Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários

Naira Silva Gama[1], Jackson Wilke da Cruz Souza[1,2]

[1]Instituto de Ciência, Tecnologia e Inovação
Universidade Federal da Bahia (UFBA), Camaçari/BA

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

nairagama@ufba.br, jackcruzsouz@gmail.com

Abstract. *Rhetorical Structure Theory (RST) is a theory that points out the rhetorical structure present in the text. Descriptive RST works are mostly dedicated to formal textual genres, resulting in a scarcity of works that explore the theory in User-Generated Content (UGC) texts. Therefore, the objective of this work is to investigate discursive signals (SDs) of RST relationships in UGC texts, more specifically in tweets from the financial market. To this end, 180 tweets were randomly selected from the DANTE-stocks corpus, which were analyzed manually, identifying the flags of the RST relationships previously noted. As a result, the typology of flags for Portuguese was updated with SDs specific to UGC texts.*

Resumo*. A Rhetorical Structure Theory* (RST) *é uma teoria que aponta a estrutura retórica presente no texto. Os trabalhos descritivos de RST majoritariamente se dedicam a gêneros textuais formais, resultando na escassez de trabalhos que explorem a teoria em textos de Conteúdo Gerado por Usuário (CGU). Assim, objetiva-se neste trabalho investigar sinalizadores discursivos (SDs) das relações RST em textos de CGU, mais especificamente em tweets do mercado financeiro. Para tanto, foram selecionados aleatoriamente 180 tweets do corpus DANTE-stocks, que foram analisados manualmente identificando os sinalizadores das relações RST anotadas previamente. Como resultado, atualizou-se a tipologia de sinalizadores para o Português com SDs específicos de textos de CGU.*

1. Introdução

A *Rhetorical Structure Theory* (RST) [Mann e Thompson 1988] é uma teoria que possibilita a visualização da estrutura retórica de um texto a partir de um conjunto de relações semânticas, e, com efeito, facilita análises textuais. [Souza, Cardoso e Rodrigues 2023] destacam que esse modelo teórico vem sendo utilizado em estudos linguísticos, diálogo e multimídia e análise de discurso, argumentação e escrita, além de aplicações de Processamento de Linguagem Natural (PLN), como avaliação de textos, sumarização e tradução automática.

Majoritariamente, as relações semânticas do modelo RST vem sendo identificadas com base em marcadores discursivos. Porém, em estudos recentes [Souza, Cardoso e Rodrigues 2023; Dantas *et al.* 2024] realizados no *corpus* CSTNews [Cardoso *et al.* 2011], chegou-se à proposição de uma série de sinalizadores linguísticos e estruturais que servem como potenciais indicadores de relações RST. Os resultados corroboram investigações internacionais na mesma temática [Taboada e Das 2013; Das e Taboada 2018; Liu e Zeldes 2022], além de indicarem particularidades que ocorreram apenas em língua portuguesa.

Ao final do estudo, Dantas *et al.* (2024) propuseram uma taxonomia em que os sinalizadores estão organizados em tipos e subtipos, a saber: (i) *Marcadores discursivos*, tendo como subtipos Preposição, conjunção e advérbio; (ii) *Morfológico*, tendo como subtipos Tempo verbal, Pronome relativo e Numeral; (iii) *Sintático*, tendo Orações relativas, Orações circunstanciais e Valência verbais como subtipos; (iv) *Semântico*, tendo como subtipos as relações Hiperonímia/Hiponímia, Sinonímia e Antonímia, Campo semântico, Conhecimento de mundo, Fonte (da informação), Verbo de comunicação, Sentido do verbo e Acrônimos; e (v) *Gráfico*, com os subtipos Travessão, Parênteses, Pontuação e Aspas.

Por terem analisado um *corpus* de texto jornalístico, o registro linguístico sob observação é o formal. Trabalhos da literatura recente indicam que diante de outros registros e modalidades da língua podem gerar outros sinalizadores discursivos [Antonio 2017; Fachada 2019; Pecuch 2021]. Assim, neste trabalho foi conduzido um estudo preliminar para encontrar sinalizadores específicos de textos de CGU. Para tanto, partiu-se de dois grandes objetivos: (i) evidenciar a amplitude em textos de CGU dos sinalizadores discursivos identificados por [Dantas *et al.* 2024]; (ii) destacar sinalizadores em potencial desse novo gênero textual sob observação.

Para tanto, este texto foi organizado em 4 seções, além desta Introdução. Na Seção 2, apresentam-se a metodologia e o *corpus* utilizado a partir dos principais objetivos deste trabalho. Na Seção 3, apresentam-se os resultados, com uma análise preliminar do conjunto de dados, os sinalizadores identificados, a uma breve comparação entre a tipologia proposta e as de outros trabalhos. Por fim, na Seção 4, as considerações finais são expostas juntamente com os trabalhos futuros.

2. Metodologia

Neste trabalho, foi utilizado o *corpus* DANTE-stocks [Di Felippo *et al.*, 2021] como conjunto de análise. Tal *corpus* é composto por postagens/*tweets* do domínio do mercado financeiro (mais especificamente sobre ações do índice Ibovespa), retirados da rede social X (antigo *Twitter*). Contabilizam-se 4.517 postagens em português, coletadas em 2014. Segundo os autores, o *corpus* está organizado em função dos seguintes critérios: (i) *Simplificação de código*, como Ausência de hífen; (ii) *Abreviação*, como Contração e Acrônimo; (iii) *Expressão de sentimento*, como Alongamento de pontuação; (iv) *Influência de língua estrangeira*, como Formação verbal; (v) *Marca de oralidade*, como Coloquialismos; (vi) *Elementos metalinguísticos* (do Twitter), como Hashtag e Menção; (vii) *Fenômeno do domínio* (Ibovespa), como Ticker e Cashtag.

Para tanto, partiu-se da classificação automática do DANTE-stocks promovida por Ramos e Souza [2024] e Pereira e Souza [2024], que indicaram três classes de textos, a saber postagens bem, mal e medianamente estruturadas, classificadas em função de aspectos semânticos, coesivos e de coerência nas construções linguísticas. A partir dessa classificação, escolheu-se aleatoriamente 180 postagens das três classes identificadas para este estudo preliminar. Em seguida, foi conduzida uma análise semi-automática utilizando a ferramenta rstWeb [Zeldes 2016], em que um anotador sem conhecimentos prévios sobre o mercado financeiro, buscou indicar possíveis relações RST nos *tweets*, bem como seus sinalizadores em potencial, conforme ilustrado na Figura 1.

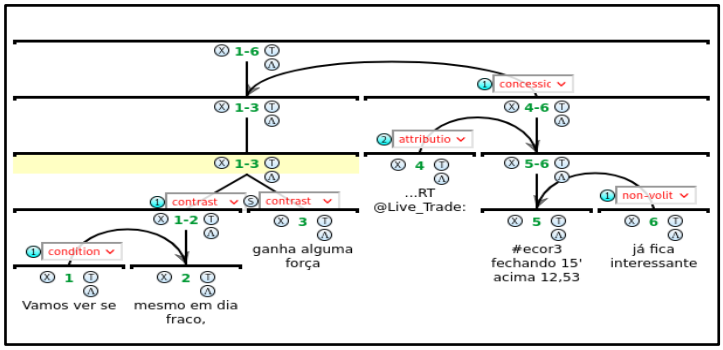

Figura 1. Postagem classificada como "medianamente estruturada".

Na Figura 1, tem-se um exemplo de postagem anotada com relações RST e seus possíveis sinalizadores. De acordo com a taxonomia proposta por Dantas *et al.* (2024), na relação *Condition*, presente entre as unidades mínimas da construção discursiva, tidas como as *Elementary Discourse Units* (EDUs) 1 e 2, tem-se o "se" como um Marcador Discursivo de subtipo Conjunção. Na relação *Contrast*, entre as EDUs 1-2 e 3, foram anotados "fraco" e "força" como Semântico e de subtipo Antonímia. Na relação *Non-Volitional-Result*, "já fica interessante" foi anotado como Sintático de subtipo oração circunstancial. Por fim, na relação *Concession*, o "#ecor3 fechando 15' acima 12,53" também foi anotado como sintático de subtipo oração circunstancial. Além disso, foram identificados novos sinalizadores. Na EDU 4, tem-se uma relação de *Attribution* sinalizada por "RT", que foi anotado com um novo tipo, a saber, Gênero (textual), e "@Live_Trade", anotado como Sintático de subtipo Citação.

3. Resultados e discussão

No Quadro 1, apresentam-se os sinalizadores em função das relações RST.

Sinalizador		**Exemplo**	**Relação RST**
Tipo	**Subtipo**		
Gênero	Link	(1) @BorisBrianCasoy Petrobras PETR4 Lula capitalizou em 2010 a R$ 26,30. Últimos 6 meses… Março na bacia das almas! *http://t.co/l00d2g2AtO*	*Elaboration*
	RT	(2) Vamos ver se mesmo em dia fraco, ganha alguma força … *RT* @Live_Trade: #ecor3 fechando 15' acima 12,53 já fica	*Attribution*
Sintático	Injunção	(3) #ecor3 *segue* seu caminho rsrsr	*Elaboration*
Semântico	Emoção	(4) #ecor3 segue seu caminho *rsrsr*	Evaluation
	Entidade	(5) *@Live_rade* MT, tá bem vale5 tb? Grato	
	Citação	(6) *@BorisBrianCasoy* Petrobras PETR4 Lula capitalizou em 2010 a R$ 26,30. Últimos 6 meses… Março na bacia das almas! http://t.co/l00d2g2AtO	*Attribution*
Gráfico	Caracter especial	(7) *Elétricas lideram altas e baixas com ajuda do governo, Vale renova mínima do ano: Ecorodovia… http://t.co/EkCKzePhxU* #infomoney #vale5	*Elaboration*
	Símbolos	(8) 11/04/14 - 17:18: *Maiores* Altas:- KROT3 4,17% R$48,75, CPLE6 3,62% R$30,94, AEDU3 3,44% R$13,55, PETR4 2,93% R$16,14, ALLL3 2,90% R$8,31.	*Elaboration*

Quadro 1. Exemplos de novos sinalizadores para textos de CGU.

Do Quadro 1, em (1), o *link* é um sinalizador da relação *Elaboration*, pois indica uma informação extra, que pode ser acessada como complemento da postagem. Em (2), RT

sinaliza a relação *Attribution*, pois indica um "retweet", ou seja, compartilhamento de uma postagem anterior. Em (3), tem-se *Elaboration* a oração imperativa indica um complemento informacional. Em (4), encontra-se uma *Evaluation*, pois o "rsrsr" indica uma emoção em relação ao que foi dito anteriormente. Em (5), observa-se uma relação de *Attribution*, visto que o autor da postagem menciona "@Live_Trade" e insere na estrutura sintática da postagem, caracterizando uma Entidade. Já em (6), também se tem uma *Attribution*, mas, nesse caso "@BorisBrianCasoy" está sendo citado na postagem, não integrando a estrutura sintática. Em (7), o caracter especial apresenta-se uma *Elaboration*, indicando um complemento à postagem. Por fim, em (8) tem-se relações *Elaboration*, visto que as indicações porcentuais e os valores monetários indicam uma informação complementar do que foi citado anteriormente.

Analisando os novos sinalizadores discursivos, nota-se algumas diferenças em relação à tipologia proposta por Dantas *el al.* (2024), como a adição do tipo Gênero e de novos subtipos, como Símbolos, Emoções e Injunção. Com relação aos trabalhos de Taboada e Das (2013) e Liu e Zeldes (2022), há aproximação quanto a proposição do tipo Gênero. Destaca-se ainda que tais trabalhos observaram a combinação entre os sinalizadores para a indicação de relações RST, algo ainda a ser realizado em pesquisas futuras.

4. Considerações finais

Observando as anotações dos textos do *corpus* do CSTNews [Cardoso et al., 2011] e do DANTE-stocks [Di Felippo *et al.* 2021], conclui-se que cada gênero textual pode apresentar características particulares. Nesse sentido, devem ser consideradas tipologias complementares uma à outra.

Os objetivos deste estudo foram alcançados, pois foi investigada a ocorrência de sinalizadores linguísticos-estruturais em *corpora* de textos de CGU com o modelo RST. Para tanto, foram utilizados métodos semiautomáticos para a identificação dos sinalizadores nas relações e, como resultado, apresentou-se uma tipologia complementar com novas categorias de tipo e subtipo de sinalizadores. Em trabalhos futuros, pretende-se ampliar a anotação do *corpus*, verificando possíveis novos sinalizadores, e consequentemente, a atualização da tipologia proposta inicialmente neste trabalho. Além disso, estes dados poderão ser utilizados em outros trabalhos que estejam relacionados ao mercado financeiro, as redes sociais ou qualquer outro ligado a linguística computacional.

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. Agradecemos também ao apoio e suporte financeiro do Conselho Nacional de Desenvolvimento Científico e Tecnológico.

5. 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. DOI: https://doi.org/10.1590/0102-445025798334674077.

Cardoso, P. C., Maziero, E. G., Jorge, M. L. C., Seno, E. M., Di Felippo, A., Rino, L. H. M., ... Pardo, T. A. (2011, October). 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).

Dantas, E., Bárbara, L.J.S., Pereira, M.A., Gama, N.S., Almeida, T.J.A., Souza, J.W.C., Cardoso, P.C.F., Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. São Carlos: Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. Disponível em https://repositorio.usp.br/item/003207370

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

Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L. S., da Silva, E. H., Roman, N. T., Pardo, T. A. (2021). Descrição preliminar do corpus dantestocks: Diretrizes de segmentação para anotação segundo universal dependencies. In *Anais do XIII Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana* (pp. 335-343).

Fachada, B. (2019). "Mas" em artigos de opinião: valores e relações retóricas. elingUP: *Revista Eletrónica de Linguística dos Estudantes da Universidade do Porto*, 8(1), p. 108-122.

Liu, Y., Zeldes, A. (2019). Discourse relations and signaling information: Anchoring discourse signals in RST-DT. *Society for Computation in Linguistics*, *2*(1), 314-317.

Mann, W.C., Thompson, S.A. (1988) Rhetorical Structure Theory: Toward a functional theory of text organization. *Text-interdisciplinary Journal for the Study of Discourse*, v.8, n.3, p. 243-281.

Pecuch, G. (2021). A relação retórica de elaboração sinalizada pelo marcador discursivo mas em aulas e em entrevistas orais. *Letras Escreve*, *11*(1), 43-57.

Ramos, I.V.M., Souza, J.W.C. (2024). Classificação automática de textos de User-Generated Content utilizando Aprendizagem de Máquina Supervisionado. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.

Pereira, M.A., Souza, J.W.C. (2024). Subsídios Linguísticos para classificação automática de textos de User-Generated Content. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC.

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

Taboada, M., & 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 (pp. 1-5).