RST Visualizer: Uma ferramenta para a análise comparativa de anotações da Teoria da Estrutura Retórica

Carlos Vitor Cardoso da Silva [1], **Jackson da Cruz Souza** [2], **Paula Figueira Cardoso**[3]

[1]Instituto de Ciências Exatas e Naturais – Faculdade de Computação (FACOMP)
Universidade Federal do Pará (UFPA) – Belém – PA – Brasil

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

carlos.silva@icen.ufpa.br, pcardoso@ufpa.br, jacksoncruz@ufba.br

Abstract. *This paper presents the RST Visualizer, a tool developed to address the need for visualizing multiple layers of annotation in discourse analysis, such as RST tree structures and discourse markers associated with each rhetorical relation at the intra-sentential level. It is a novel tool designed to support research that requires qualitative analysis through the visual comparison of different types of annotation.*

Resumo. *Este artigo apresenta a ferramenta RST Visualizer, que surgiu da necessidade de visualizar diferentes camadas de anotações sobre a análise discursiva de textos, tais como sua representação no formato de árvore RST e sinalizadores discursivos de cada relação retórica no nível intra-sentencial. Trata-se de uma ferramenta inédita para apoiar pesquisas que precisam realizar análises qualitativas por meio da comparação visual de diferentes anotações.*

1. Introdução

A Rhetorical Structure Theory (RST) é uma teoria que busca compreender as relações entre diferentes partes de um discurso, estabelecendo como os segmentos textuais se conectam para formar uma mensagem coerente [Mann and Thompson 1988]. A premissa central é que um texto coerente é constituído por unidades mínimas de discurso, cada uma desempenhando um papel na construção do sentido global do texto. Essas unidades se relacionam por meio de relações retóricas (ou relações de coerência), organizadas em uma estrutura discursiva frequentemente representada como uma árvore hierárquica.

Na literatura, diversos estudos investigaram como as relações da RST contribuem para a compreensão do texto [Cardoso et al. 2024, Liu and Zeldes 2019, Das and Taboada 2018, Pardo 2005]. Pesquisas recentes destacam que a identificação de conectivos que sinalizam tais relações pode facilitar o processamento do texto. No entanto, a ausência de um marcador discursivo (MD) prototípico não necessariamente compromete sua interpretação. Nesse contexto, os autores [Das and Taboada 2018, Cardoso et al. 2024] foram além da análise de conectivos explícitos e anotaram diferentes tipos de pistas linguísticas — denominadas sinalizadores discursivos (SD) — que influenciam a escolha de uma relação retórica por parte de anotadores humanos.

Com os dados anotados, surge a necessidade de visualizar essas marcações de forma interativa, a fim de facilitar análises variadas. Diante da complexidade envolvida e

da necessidade de análises comparativas mais refinadas, este artigo apresenta a ferramenta RST Visualizer, desenvolvida para apoiar análises qualitativas por meio da comparação visual de diferentes anotações, possibilitando a identificação de pontos de concordância, bem como eventuais equívocos ou divergências entre anotadores.

Este trabalho está organizado da seguinte forma: na Seção 2, apresentam-se os trabalhos relacionados. Na Seção 3, descreve-se o desenvolvimento da ferramenta RST Visualizer. Na Seção 4, apontam-se considerações finais.

2. Trabalhos relacionados

Desde a sua criação, as pesquisas com RST avançaram em diferentes frentes, como desenvolvimento de corpora, analisadores discursivos automáticos e aplicações. No entanto, houve menos progresso na criação de interfaces online, colaborativas e atualizadas para anotação, o que facilitaria o desenvolvimento de novos conjuntos de dados anotados manualmente [Zeldes 2016]. Por muito tempo, a ferramenta de anotação mais utilizada foi a RSTTool[1] [O'Donnell 2000] que permite segmentar um texto em proposições, conectá-las com relações RST e visualizar a análise no formato de árvore. Recentemente, pesquisadores têm utilizado a ferramenta rstWeb[2] [Zeldes 2016] com as mesmas funcionalidades da RSTTool, além de outras funcionalidades. Sua principal é que o anotador pode acessá-la remotamente a partir de um servidor, além de permitir a anotação de SD em todos os níveis da árvore, o que favorece o desenvolvimento de estudos sobre como as relações são sinalizadas no discurso.

Além das ferramentas de anotação, trabalhos recentes investigam pistas linguísticas que apontam a escolha de uma relação discursiva. Para a língua inglesa, têm-se os trabalhos de [Das and Taboada 2018, Liu and Zeldes 2019]. Um dos achados é que uma relação pode ser sinalizada 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). Para o português brasileiro, [Rodrigues et al. 2023, Cardoso et al. 2024], anotaram SD no corpus CSTNews [Cardoso et al. 2011] e classificaram-nas em cinco grandes categorias: sintáticos, semânticos, gráficos, morfológicos e marcadores discursivos. Cada categoria ainda possui outras subcategorias.

Para gerar análises qualitativas, essas pesquisas apontam a necessidade de ferramentas que permitam ao pesquisador não apenas visualizar a estrutura do discurso, mas também examinar detalhadamente os elementos linguísticos associados às anotações. Dessa forma, observa-se uma lacuna no ecossistema de ferramentas disponíveis: a ausência de uma solução interativa que permita a comparação direta entre múltiplas anotações RST em um mesmo texto. Este trabalho busca preencher essa lacuna com o desenvolvimento da ferramenta RST Visualizer, voltada à análise qualitativa e comparativa de anotações RST, com suporte à exploração de diferentes tipos de SD.

[1]http://www.wagsoft.com/RSTTool/
[2]https://gucorpling.org/rstweb/info/

3. Desenvolvimento da ferramenta RST Visualizer

A ferramenta RST Visualizer[3] é uma aplicação web interativa desenvolvida para explorar e analisar textos anotados com relações retóricas (RST) e seus sinalizadores discursivos (SD). Seu principal diferencial é simplificar a busca por documentos e por relações intra-sentenciais e seus marcadores.

Desenvolvida com Angular[4] na interface (front-end) e NestJS[5] no servidor (back-end), a aplicação utiliza uma API GraphQL[6] para comunicação. O front-end foi escrito em TypeScript com a biblioteca visual DaisyUI[7].

Atualmente, as principais funcionalidades implementadas na ferramenta são: (1) Navegação e seleção de documentos *rs3*; (2) Visualização detalhada do conteúdo textual de um documento; (3) Apresentação das relações intra-sentenciais identificadas e seus sinalizadores; (4) Visualização do texto original formatado; (5) Filtragem de relações com base nos tipos e subtipos de sinalizadores associados; (6) Possibilidade de localizar e destacar no texto os segmentos correspondentes a uma relação.

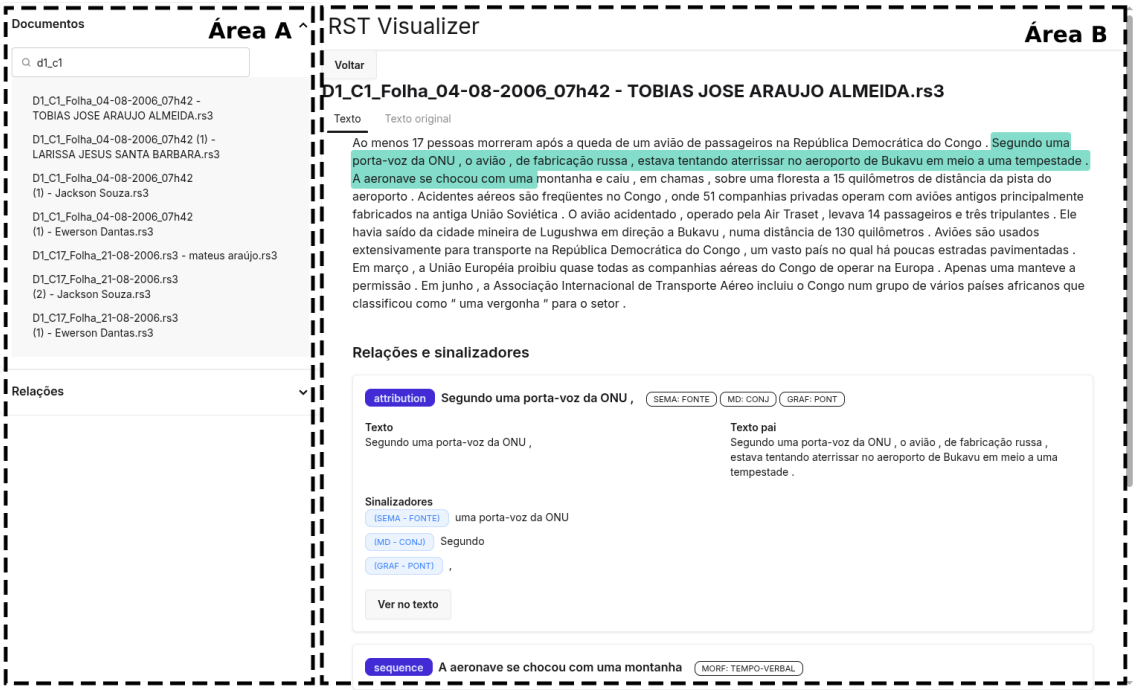

Figura 1. Página de visualização de um documento, destacando uma relação RST e seus sinalizadores.

A interface, ilustrada na Figura 1, organiza essas funcionalidades. Na Área A, o usuário pode navegar e filtrar a lista de documentos disponíveis, atendendo à Funcionalidade (1). Ao selecionar um documento, seu conteúdo é exibido na Área B, onde as Funcionalidades (2) e (3) são apresentadas: o texto anotado e uma lista detalhada de

[3]https://github.com/carloscardoso05/rst-visualizer
[4]https://angular.dev/
[5]https://nestjs.com/
[6]https://graphql.org/
[7]https://daisyui.com/

suas relações intra-sentenciais. Cada relação pode ser expandida para exibir detalhes dos sinalizadores, e um botão "Ver no texto"permite destacar o trecho correspondente no documento, conforme a Funcionalidade (6). Para contornar a perda de formatação que ocorre no processo de anotação, uma aba “Texto original” exibe o texto-fonte, atendendo à Funcionalidade (4).

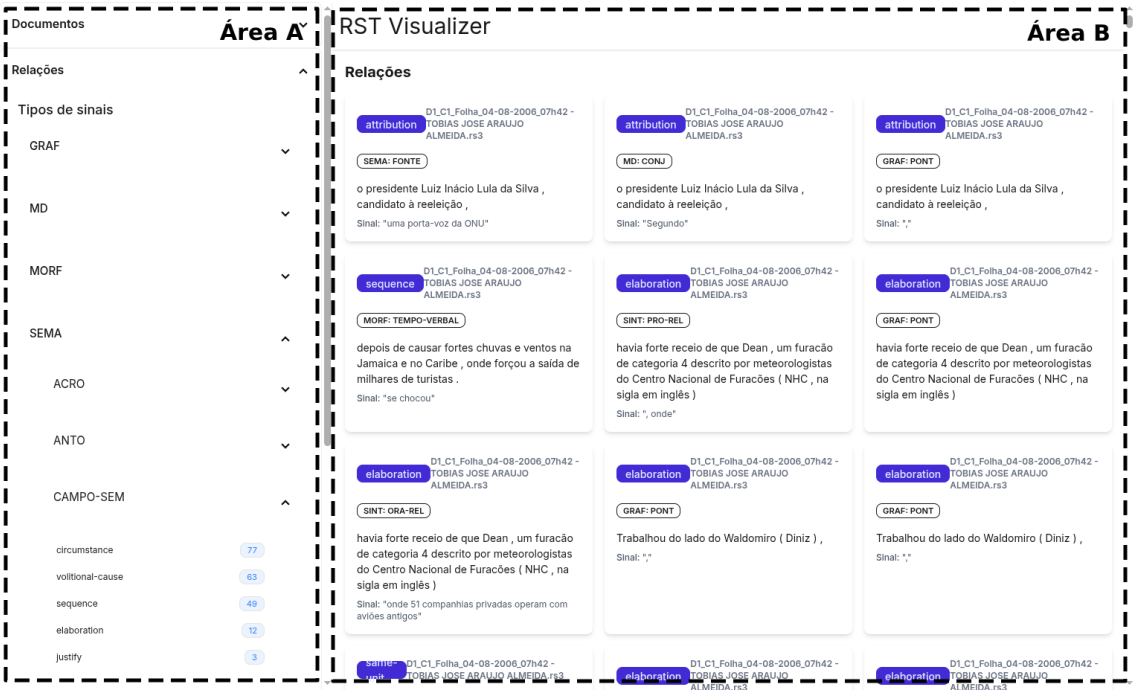

Figura 2. Página de visualização de relações e barra de pesquisa de relações na lateral.

A Figura 2 demonstra a Funcionalidade (5), onde uma barra lateral (Área A) permite ao pesquisador filtrar as relações exibidas (Área B) com base em tipos e subtipos de sinalizadores, refinando a análise.

4. Considerações finais

A identificação de relações intra-sentenciais nos textos anotados apresentou um desafio significativo, demandando a criação de um algoritmo em múltiplas etapas. O ponto de partida foi a análise da estrutura dos arquivos *.rs3* para viabilizar sua conversão em estruturas de dados manipuláveis. Com os dados devidamente organizados, aplicou-se um conjunto de regras para detectar as relações desejadas.

O desenvolvimento da interface gráfica constituiu o segundo desafio, pois era necessário atender a três requisitos principais: organizar o texto de forma visualmente agradável; garantir que cada token pudesse ser identificado individualmente, de modo a possibilitar o destaque das relações; e, por fim, a interface deveria permitir que o usuário realizasse buscas por tipos e subtipos de sinalizadores em todo o conjunto de textos.

No momento, a principal limitação da ferramenta é a impossibilidade de abrir mais de um documento simultaneamente. No entanto, isso já está sendo desenvolvido para ser incluído em versões futuras, bem como a visualização do documento como árvore e destacamento das relações na árvore.

Referências

Cardoso, P., Souza, J., Rodrigues, R., Dantas, E., Santa Bárbara, L., Araújo, M., Gama, N., Almeida, T., and Cruz, G. (2024). A linguagem em foco: Anotação de sinalizadores discursivos em textos jornalísticos. In *Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 247–256. SBC.

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., and 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. sn.

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

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

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

O'Donnell, M. (2000). RSTTOOL 2.4-a markup tool for Rhetorical Structure Theory. In *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*, pages 253–256.

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

Rodrigues, R., da Cruz Souza, J. W., and Cardoso, P. C. F. (2023). Sinalizadores retórico-discursivos: revisitando a anotação rst no córpus CSTNews. In *Simpósio Brasileiro De Tecnologia da Informação e da Linguagem Humana (STIL)*, pages 249–257. SBC.

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.