Anotação *Enhanced Rhetorical Structure Theory* em *textos de User-Generated Content*

Mateus Araújo Pereira[1], Jackson Wilke da Cruz Souza[2]

[1]Instituto de Letras (ILUFBA)
Universidade Federal da Bahia (UFBA) - Salvador/BA

[2] Instituto de Ciência, Tecnologia e Inovação (ICTI)
Universidade Federal da Bahia (UFBA) - Salvador/BA

mteuspereiraa@outlook.com, jackcruzsouza@gmail.com

Abstract. *This study examines the application of Enhanced Rhetorical Structure Theory (eRST) to User-Generated Content (UGC), specifically tweets within the financial market domain. Using the rstWeb tool, the annotation was conducted based on the Dante-Stocks corpus. The results highlight the recurrence of specific eRST relations depending on the tweets' semantic-syntactic structure and suggest possible adaptations of annotation guidelines originally developed for the English language.*

Resumo. *Este estudo analisa a aplicação da Enhanced Rhetorical Structure Theory (eRST) em textos de User-Generated Content (UGC), especificamente tweets relacionados ao mercado financeiro. Com o auxílio da ferramenta rstWeb, a anotação foi realizada a partir do corpus Dante-Stocks. Os resultados evidenciam a recorrência de determinadas relações eRST, a depender da estruturação semântico-sintática dos tweets, além de apontarem possíveis adaptações das diretrizes de anotação originalmente formuladas para a língua inglesa.*

1. Introdução

O crescimento da comunicação digital, especialmente em plataformas *online*, transformou a forma como os usuários produzem e compartilham conteúdo textual. Os enunciados vinculados ao ambiente digital são marcados pela brevidade, espontaneidade, fragmentação textual e o uso de estruturas não convencionais, comumente acompanhada de elementos multimodais - *emojis e hashtags* - como corrobora Sanguinetti *et al.* (2022). Essas características apontam para uma dificuldade aos modelos tradicionais de análise textual, demandando abordagens teóricas e metodológicas adaptadas às particularidades desse gênero textual (GT). De acordo com Sanguinetti *et al.* (2022), os estudos em Processamento de Linguagem Natural (PLN), têm buscado ferramentas para acompanhar essa transformação.

Entre as abordagens teóricas para a análise da estrutura discursiva, destaca-se a *Rhetorical Structure Theory* (RST), proposta por Mann e Thompson (1987), é um modelo que descreve o discurso em unidades (*Elementary Discourse Units* – EDUs) hierarquicamente organizadas por meio de relações retóricas. Tais unidades podem ser classificadas ora como núcleo (quando apresentam a informação principal), ora como satélite (quando apresentam informações complementares ou subjacentes aos núcleos). Essas relações são responsáveis por estabelecer vínculos de coesão e coerência entre as partes do discurso. Embora a RST tenha sido idealizada para a análise de gêneros formais, como os textos jornalísticos e acadêmicos, seu uso vem se expandindo para contextos discursivos mais dinâmicos e informais. Pereira e Souza (2024) têm explorado a

aplicabilidade da RST nesse tipo de discurso, como é o caso de textos gerados por usuários – os chamados *User-Generated Content* (UGC). Krumm, Davies e Narayanaswami (2008) definem UGC, como quaisquer conteúdos desenvolvidos por usuário em ambientes *online*, integrando um conjunto de formatos, como textos, fotos, vídeos, comentários em fóruns ou redes sociais.

Nesse contexto, este estudo é motivado pela escassez de pesquisas voltadas à aplicação da RST em textos de UGC, especialmente para a língua portuguesa. A literatura existente concentra-se em textos formais (p.ex. textos jornalísticos), contribuindo, assim, para uma lacuna teórica no que se refere a outros textos. No âmbito da RST, textos de UGC demandam estratégias específicas quanto à segmentação de EDUs, identificação das relações RST e de seus possíveis sinalizadores, tornando, assim, a tarefa de anotação desafiadora frente aos critérios tidos como *status quo* da RST.

Embora ainda não existam estudos que adotem exatamente a mesma abordagem proposta nesta pesquisa, alguns trabalhos apresentam proximidades metodológicas e teóricas, que foram utilizados como referências e consultas ao longo do desenvolvimento deste trabalho. Esses estudos forneceram direções, mostrando a possibilidade de adaptação do modelo teórico da RST na análise de textos informais e dinâmicos, como os encontrados nas redes sociais.

Nesse contexto, Zeldes *et al.* (2025) propõe uma versão enriquecida da teoria, nomeando-se *Enhanced Rhetorical Structure Theory* (eRST), em que são discutidas novas diretrizes de segmentação, novas relações de sentido e estratégias de sinalização dessas relações. Com isso, a eRST oferece uma abordagem possível de representação da organização discursiva em textos variados, como os de UGC, sem perder a funcionalidade em aplicações computacionais.

Diante desse cenário, o objetivo desta pesquisa foi investigar de que modo a eRST pode ser aplicada em *tweets* do domínio do mercado financeiro. Para tanto, utilizou-se o *corpus* Dante-Stocks [Di Felippo *et al.*, 2021], composto por mais de 4.000 *tweets* relacionados ao mercado financeiro brasileiro. Dentre esses, 60 *tweets* foram selecionados para a anotação, focando na decisão tomada sobre a segmentação das EDUs e da proposição das relações e-RST.

Este trabalho está organizado em quatro seções, além desta Introdução. Na Seção 2**,** descreve-se a metodologia adotada para a anotação dos *tweets*, detalhando o funcionamento da ferramenta utilizada no processo, os critérios de seleção dos dados e a dinâmica entre os anotadores. Na Seção 3, são discutidos os resultados, incluindo as características estruturais mais recorrentes, os desafios enfrentados durante o processo de anotação e as ideias teóricas observadas. Por fim, nas considerações finais, discutimos os limites do modelo atual, a relevância da dimensão pragmática na análise desses textos e apontamentos sobre trabalhos futuros.

2. Metodologia

Por conta dos objetivos traçados nesta pesquisa, utilizou-se o *corpus* Dante-Stocks [Di Fellipo *et al.*, 2021], composto por pouco mais de 4.000 *tweets* extraídos da plataforma X/Twitter, todos vinculados ao domínio do mercado financeiro brasileiro. A partir desse conjunto, foram extraídos 60 *tweets* divididos igualmente entre as categorias *Bem estruturado* (BE), *Mal estruturado* (ME) e *Mediamente estruturados* (MDE), conforme propõe Pereira e Souza (2024). O processo de anotação foi conduzido por dois anotadores, previamente preparados quanto ao modelo eRST [Zeldes *et al.*, 2025] e familiarizados com as características do referido gênero textual. Para a anotação das relações discursivas

foi utilizada a ferramenta rstWeb [Zeldes, 2016], desenvolvida para facilitar tanto a segmentação dos textos em EDUs, quanto a atribuição de relações retóricas entre essas unidades.

Embora a anotação síncrona possa apresentar vieses decorrentes da dinâmica do grupo e da pressão para alcançar um possível consenso entre os anotadores (mais e/ou menos experientes), que pode levar à conformidade e reduzir a diversidade de interpretações, optou-se no presente estudo por utilizá-la. Essa estratégia metodológica foi adotada com o objetivo de favorecer sanar imediatamente dúvidas, interpretações e percepções quanto à identificação e classificação da anotação eRST. Assim, quando houve dúvidas, os anotadores discutiram a fim de chegar num consenso, buscando promover consistência à anotação. Além disso, foram utilizadas as diretrizes de anotação de Zeldes *et al.* (2025), fazendo-se adaptações e proposições pertinentes, já que o gênero textual foco deste trabalho não estava contemplado nos apontamentos feitos pelos autores.

3. Resultados e Discussões

Com relação à segmentação das EDUs, foram identificadas situações em que a segmentação textual e estrutural do *tweet* comprometeram a aplicação direta das relações previstas pela eRST. Esses casos foram enunciados que apresentaram (i) *construções fragmentadas*, em que o conteúdo fazia menção a outra postagem; (ii) *ausência de conectores explícitos*, em que as relações ora interpretadas apresentavam aspectos inovativos frente às diretrizes de anotação; (iii) *presença de elementos multimodais*, em que usos específicos de textos de UGC e do próprio GT podem ser utilizados como sinalizadores discursivos não convencionais de relações eRST. Nesses casos, ignoraram-se critérios sintáticos previstos nas diretrizes da eRST. Tais critérios foram pautados em GTs cujo registro linguístico é o culto.

Em "[RT]**1** [[@joanarauhl :]**2** [a joana ta viciadinha num jogo]**3**]**2-3**", por exemplo, há duas relações que demonstram a atribuição do conteúdo veiculado a uma fonte. Entre as EDUs 1 e 2-3, a relação é *Attribution-Negative*, já que a EDU 1 não é a fonte do discurso; ao passo que entre as EDUs 2 e 3 há a relação *Attribution-Positive*, pois o conteúdo da EDU 3 tem como fonte a entidade que está mencionada na EDU 2. Entretanto, em nenhuma das EDUs há marcas sintáticas usuais da norma culta sobre a atribuição de conteúdo (como verbos discendi, por exemplo). Além disso, compreendendo a dupla função de atribuição de conteúdo, as EDUs 1 e 2 foram separadas, em unidades distintas, rompendo, novamente, com a noção de construção oracional de manuais de anotação RST. Neste trabalho, é possível que decisões como essas tenham sido tomadas com mais facilidade porque a representação arbórea da eRST está atrelada a unidades discursivas e não a unidades lexicais.

Notou-se a recorrência de estruturas discursivas similares entre diferentes tweets, mesmo com conteúdo diferente. Nesse caso, este foi outro quesito que facilitou o processo de anotação, já que a decisão sobre uma estrutura reiterada tende a convergir para uma maior concordância em relação a estruturas pouco frequentes no corpus.

De acordo com a Figura 1, considerando a soma entre as classes BE, ME e MDE, as relações mais recorrentes no *corpus* foram *Elaboration-Additional* (44) e *Organization-preparation* (43), *Attribution-positive* (34) e *Joint-List* (29). Esse dado corrobora os apontamentos de Pastor, Oostdijk e Larson (2024), que defendem que são relações que ocorrem mais em textos cuja construção seja fragmentada e informal, tal como os de UGC. Destaca-se que algumas relações (como *Contingency-Condition* e *Explanation-Justify*) aconteceram apenas em um dos três tipos de texto, ao passo que outras relações (*Adversative-anthitesis* e *Mode-Means*) não ocorreram em um dos tipos. Em especial, a relação *Joint* foi proeminente em enunciados paratáticos, em que as ideias são organizadas por justaposição ou inferência, sem o uso de conectores explícitos. Outro aspecto importante dos textos analisados é que são provenientes de um GT que garante explicitude de opinião, o que justifica a ocorrência da relação *Organization-Phatic*, ainda que, em alguns casos, ao final dos *tweets*.

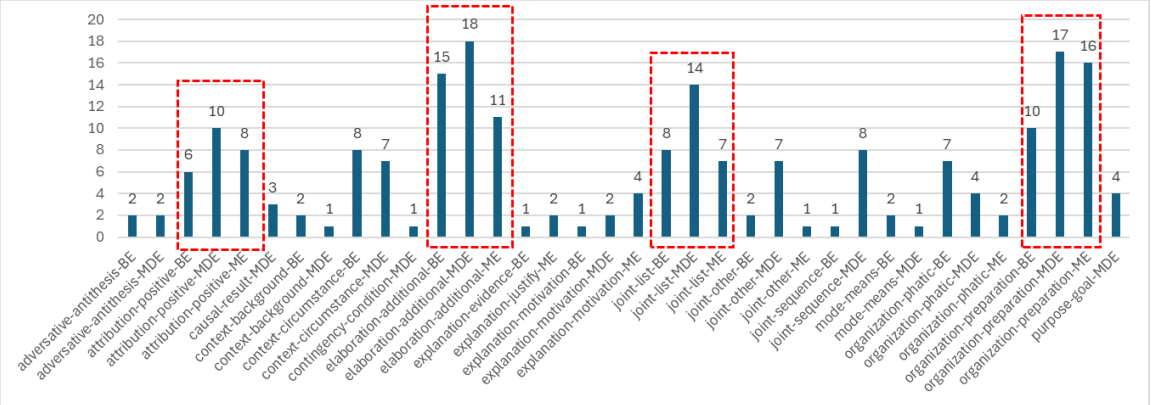

Figura 2. Distribuição das relações eRST na amostra do *corpus* Dante-stocks.

5. Considerações finais

A anotação RST em textos de UGC, especialmente em *tweets*, demonstrou tanto a aplicabilidade da teoria quanto suas limitações diante de características desse GT. Os resultados apontam que, apesar de a recorrência de determinadas estruturas discursivas ter facilitado a anotação e contribuído para a consistência do processo, também foram identificadas situações em que as relações da eRST não se ajustaram harmoniosamente à organização sintática dos *tweets*, exigindo adaptações no processo de anotação.

Ainda, destaca-se a importância da decisão metodológica de realizar uma anotação síncrona. Tal decisão possibilitou a resolução imediata de dúvidas e divergências, promovendo decisões colegiadas mais criteriosas e coerentes, lidando com os desafios interpretativos dos textos, sobretudo naqueles de baixa estruturação formal.

Cabe ainda pontuar que apesar de a identificação de EDUs e de relações retóricas serem tarefas distintas, na verdade, são tarefas que ocorrem concomitantemente, ainda que não explicitamente. Isso se deve ao fato de o anotador, ao separar as EDUs, de certa forma, já interpreta e prevê possíveis relações entre as unidades.

Nesse contexto, em trabalhos futuros, prevê-se a avaliação automática da concordância sobre a segmentação e escolha das relações eRST, a ampliação da anotação e a versão de um manual de anotação em português, contemplando reflexões sobre textos de UGC.

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. Além disso, agradecemos ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e a Fundação de Amparo à Pesquisa do Estado da Bahia (FAPESB) pelo financiamento e suporte.

Referências

Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., & Pardo, T. (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). Porto Alegre: SBC. doi: https://doi.org/10.5753/stil.2021.17813

Krumm, J., Davies, N., & Narayanaswami, C. (2008). "User-generated content". In IEEE

Pervasive Computing, 7(4), 10–11. https://doi.org/10.1109/MPRV.2008.85

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

Pastor, M.,Oostdijk, N., & Larson, M. (2024). "The Contribution of Coherence Relations to Understanding Paratactic Forms of Communication in Social Media Comment Sections". https://hal.science/hal-04536610v1/document

Pereira, M., & Souza, J. (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, (pp. 429-433). Porto Alegre: SBC. doi: https://doi.org/10.5753/stil.2024.245132

Sanguinetti, M., Bosco, C., Cassidy, L., Çetinoğlu, Ö., Cignarella, A. T., et al. (2022). Treebanking user-generated content: A UD based overview of guidelines, corpora and unified recommendations. Language Resources and Evaluation, 57(2), 493–544. https://doi.org/10.1007/s10579-022-09581-9

Zeldes, A., Aoyama, T., Liu, Y. J., Peng, S., Das, D., & Gessler, L. (2025). eRST: A signaled graph theory of discourse relations and organization. Computational Linguistics, 51 (1), 23–72. Doi: https://doi.org/10.1162/coli_a_00538

Zeldes, A. (2016). "rstWeb - a browser-based annotation interface for Rhetorical Structure Theory and discourse relations". In Proceedings of NAACL-HLT 2016 System Demonstrations