tipos semânticos, ou seja, enquanto um anotador seleciona entidade como o sinal relevante para uma certa relação, o outro anotador a anota como sendo semântica. Os autores observaram que muitos dos atributos de entidade e características semânticas na verdade se sobrepõem. Dessa forma, essa dificuldade acontece também para a língua inglesa.

Assim como [Liu 2019, Das e Taboada 2018] relatam, também observamos no corpus de estudo várias relações que não tinham um token explícito para servir de sinalizador. Esses casos foram registrados como CPD. Por outro lado, as primeiras análises revelaram que alguns SD são altamente indicativos, enquanto outros são genéricos ou ambíguos. Assim, para obter uma compreensão mais precisa, é necessário considerar os contextos ao redor dos SD para desambiguá-los.

b) Considerações sobre dificuldades e limitações encontradas

A anotação das relações RST é um processo que se baseia na interpretação do analista. Assim, a depender dessa interpretação serão indicadas determinadas relações RST em detrimento de outras, resultando, então, em diferentes sinalizadores para essas relações. Neste estudo, a identificação de SDs foi feita por um grupo majoritariamente diferente de quem fez a anotação RST, com uma distância temporal considerável entre as duas tarefas. Esse fato, portanto, pode ter sido um dificultador para o grupo que fez a indicação dos sinalizadores.

Além disso, os anotadores destacaram que algumas relações RST utilizadas no

CSTNews são mais difíceis de interpretar, e consequentemente, torna-se um desafio apontar SD específicos, como exemplificado em (2).

> (2) (...) [com o Programa Fome Zero, conseguiu atingir o primeiro ponto das Metas do Milênio - erradicar a fome -, com dez anos de antecedência,]^A [reduzindo em mais da metade a pobreza extrema.]^B

O trecho (2b) em relação ao trecho (2a) apresenta a relação *Volitional result*, ou seja, o resultado ocasionado foi não intencional. Nesse caso em específico, os anotadores indicaram que o sentido do verbo "reduzindo" seria o indicativo do resultado, porém sem menção ao aspecto volitivo. Destaca-se que a maioria dos *rols* de relações para outras línguas não preveem diferença nesse aspecto.

Outro aspecto que parece ter apresentado dificuldade aos anotadores foi o fato de o manual de anotação ter sido desenvolvido com base no estudo de [Rodrigues et al. 2023] e os resultados da fase de treinamento. Como citado, relações e sinalizadores que não estavam previstos e que ocorreram ao longo do *corpus* podem ter ocasionado certos equívocos entre os anotadores. Ademais, o fato de o manual indicar certa correlação entre SDs e relações pode ter condicionado o olhar dos anotadores, como demonstrado em (3).

> (3) [nesta terça deve se encontrar com o relator do caso na Câmara, deputado José Carlos Araújo (PR-BA)]^A [para tratar do assunto.]^B

De acordo com [Cardoso et al. 2011], a sentença entre (3a) e (3b) é de *Purpose*. O manual de anotação de SDs utilizou esse exemplo e indicou que a preposição "para" pode ser utilizada para identificar essa relação. Entretanto, o objetivo entre os segmentos pode também ser evidenciado por meio de "oração final" presente em (3b). Nesse caso, é possível que os anotadores tenham sido condicionados a partir de determinados pressupostos sobre as relações, ainda que tenham sido estimulados a indicarem em formulário eletrônico outros possíveis SDs e definições não previstos no manual.

Por fim, cabe pontuar que no repositório *online* do projeto de pesquisa "RST além dos marcadores discursivos"[2] disponibilizamos para consulta o *corpus* com a versão unificada entre os anotadores, a anotação de SDs e a planilha completa da concordância dos anotadores.

5. Considerações Finais

Neste trabalho buscamos detalhar a metodologia empregada na identificação de SDs em textos jornalísticos a partir da taxonomia proposta por [Dantas et al. 2024]. Destacamos que um estudo com essa abordagem em PB ainda não havia sido realizado, ao contrário do que já ocorre em outros idiomas, especialmente o inglês.

Os resultados relatados podem subsidiar outras análises em estudos futuros. Um desses estudos se concentra na investigação quali e quantitativa da correlação entre SDs e as relações RST, algo já iniciado por [Rodrigues et al. 2023] e tal como outros trabalhos fizeram [Liu 2019, Das e Taboada 2018, Pardo 2005]. Outro estudo será em relação à concordância de aspectos da anotação, como tipos (sintático e semântico, por exemplo) e subtipos (pronome relativo e conhecimento de mundo, por exemplo) dos sinalizadores. Ao final desses estudos será possível fazer o levantamento da distribuição dos SDs no

[2]Disponível em `https://sites.google.com/view/rst-poetisa/`

corpus, bem como observar quais são mais ou menos consensuais entre os anotadores.

Dados os apontamentos críticos realizados sobre as limitações identificadas, destaca-se que este trabalho apresenta potencial de servir de diretriz de investigações de análises sobre as relações RST e seus SDs e aprimoramento de ferramentas e recursos para anotação de *corpus*. Tais aspectos são de extrema importância ao alargar a anotação a escalas maiores buscando não apenas ampliar a quantidade de textos, mas também diversificar os gêneros textuais a serem considerados.

6. 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.

.

Referências

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., Nunes, M. d. G. V., e Pardo, T. A. (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*, pages 88–105.

Dantas, E., Bárbara, L. d. J. S., Pereira, M. A., Gama, N. S., Almeida, T. J. A., Souza, J. W. d. C., Cardoso, P. C. F., e Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. Série de Relatórios Técnicos do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Das, D. e Taboada, M. (2018). RST signalling corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52:149–184.

Duran, M. S., Nunes, M. d. G. V., Lopes, L., e Pardo, T. A. S. (2022). Manual de anotação como recurso de processamento de linguagem natural: o modelo universal dependencies em língua portuguesa. *Domínios de Lingu@ gem*, 16(4):1608–1643.

Hovy, E. e Lavid, J. (2010). Towards a 'science'of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1):13–36.

Krippendorff, K. (2011). Computing krippendorff's alpha-reliability.

Liu, Y. (2019). Beyond the Wall Street Journal: Anchoring and comparing discourse signals across genres. *arXiv preprint arXiv:1909.00516*.

Liu, Y. e Zeldes, A. (2019). Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Society for Computation in Linguistics*, 2(1).

Mann, W. C. e Thompson, S. A. (1987). *Rhetorical Structure Theory: A theory of text organization*. University of Southern California, Information Sciences Institute Los Angeles.

Marcu, D. (2000). The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational linguistics*, 26(3):395–448.

Maziero, E. G. (2016). *Análise retórica com base em grande quantidade de dados*. PhD thesis, Universidade de São Paulo.

Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. PhD thesis, Universidade de São Paulo.

Pedro, W. e Vale, O. (2018). Comentcorpus: o uso de mecanismos linguísticos na detecção de ironia e sarcasmo para o português do Brasil em um corpus opinativo. *Linguística de corpus: perspectivas. Porto Alegre: Instituto de Letras da Universidade Federal do Rio Grande do Sul*, pages 19–40.

Pustejovsky, J. e Stubbs, A. (2012). *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. O'Reilly Media, Inc.

Rodrigues, R., Souza, J. W., e Cardoso, P. C. F. (2023). Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTnews. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 249–257. SBC.

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

Zeldes, A. (2016). rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 1–5.