AVALIAÇÃO DE LLMS NA IDENTIFICAÇÃO DE SINALIZADORES DISCURSIVOS EM TEXTOS JORNALÍSTICOS ANOTADOS COM RST

Evaluation of LLMs in Identifying Discourse Signals in News Texts Annotated with RST

DOI: 10.14393/LL63-v41S-2026-11

Roana Rodrigues[*]

Paula Christina Figueira Cardoso[**]

Jackson Wilke da Cruz Souza[***]

[*] Doutora em Linguística pelo Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos. Professora do Departamento de Letras Estrangeiras e do Programa de Pós-Graduação em Letras da Universidade Federal de Sergipe. ORCID: 0000-0002-7748-8716. E-mail: roana(AT)academico.ufs.br

[**] Doutora em Computação pelo Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo/São Carlos. Professora na Faculdade de Computação da Universidade Federal do Pará. ORCID: 0000-0003-3621-8960. E-mail: pcardoso(AT)ufpa.br

[***] Doutor em Linguística pelo Programa de Pós-Graduação em Linguística da Universidade Federal de São Carlos. Professor no Instituto de Ciência, Tecnologia e Inovação e do Programa de Pós-Graduação em Língua e Cultura da Universidade Federal da Bahia. ORCID: 0000-0003-1881-6780. E-mail: jackcruzsouza(AT)gmail.com

RESUMO: Neste estudo, investigamos a aplicabilidade de *Large Language Models* (LLMs) na anotação de sinalizadores discursivos (SDs) que contribuem com a identificação de relações de coerência em textos jornalísticos do português brasileiro, anotados segundo a *Rhetorical Structure Theory* (RST). Utilizamos os modelos GPT-4 e LLaMA-4 para analisar três relações retóricas presentes no *corpus* CSTNews, previamente anotado por especialistas humanos. A análise comparativa revelou uma convergência expressiva quanto à identificação de SDs considerados prototípicos de cada relação. O GPT-4 apresentou anotações mais granulares, incluindo SDs que não contribuíam diretamente para a identificação da relação; já o LLaMA-4 produziu anotações mais sucintas, com predominância de SDs únicos. Os modelos sugeriram possíveis novos sinalizadores não catalogados. Concluímos que, embora não substituam a anotação humana, os LLMs podem atuar como ferramentas auxiliares, otimizando o tempo e possibilitando a ampliação da análise linguística em contextos computacionais.

PALAVRAS-CHAVE: *Rhetorical Structure Theory*. Sinalizadores discursivos. *Large Language Models*. Anotação de *corpus*. Processamento de Linguagem Natural.

ABSTRACT: In this study, we investigated the applicability of Large Language Models (LLMs) in the annotation of signal markers (SMs) that contribute to the identification of coherence relations in Brazilian Portuguese news texts, annotated according to the Rhetorical Structure Theory (RST). We used the GPT-4 and LLaMA-4 models to analyze three rhetorical relations in the CSTNews *corpus*, previously annotated by human experts. The comparative analysis revealed a significant convergence in the identification of SMs considered prototypical of each relation. GPT-4 produced more fine-grained annotations, including SMs that did not directly contribute to identifying the relation, whereas LLaMA-4 generated more concise annotations, predominantly assigning single SMs. The models also suggested potential new, non-catalogued signals. We concluded that, although they do not replace human annotation, LLMs can serve as auxiliary tools, optimizing time and enabling the expansion of linguistic analysis in computational contexts.

KEYWORDS: Rhetorical Structure Theory. Discourse signals. Large Language Models. Corpus annotation. Natural Language Processing.

1 Introdução

Nas pesquisas linguísticas, o *corpus* é um elemento fundamental, já que se caracteriza por ser uma amostra representativa do uso real de uma língua. Como sabido, para a Linguística de *Corpus*, tal *amostra* possui um tamanho finito, que, a partir de textos autênticos, apresenta o máximo de representatividade de uma determinada língua ou variedade linguística. Além disso, o *corpus* possui formato eletrônico e se caracteriza pela sua *reutilização*, ou seja, apesar de ser construído para uma determinada finalidade de pesquisa, pode ser disponibilizado para consultas e aplicações em outras investigações (Aluísio; Almeida, 2006).

A *anotação* é um processo posterior à compilação do *corpus*, no qual há o seu enriquecimento com a adição de informações linguísticas variadas, de acordo com os objetivos e a fundamentação teórico-metodológica de cada estudo. Há *corpora* compilados com anotações em diferentes níveis linguísticos contemplando diversos fenômenos. O processo de anotação manual é demorado e exige tempo e mão de obra qualificada. Sua funcionalidade contribui não só com os estudos descritivos das línguas, mas também com outras áreas do conhecimento, sendo, por exemplo, um recurso basilar para as pesquisas em Processamento de Língua Natural (PLN), já que é possível realizar análises estatísticas que permitem a prospecção e/ou regressão de comportamentos linguísticos.

No contexto do PLN, anotações de alta qualidade desempenham um papel fundamental no treinamento e na avaliação de modelos de aprendizado de máquina. Como se trata de uma tarefa que ao mesmo tempo é fundamental e laboriosa, atualmente, observa-se um número crescente de estudos que, fundamentados no desempenho notável dos *Large Language Models* (doravante LLMs) em diversas tarefas de PLN, têm investigado seu uso como anotadores automáticos de dados (Muermans; Kosseim, 2022; Barbosa; Campelo, 2024; Wang *et al.*, 2024).

Os LLMs, tais como o GPT e o LLaMA, são sistemas de Inteligência Artificial (IA), treinados em *corpora* extensos, que têm revolucionado a geração de informação, com a produção de "textos coerentes e contextualmente relevantes em uma ampla gama de tópicos" (Barbosa; Campelo, 2024, p. 1). Por outro lado, conforme salientam Nasution *et al.* (2024), embora as anotações geradas por LLMs sejam mais econômicas e eficientes de serem obtidas, elas frequentemente apresentam erros em tarefas complexas ou específicas de domínio e

podem introduzir vieses quando comparadas às anotações realizadas por especialistas humanos.

Como visto, a anotação linguística é uma tarefa árdua a ser desenvolvida tanto pelos seres humanos, quanto pelas máquinas. Somado a isso, sobre o português brasileiro (PB), há ainda poucos *corpora* anotados em nível discursivo disponibilizados, cuja complexidade é ainda maior. Para o PB, temos como referência o CSTNews (Cardoso *et al.*, 2011), um *corpus* jornalístico anotado discursivamente, com base na *Rhetorical Structure Theory* (RST), uma teoria de base funcionalista, fundamentada por Mann e Thompson (1987), que propõe a descrição das relações de coerência textual das línguas naturais.

Em recente trabalho (Cardoso *et al.*, 2024), foi retomada a anotação das relações de coerência (também denominadas *relações retóricas*) do CSTNews, a fim de se investigarem e anotarem as pistas linguísticas, nomeadamente *sinalizadores discursivos* (SDs) que contribuem com a identificação dessas relações. Trata-se de uma atividade primordial dos estudos da então chamada eRST (*Enhanced Rhetorical Structure Theory)*, proposta por Zeldes *et al.* (2024), com foco no aprimoramento da teoria.

Portanto, este trabalho surge da concatenação de dois eventos principais: a agenda dos projetos de pesquisa em desenvolvimento,[1] que visam à produção de recursos linguísticos do PB para o PLN, com a produção, divulgação e disponibilização de *corpus* anotado em nível discursivo; e a preocupação atual do estado da arte em avaliar a performance de LLMs em diferentes tarefas, sendo, mais especificamente, a investigação da tarefa de anotação automática de *corpus* com base na RST.

Sendo assim, com o desenvolvimento desta pesquisa espera-se discutir os seguintes pontos: (i) quais as convergências e divergências entre a anotação humana e as anotações geradas pelos LLMs?; e (ii) quais as particularidades da geração de informação apresentada pelos LLMs investigados: GPT-4 (Achiam *et al.*, 2024) e LLaMA-4 (Huang *et al.*, 2023)?.

Para tanto, este trabalho se organiza da seguinte maneira: na Seção 2, são apresentados os fundamentos da RST, assim como reflexões a respeito do uso de LLMs para a descrição linguística; na Seção 3, são detalhados os procedimentos metodológicos adotados na pesquisa;

[1] Os autores deste trabalho integram os projetos POeTiSA (https://sites.google.com/icmc.usp.br/poetisa) e RST *beyond discourse markers* (https://sites.google.com/view/rst-poetisa), com foco na produção de recursos linguísticos em nível discursivo para o PB. Sites acessados em maio de 2025.

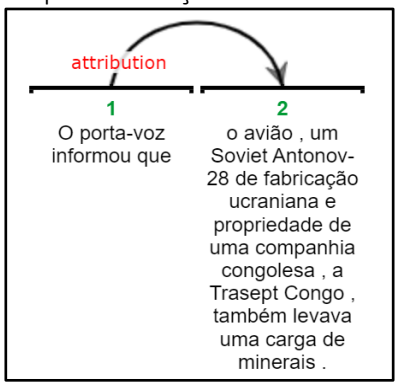
na Seção 4, discutimos os resultados obtidos. Por fim, expomos as contribuições deste estudo e os trabalhos futuros a serem desenvolvidos.

2 A anotação em nível discursivo: a respeito da anotação humana e do uso de LLMs

Para a análise do desempenho de LLMs na tarefa de anotar pistas linguísticas (SDs), em nível discursivo, utilizamos como base teórico-metodológica a *Rhetorical Structure Theory* - RST (Mann; Thompson, 1987). Com base na RST, tem-se a anotação, análise e descrição entre duas proposições de texto: *Elementary Discourse Units* (EDUs), que se constituem, geralmente, por um *núcleo* (N), que contém a informação que explicita a intenção principal do falante/escritor, e um *satélite* (S), que possui um dado adicional que influencia na interpretação da informação nuclear por parte do ouvinte/leitor. Além disso, é possível haver relações constituídas por N-N, tais como as relações *Contrast, List* e *Sequence.* De maneira geral, para a identificação das relações são consideradas as restrições sobre N, as restrições sobre S, as restrições da relação N+S e o efeito desencadeado.

Para o PB, com base na RST é possível identificar 32 relações de coerência,[2] conforme proposto em Pardo (2005). Na Figura 1, tem-se um exemplo da relação *Attribution*, retirado do *corpus* CSTNews.

Figura 1 – Exemplo da relação de coerência *Attribution.*



Fonte: *Corpus* CSTNews anotado.

[2] Segundo Pardo (2005), as relações retóricas/de coerência do PB são: *Antithesis, Attribution, Background, Circumstance, Comparison, Concession, Conclusion, Condition, Contrast, Elaboration, Enablement, Evaluation, Evidence, Explanation, Interpretation, Joint, Justify, List, Means, Motivation, Non-Volitional Cause, Non-Volitional Result, Otherwise, Parenthetical, Purpose, Restatement, Same-Unit, Sequence, Solutionhood, Summary, Volitional Cause* e *Volitional Result*.

A relação *Attribution* apresenta um N (EDU 1) e um S (EDU 2), cujo efeito desencadeado é que o leitor reconheça que a informação contida na segunda EDU está sendo atribuída à fonte presente na primeira. Como a organização das relações se dá de maneira hierárquica, a teoria prevê uma representação em formato de árvore, que, conforme o tamanho do texto, pode ter mais ramificações e/ou níveis.

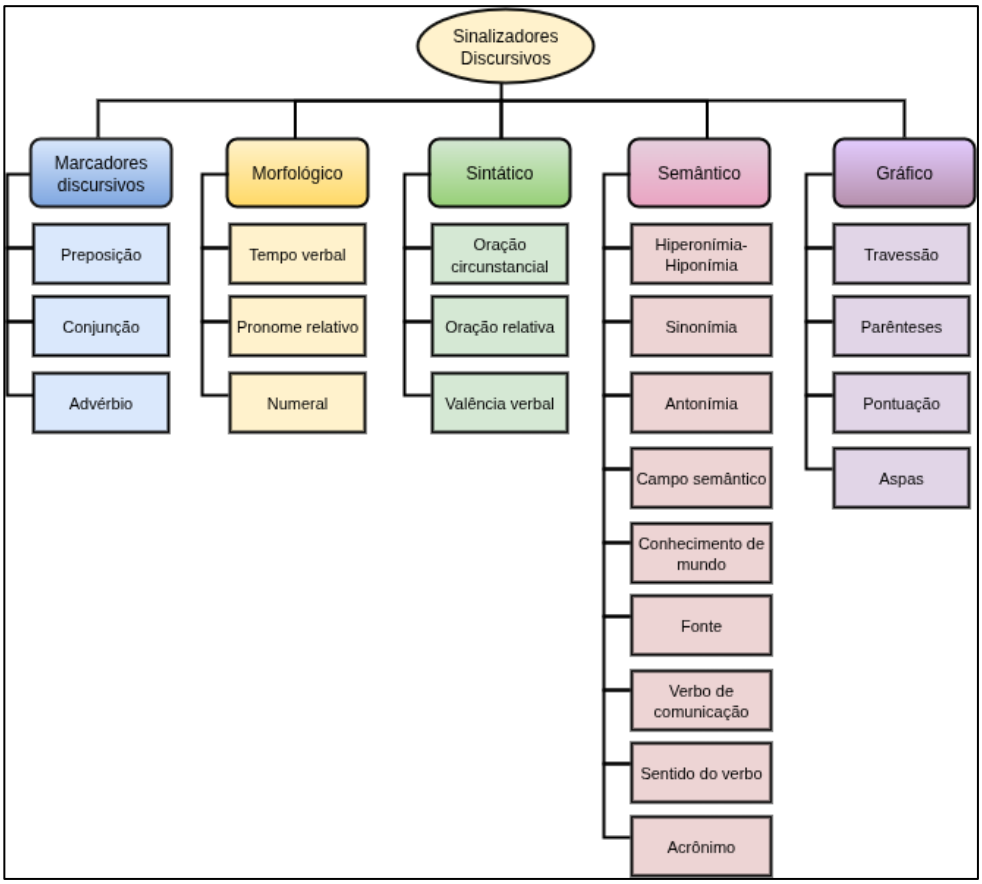
O *corpus* CSTNews (Cardoso *et al.*, 2011), utilizado nesta investigação, é composto por 50 *clusters* (conjuntos de textos) constituídos, cada um, por 2 ou 3 notícias retiradas das seções de Esporte, Mundo, Dinheiro, Política, Ciências e Cotidiano de diferentes jornais brasileiros, totalizando 140 textos. Além disso, o *corpus* inclui diversas camadas de anotação, sendo a anotação RST o foco deste trabalho. Conforme se verifica na Figura 1, a primeira anotação realizada consistiu na segmentação e no apontamento da relação retórica estabelecida entre as EDUs. Já em trabalho mais recente, anotou-se, com o uso da ferramenta rstWeb (Zeldes, 2016), os SDs que contribuem com a interpretação da relação indicada. No Manual de Anotação (Dantas *et al.*, 2024),[3] foi proposta uma taxonomia desses SDs, conforme replicamos na Figura 2.

Sendo assim, entende-se que os SDs podem ser dos seguintes tipos, para além dos marcadores discursivos prototípicos: *gráficos*, *morfológicos*, *sintáticos* e *semânticos*, podendo aparecer sozinhos ou combinados, a depender das EDUs e das relações. Tal taxonomia coincide em muitos aspectos com os trabalhos que vêm sendo desenvolvidos para a língua inglesa (Das; Taboada, 2018; Zeldes *et al.*, 2024) e, embora utilizada nos projetos de pesquisa atuais, está em constante revisão.

A elaboração da proposta de taxonomia e a anotação do *corpus* foi uma atividade que levou aproximadamente quatro anos de trabalho (de 2021 a 2024), considerando o período de implementação do projeto, revisão sistemática da literatura e primeiras rodadas de anotação. Nesse contexto, é notável como o uso de LLMs na anotação de dados, além de prover agilidade à tarefa, pode apontar inconsistências no processo e indicar possíveis caminhos para solucioná-las a partir de *prompts* adequados.

[3] Informações mais detalhadas a respeito da taxonomia estão disponíveis em Souza, Cardoso e Rodrigues (2025).

Figura 2 – Taxonomia dos Sinalizadores Discursivos para o PB.



Fonte: Dantas *et al.* (2024).

Várias pesquisas em PLN têm se dedicado à avaliação do desempenho de modelos de LLMs em diferentes aplicações. Especificamente sobre o desempenho de LLMs no processamento em nível discursivo, retomamos os trabalhos de Muermans e Kosseim (2022) e Chan *et al.* (2024) sobre o inglês, e de Barbosa e Campelo (2024) sobre o PB.

Segundo Chan *et al.* (2024), para compreender o texto em linguagem natural em um nível mais profundo, é crucial para um LLM capturar e compreender as relações intersentenciais de nível superior do texto, o que envolve dominar relações mais complexas e abstratas além das formas superficiais. Dado o desempenho do ChatGPT em outras tarefas, os pesquisadores avaliaram quantitativamente o desempenho do modelo em compreender relações intersentenciais, mais precisamente as relações temporais, causais e discursivas entre duas frases. Os autores conduziram diversas avaliações em onze conjuntos de testes sobre as relações, a partir de três cenários de *prompt*: *zero-shot*, quando o modelo realiza uma tarefa

sem ver exemplos; *zero-shot engineering*, quando o modelo realiza a tarefa também sem exemplos, mas com maior estruturação das instruções dadas; e *in-context learning*, no qual o modelo aprende com exemplos incluídos no próprio *prompt*.

Os autores também observaram e avaliaram o desempenho do modelo nas várias intrarrelações de cada relação intersentencial (por exemplo, a relação *antes* e *depois* em Relações Temporais). Como resultado, os autores identificaram o excelente desempenho do ChatGPT na detecção e no raciocínio sobre relações causais, além do fácil reconhecimento das relações de discurso devido ao uso de conectivos discursivos explícitos no contexto. No entanto, verificaram-se dificuldades em casos sem conectivos para tarefas discursivas implícitas, e na identificação da ordem temporal entre dois eventos. Tal fato pode ser atribuído ao *feedback* humano inadequado durante o processo de treinamento do modelo.

Vê-se que os conectivos são elementos-chave para a compreensão e geração de dados em nível discursivo. Para Muermans e Kosseim (2022), identificar os conectivos discursivos (*but/mas*, *if/se* etc.) é fundamental para tarefas de PLN que dependem da análise do discurso para entender como os elementos textuais são relacionados entre si, tais como geração de texto, sistemas de diálogo e sumarização. Os autores relatam experimentos para identificar conectivos discursivos multilíngues (inglês, turco e mandarim) e mostram como modelos de BERT (modelo de linguagem também baseado em IA, que entende o significado das palavras em seu uso, considerando o contexto frasal) específicos para cada idioma parecem ser suficientes mesmo com poucos dados de treinamento próprios para cada tarefa.

Sobre o PB, Barbosa e Campelo (2024) avaliaram a capacidade de diversos LLMs, tais como o GPT-4o e o Claude Opus, em analisar a coerência textual. A pesquisa examinou especificamente a coerência local e global, comparando o desempenho dos modelos em interações via API e por meio de interfaces de chat. Para avaliação da *coerência local*, que é o foco deste trabalho, Barbosa e Campelo (2024) utilizaram o teste de embaralhamento, que compara a coerência de um texto em sua ordem original com uma versão embaralhada. Esse teste foi aplicado a uma variedade de textos de quatro *corpora*: COCA (Davies, 2008), GCDC (Lai; Tetreault, 2018), DDisCo (Mikkelsen *et al.*, 2022) e CSTNews (Cardoso *et al.*, 2011), permitindo avaliar a capacidade dos modelos na identificação da sequência coerente de sentenças. A metodologia garantiu uma análise robusta da coerência local em diferentes

contextos textuais. Em termos de indicador de desempenho (medida-F), todos os modelos tiveram bons resultados, com valores superiores a 70% para ambas as abordagens. Contudo, os modelos GPT-4o e o Claude Opus superaram consistentemente os demais LLMs avaliados.

Como apresentado pelos autores, a literatura sobre o uso de LLMs relata que tais modelos conseguem gerar texto coerente em uma ampla variedade de assuntos, o que indica que eles têm capacidade refinada para discernir a coerência ou a falta dela. Wang *et al.* (2024) destacam que, em vez de substituir completamente os anotadores humanos por LLMs, é mais eficaz combinar as competências desses modelos com o conhecimento dos especialistas humanos, a fim de assegurar anotações precisas e confiáveis.

Os autores descrevem uma abordagem colaborativa multietapas entre humanos e LLMs, na qual: (i) LLMs geram etiquetas e fornecem explicações; (ii) um verificador/analista avalia a qualidade das etiquetas geradas pelos LLMs; e (iii) anotadores humanos reanotam um subconjunto de etiquetas com pontuações de verificação mais baixas. Para facilitar a colaboração entre humanos e LLMs, Wang *et al.* (2024) utilizaram a capacidade que os LLMs têm de racionalizar suas decisões. Portanto, as explicações geradas por LLMs podem servir como informações adicionais ao modelo do verificador/analista, bem como ajudar os humanos a entenderem melhor as etiquetas de LLMs.

Apesar de concordarmos com os autores sobre a necessidade de uma relação harmoniosa e colaborativa entre o conhecimento produzido por humanos e a utilização de LLMs, nos distanciamos, neste trabalho, das etapas prescritas por Wang *et al*. (2024), já que o processo de criação de etiquetas e a primeira rodada de anotação foram realizados por humanos. Assim, o objetivo aqui é avaliar o desempenho dos LLMs na execução dessa mesma tarefa.

3 Processos metodológicos

Como demonstrado, as EDUs utilizadas neste trabalho são provenientes do *corpus* CSTNews (Cardoso *et al.*, 2011). A escolha desse *corpus* baseia-se no fato de que, em PB, há poucos recursos linguísticos previamente anotados segundo o modelo RST. Em Dantas *et al.* (2024) foi criado um subconjunto desse mesmo *corpus* (totalizando 50 textos) e anotado por 8

especialistas[4] com a taxonomia dos SDs, em relações *intrassentenciais* (que ocorrem dentro de uma mesma sentença).

Para o presente trabalho, foi selecionada uma amostra mapeando-se as EDUs em que as relações *Attribution*, *Circumstance* e *Elaboration* ocorreram. A escolha por estas relações está no fato de elas caracterizarem a construção discursiva de textos jornalísticos, a saber: reportar eventos; complementar informações; e circunstanciar os fatos narrados. Ao final, foram analisados 20 casos de *Attribution*, 20 de *Elaboration* e 5 de *Circumstance*.

Os dados foram organizados em planilhas, separando-se cada uma das EDUs, com a indicação da relação entre elas e a decisão da anotação humana. Em um segundo momento, foram escolhidos os LLMs GPT-4 (Achiam *et al.*, 2024) e LLaMA-4 (Huang *et al.*, 2023) para gerarem a anotação dos dados da amostra. A escolha por esses modelos de língua ocorreu devido a sua recorrência em experimentos similares na área de PLN e por possuírem especificidades interessantes ao se considerar que o LLaMA-4 foi treinado com dados do português e o GPT-4, com dados multilíngues. Os resultados gerados pelos LLMs também foram acrescentados na planilha, para facilitar o trabalho de análise e comparação das anotações.

Nos experimentos foram utilizadas as interfaces *web* públicas dos modelos selecionados sem ajuste de hiperparâmetros. A técnica utilizada para a elaboração do *prompt* foi a *few-shot learning* (Sivarajkumar *et al.*, 2024), ou seja, a elaboração do *prompt* com a apresentação de poucos exemplos, já que a caracterização de SDs em função das relações RST pode ser ambígua e/ou causar discordâncias, como ocorre naturalmente entre anotadores humanos.

No Quadro 1, tem-se duas versões dos *prompts* elaborados. Ao utilizar a primeira versão, sem explicar o que seriam as categorias de SDs, os resultados foram muito superficiais, sendo que os LLMs não indicavam com maior precisão os sinalizadores. Além disso, na primeira versão não se contemplava a indicação de novos sinalizadores para além dos que já estavam previstos na taxonomia. Assim, após essa observação, o *prompt* foi aprimorado, chegando-se à sua segunda versão. É importante ressaltar que, visando minimizar eventuais alucinações

[4] Os especialistas que realizaram a anotação dos SDs são professores e estudantes de diferentes áreas, tendo como base os interesses relacionados ao Processamento de Língua Natural, a saber: Letras, Linguística, Ciência da Computação e Formação Interdisciplinar em Ciência, Tecnologia e Inovação. Todas as pessoas envolvidas fazem parte do projeto *RST beyond discourse markers* e realizaram o processo de anotação sob supervisão de anotadores mais experientes, que já trabalharam com anotação, inclusive em RST, em pesquisas anteriores.

causadas pelos LLMs, decidiu-se fazer a anotação automática de cada sentença da amostra individualmente.

Quadro 1 – *Prompts* utilizados para anotação automática

1ª versão do *prompt*	2ª versão do *prompt*
Considerando a Rhetorical Structure Theory e a proposta de sinalizadores discursivos, identifique os sinalizadores presentes entre os segmentos identificados como EDU_1 e EDU_2 em que ocorre a relação Circumstance. [EDU_1] Depois de 20 dias de tempo seco, [EDU_2] voltou a chover na capital paulista. *Considerando a Rhetorical Structure Theory e a proposta de sinalizadores discursivos, identifique os sinalizadores presentes entre os segmentos identificados como EDU_1 e EDU_2 em que ocorre a relação Circumstance.* *[EDU_1] Depois de 20 dias de tempo seco,* *[EDU_2] voltou a chover na capital paulista.*	*Considere a seguinte proposta de tipologia de sinalizadores discursivos de Dantas et al. (2024) para o português com base na teoria RST foi a seguinte:* ● *Sinalizadores Gráficos: Travessão, Parênteses, Pontuação e Aspas;* ● *Sinalizadores Semânticos: Hiperonímia - Hiponímia, Sinonímia, Antonímia, Campo semântico, Conhecimento de mundo, Fonte, Verbo de comunicação, Sentido do verbo e Acrônimo* ● *Sinalizadores sintáticos: Oração circunstancial, Oração relativa, Valência verbal* ● *Sinalizadores Morfológicos: Tempo verbal, Pronome relativo e Numeral* ● *Marcadores discursivos: Preposição, Advérbio e Conjunção.* *Com base nisso, indique no texto abaixo quais os possíveis tipos e subtipos de sinalizadores das relações RST já indicadas entre as EDUs. Em seguida, proponha a classificação dos sinalizadores conforme a proposta taxonômica de Dantas et al. (2024). Caso haja outros sinalizadores que não estejam contemplados na proposta, identifique-os.*

Fonte: elaborado pelos autores.

4 Resultados

Com base na exploração dos dados gerados pelas LLMs, foi possível identificar pontos de encontro e distanciamento entre as bases estudadas e entre elas e a anotação humana, o que possibilitou várias reflexões, dentre as quais se destacam a revisão da proposta de taxonomia utilizada (Dantas *et al.*, 2024) e o *prompt* criado para a tarefa. A fim de organizar a discussão, apresentaremos os dados de acordo com cada uma das relações estudadas: *Attribution*, *Elaboration* e *Circumstance*.

Attribution é uma relação constituída por um núcleo que apresenta uma expressão, fala ou pensamento (*mensagem*) e um satélite que expõe quem produz essa informação (*fonte*), tendo, portanto, como efeito o fato de o leitor ser informado sobre a mensagem e quem a produziu. Mesmo não havendo total concordância entre os anotadores humanos, nem a relação se estabelecer a partir de uma única construção, é possível afirmar que a relação *Attribution* é majoritariamente identificada pelos humanos por meio dos seguintes SDs: de tipo semântico (*verbo de comunicação* e *fonte*) e de tipo marcador discursivo (*conjunção*). Em comum com a anotação humana, em todas as 20 relações *Attribution* estudadas, os modelos (GPT-4 e LLaMA-4) anotaram o *verbo de comunicação* como um SD que identifica essa relação. Destaca-se que os LLMs apresentaram os resultados em formato de relatório indicando o SD, além de realizar comentários sobre eles. No Quadro 2, tem-se um exemplo de anotação gerada considerando apenas os SD indicados.

Quadro 2 – Exemplo de anotação automática da relação *Attribution*

Segmento textual	Modelo	
	GPT-4	LLaMA-4
[EDU1] Em seu programa semanal de rádio, Café com o Presidente, Lula disse que [EDU 2] este será o maior investimento em saneamento "da história do País".	verbo de comunicação + aspas + fonte + conjunção	verbo de comunicação

Fonte: elaborado pelos autores, a partir da anotação do D2_C6[5] do CSTNews.

Conforme demonstrado no Quadro 2, percebe-se que o GPT-4 considera SDs de diferentes tipos, ao passo que o LLaMA-4 é bastante conciso na indicação; comportamento esse que se repete na anotação de toda a amostra deste trabalho. Embora o LLaMA não evidencie diferentes sinalizadores, nos comentários do relatório há a indicação de que o *verbo de comunicação* pressupõe a *fonte da informação*, sendo desnecessário, portanto, separar dois sinalizadores que, ao cabo, possuem a mesma função. O rótulo *fonte* foi aplicado pelo LLaMA apenas em casos em que a EDU não apresentava um *verbo de comunicação*. A proposta de revisão da taxonomia dos SDs está em desenvolvimento em um trabalho paralelo. Em reuniões

[5] Os códigos apresentados ao final de cada exemplo indicam o documento (D) e o cluster (C) a que o exemplo pertence no corpus CSTNews.

anteriores, a equipe já havia discutido a possibilidade de simplificar a anotação das relações de *Attribution* por meio da utilização de um único rótulo. Nesse sentido, a sugestão apresentada pelo LLaMA parece apontar e corroborar os caminhos futuros da investigação.

O GPT-4 apresenta uma descrição minuciosa dos sinalizadores. Em alguns casos, como em (1), foram apontados cinco SDs que indicariam a relação *Attribution*: *fonte* (*O ministro da saúde egípcio, Hatem El-Gabaly*), *vírgula* (*após a informação da fonte*)*, verbo de comunicação* (*informou*)*, advérbio* (a construção temporal: *nesta segunda-feira*) e *conjunção* (*que*).

(1) [EDU 1] O ministro da Saúde egípcio, Hatem El-Gabaly, informou nesta segunda-feira que
[EDU 2] 57 pessoas morreram
[EDU 3] e 128 ficaram feridas no choque entre dois trens de passageiros no delta do Nilo, ao norte do Cairo. [D2_C14]

Em análise, questionou-se se o GPT realmente entendeu a tarefa proposta neste trabalho ou se apenas aplicou de maneira indistinta a taxonomia . Apesar do aparente excesso de rótulos, os resultados possibilitaram perceber como o apontamento do SD de tipo gráfico *aspas* é relevante para a identificação da relação *Attribution*. Na amostra analisada, o GPT-4 indicou esse SD como relevante não apenas em casos de discurso direto, predominantemente apontado na anotação humana, como em (2), mas também em discursos indiretos com breves citações diretas em seu interior, como *"da história do País"* no Quadro 2 e *"um verdadeiro canteiro de obra"* no exemplo em (3), casos esses não identificados na anotação humana:

(2) [EDU 1] "Confirmamos que água contendo uma pequena quantidade de material radioativo vazou da instalação",
[EDU 2] explicou Shougo Fukuda, um porta-voz da Tokyo Electric. [D4_C32]

(3) [EDU 1] O presidente Luiz Inácio Lula da Silva disse, nesta segunda-feira, 6 que
[EDU 2] irá transformar o Brasil em "um verdadeiro canteiro de obra" [D2_C6]

Sendo assim, a anotação feita pelo GPT permitiu explorar as diferentes configurações da relação *Attribution* e propor alternativas para as suas especificidades. Por outro lado, no entanto, o aparente excesso de rótulos apontados produziu também equívocos, como em (4),

no qual as aspas não têm relação com a *mensagem* (“*que ficou triste com as vaias que recebeu…”*) proferida pela *fonte* (“*O presidente Luiz Inácio Lula da Silva”*):

(4) [EDU 1] O presidente Luiz Inácio Lula da Silva afirmou nesta segunda-feira, durante o programa de rádio "Café com o Presidente",
[EDU 2] que ficou triste com as vaias que recebeu durante a abertura oficial da 15ª edição dos Jogos Pan-Americanos, [D1_C49]

Em (4), “Café com o Presidente” faz remissão ao nome do programa de rádio no qual a *fonte* emitiu a *mensagem*. Portanto, embora seja um SD relevante para a identificação da *Attribution*, o GPT-4 não parece ter compreendido e distinguido os seus usos apenas quando relevante para a identificação dessa relação; daí a importância da revisão e análise humana, seja sobre a anotação em si, seja, inclusive, no estudo do *prompt* que levou a IA a apresentar tais resultados.

Elaboration é uma relação constituída por um núcleo que apresenta um conteúdo principal e um satélite que expressa o detalhamento ou informação adicional em relação ao núcleo. Tal como na relação *Attribution*, nesta os anotadores humanos também não são unânimes na indicação dos sinalizadores. Porém, de maneira geral, identificaram que a *Elaboration* pode ser sinalizada por tipo sintático (*oração relativa*), tipo morfológico (*pronome relativo*) e tipo gráfico (*vírgula*). No Quadro 3, apresenta-se um exemplo de anotação da relação *Elaboration* feita pelos LLMs.

Quadro 3 – Exemplo de anotação automática da relação *Elaboration.*

Segmento textual	Modelo	
	GPT-4	LLaMA-4
[EDU 1] A região de Niigata sofreu no dia 23 de outubro de 2004 um terremoto de magnitude 6,8 [EDU 2] que deixou 67 mortos e mais de 3.000 feridos.	pronome relativo + oração relativa + numeral + campo semântico	oração relativa

Fonte: elaboração própria, a partir da anotação do D4_C32 do CSTNews.

As EDUs apresentadas no Quadro 3 tiveram como anotação humana *pronome relativo* e a *oração relativa*. Um dos anotadores humanos apontou ainda o sinalizador de tipo semântico

conhecimento de mundo, o que indica haver alguma relação entre a alta magnitude do terremoto e o número de vítimas. Os LLMs também indicaram a oração relativa como SD, o que sugere ser esse o rótulo mais relevante e produtivo dessa relação.

Tem-se, portanto, uma anotação mais concisa proposta pelo LLaMA, com a apresentação de apenas um SD (*oração relativa*) para identificar a *Elaboration*. Isso designa um reiterado comportamento de economia desse modelo, já que efetivamente o sinalizador de tipo morfológico (*pronome relativo*) e o sinalizador de tipo sintático (*oração relativa*) se sobrepõem, não havendo, de fato, ao menos em uma análise preliminar, a necessidade de existência e/ou indicação dos dois SDs na anotação do *corpus*.

Por outro lado, a anotação proposta pelo GPT é mais detalhada, dando atenção ao numeral (*6,8*, *67* e *3.000*) e ao campo semântico (*terremoto* e *magnitude; mortos* e *feridos*), como rótulos relevantes para a identificação da relação. A indicação do *campo semântico* parece dialogar com o apontamento de *conhecimento de mundo* explorado por um dos anotadores humanos. Ainda, mantém-se o questionamento feito sobre a possibilidade de tais SDs realmente serem fundamentais para a identificação da relação RST em questão ou se o LLM apenas aplica o máximo de rótulos possível nas EDUs em análise.

Verificou-se ainda que os LLMs apresentaram rótulos em todas as três relações estudadas não previstos na taxonomia de Dantas *et al.* (2024), conforme o exemplo (5):

(5) (...)
[EDU 3] Lula disse que, na próxima semana,
[EDU 4] quando voltar da viagem
[EDU 5] que está fazendo pela América Central,
[EDU 6] irá começar a anunciar as obras de infra-estrutura em transporte
[EDU 7] como estradas, ferrovias, gasodutos a portos e aeroportos. [D2_C6]

Em (5), as relações de *Elaboration* se estabelecem entre as EDUs 5 e 4 e entre as EDUs 7 e 4-6. Na EDU 7, tanto a anotação humana quanto o GPT apontaram "*como*" (*conjunção*) como um sinalizador relevante para a interpretação da relação *Elaboration* no segmento textual. Por sua parte, o LLaMA indicou o SD de tipo semântico (*hiperonímia*), estabelecendo relação entre *obras de infraestrutura* e os tipos de tais obras (*estradas, ferrovias, gasodutos, portos* e *aeroportos*). Além disso, ele apontou também, para o mesmo fragmento, a subclasse *exemplificação*, não existente na tipologia trabalhada.

Em toda a amostra a apresentação de um SD não proposto na tipologia aconteceu duas vezes no LLaMA, na indicação de *exemplificação* e de uma *oração subordinada*. Na tipologia utilizada como base, não há o termo *subordinada* em nenhum momento, embora esteja prevista a anotação, de tipo sintático, de orações relativas e circunstanciais. Com o GPT, o comportamento foi similar, com o apontamento do rótulo *conjunção subordinativa*, gerando uma anotação com maior grau de granularidade (já que a taxonomia prevê apenas a subclasse *conjunção*); e das subclasses *oração subordinada* e *elemento referencial*, sendo o último apresentado no exemplo (6):

(6) (...)
[EDU 4] a colisão ocorreu
[EDU 5] quando o trem número 808
[EDU 6] que circulava em alta velocidade
[EDU 7] se chocou com a parte traseira de outro,
[EDU 8] que vinha da de Qaliub. [D2_C14]

Em (6), tem-se duas relações: *Elaboration*, entre as EDUs 6 e 5 e entre as EDUs 8 e 7. Nessas relações, marcadas sobretudo pelas orações relativas, o GPT indicou o elemento referencial (*de outro*) como um SD relevante para a interpretação do efeito produzido no texto. Entende-se que *de outro* remete àa *trem*, elemento que aparece na EDU 5. A anotação feita pelo GPT chama a atenção para a maneira como a questão referencial (anafórica e catafórica), tão cara aos estudos da coerência e coesão textuais, está (ou não) sendo contemplada na taxonomia. Além disso, o estudo de elementos (e relações) entre unidades não adjacentes é um dos principais focos dos trabalhos mais recentes da eRST (Zeldes *et al.*, 2024).

Por fim, a relação *Circumstance* se estrutura sobre um núcleo que apresenta um evento principal e um satélite que apresenta a circunstância (de tempo ou modo, por exemplo) sob a qual o evento é realizado. De maneira geral, os anotadores humanos indicam que essa relação é indicada pela presença de um SD de tipo sintático (*oração circunstancial*). Em consonância com essas observações, os LLMs indicaram oração circunstancial e sinalizadores semânticos (como sentido do verbo ou campo semântico) como pistas de *Circumstance*.

Quadro 4 – Exemplo de anotação automática da relação *Circumstance*

Segmento textual	Modelo	
	GPT-4	LLaMA-4
[EDU 1] Antes de participar do revezamento, [EDU 2] Thiago havia conquistado o ouro na disputa dos 400 metros medley.	oração circunstancial + tempo verbal + preposição + campo semântico	advérbio + tempo verbal

Fonte: elaboração própria a partir da anotação do D2_C38 do CSTNews.

Para o GPT, a EDU 1 é *oração adverbial* que indica circunstância; além disso as expressões *"revezamento"* e *"400 metros medley"* remetem a um *campo semântico* específico. Os modelos divergem quanto à classificação do MD explícito: o GPT aponta *antes de* como preposição, enquanto o LLaMA classifica como *advérbio*. Embora os LLMs tenham demonstrado maior precisão nos casos de subclassificação, se comparado à anotação humana, este caso reitera a necessidade de revisão e análise da anotação.

Vale destacar que a relação *Circumstance* é pouco frequente no *corpus*, o que pode ter influenciado a divergência entre as anotações. Ainda assim, nos cinco casos estudados, o *tempo verbal* parece ser um indicativo relevante para o efeito de *circunstância*. Considerando o Quadro 4, ambos os modelos reconhecem a presença de uma marca por meio do tempo verbal. Para o modelo LLaMA, o uso do pretérito mais-que-perfeito ("*havia conquistado*") em relação com o infinitivo "*participar*" ajuda a estabelecer a sequência temporal e, por conseguinte, a relação circunstancial de tempo. Assim como nos dados anteriormente expostos, o modelo apresenta uma anotação mais enxuta, enquanto o GPT gera, em aparente excesso, os rótulos previstos na taxonomia.

Da anotação dos SDs das relações *Circumstance* destaca-se também o caso em (7):

(7) [EDU 1] Depois que terminou o evento,
[EDU 2] várias pessoas vieram dizer que tinha sido organizado,
[EDU 3] que gente tinha recebido o convite. [D1_C49]

A anotação humana da relação estabelecida entre as EDUs 1 e 2-3, ilustradas no exemplo (7), é apenas de *Circumstance*. No entanto, os dados gerados pelo GPT possibilitam questionar tal anotação realizada já há bastante tempo por Cardoso *et al.* (2011) ao indicar:

oração circunstancial, *conjunção subordinativa*, *verbo de comunicação*, *oração subordinada* e *fonte*. Diante do resultado, verifica-se que o LLM, mesmo sendo informado de que se trata de um texto constituído pela relação *Circumstance*, apontou SDs característicos da relação *Attribution*, ao entender que *várias pessoas* (*fonte*) emitiram uma informação (*mensagem*). Esse tipo de resultado demonstra como os SDs são fundamentais para o estudo das relações de coerência, desde a segmentação das EDUs até a identificação dos efeitos de interpretação.

Em suma, pode-se afirmar que o GPT-4, apesar de alucinações terem sido notadas quanto à *funcionalidade* (indicando palavras ou construções que não exerciam a função de SD, de fato) ou à *pertinência* (indicando sinalizadores que não contribuem para a indicação de determinada relação), apresentou uma anotação mais minuciosa, desencadeando a maior parte das reflexões propostas neste trabalho, além de demonstrar certa convergência ao que foi indicado pelos humanos. É possível que suas limitações possam ser superadas realizando ajustes nos hiperparâmetros (como a *temperatura*, para que os resultados possam ser mais próximos às escolhas humanas e/ou mais criativas/inventivas na identificação de novas pistas) e aprimoramento no treinamento dos modelos, pois, em nenhum momento, foi dado algum *feedback* sobre as escolhas realizadas automaticamente.

Em relação ao LLaMA, embora tenha retornado poucas alucinações, suas indicações foram sempre mais concisas, com a anotação de um pequeno número de SDs para cada relação, incluindo, em grande parte dos casos, a apresentação de sinalizadores únicos para a identificação das relações. Dessa forma, verificou-se que o seu produto converge apenas parcialmente com a anotação humana, que prezou pela anotação de SDs combinados.

5 Considerações finais

Diante dos resultados e das escolhas metodológicas, é possível afirmar que o trabalho conjunto - humano e LLM - é muito produtivo na tarefa de anotação e, principalmente, de sua revisão, proporcionando reflexões e aperfeiçoamento na taxonomia proposta.

Como visto, em determinadas situações, o uso de LLMs (sobretudo o LLaMA) apresentou alternativas de simplificação da taxonomia proposta, anotando a relação *Attribution* apenas com o SD *verbo de comunicação* e a relação *Elaboration*, com o SD *oração relativa*. Além disso, ambos LLMs propuseram pistas não consideradas na anotação manual, a

saber: *exemplificação, conjunção ou oração subordinada* e *elemento referencial*; casos estes que já estão sob análise em trabalhos em desenvolvimento de revisão da taxonomia utilizada.

Por outro lado, deste estudo, destacam-se duas limitações, que se configuram como provocações para estudos posteriores. A primeira diz respeito à amostra utilizada, a qual abordou apenas três das relações RST para o PB, circunscritas apenas em segmentos intrassentenciais. Nesse sentido, em estudos futuros, caberá investigar como os LLMs se comportam diante das outras relações RST, além de ampliar o escopo para relações interssentenciais. A segunda se refere à dificuldade de identificação de SDs em algumas EDUs, se comparada à anotação humana. Uma possibilidade de investigação será analisar se modificações na redação do *prompt* alteram positivamente os resultados dessa identificação. Concomitantemente a isso, poderá ser desenvolvido um sistema via API, que seja pré-treinado com exemplos de anotação de SDs, além do ajuste de hiperparâmetros, permitindo análises mais profundas que permitam extrair mais explicações e descrições do ponto de vista qualitativo.

De modo geral - e conforme os objetivos desta pesquisa - foi possível apresentar os pontos comuns e divergentes entre a anotação humana e a gerada pelos LLMs investigados, além de discutir as particularidades e contribuições da geração de informação apresentada pelos LLMs.

Agradecimentos

Este trabalho foi realizado no Centro de Inteligência Artificial da Universidade de São Paulo (C4AI – http://c4ai.inova.usp.br/), com apoio da Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP, processo nº 2019/07665-4) e da IBM Corporation. O projeto também contou com apoio do Ministério da Ciência, Tecnologia e Inovação, com recursos da Lei nº 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, foi apoiado por recursos financeiros da Chamada PRPPG-UFBA 010/2024 – Programa de Apoio a Docentes/Pesquisadores em Início de Carreira 2024 – e da Chamada FAPESB/CNPq 004/2023 – Programa Primeiros Projetos.

Referências

ACHIAM J. *et al.* GPT-4 technical report. Cornell University. **arXiv**. 10.48550/arXiv.2303.08774, 2024. DOI: https://doi.org/10.48550/arXiv.2303.08774.

ALUÍSIO, S. M.; ALMEIDA, G. M. B. O que é e como se constrói um *corpus*? Lições aprendidas na compilação de vários *corpora* para pesquisa linguística. **Calidoscópio**, v. 4, n. 3, p. 156-178, 2006.

BARBOSA, B. K. S.; CAMPELO, C. E. C. LLMs as Tools for Evaluating Textual Coherence: A Comparative Analysis. *In*: **Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)**. SBC, 2024. p. 278-287. DOI: https://doi.org/10.5753/stil.2024.245379.

CARDOSO, P. C. F. *et al*. 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**. Cuiabá/Brazil, 2011. p. 88-105.

CARDOSO, P. *et al*. A Linguagem em Foco: Anotação de Sinalizadores Discursivos em Textos Jornalísticos. *In*: **Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL)**. SBC, 2024. p. 247-256. DOI: https://doi.org/10.5753/stil.2024.245329.

CHAN, C.; CHENG, J.; WANG, W.; JIANG, Y.; FANG, T.; LIU, X.; SONG, Y. Exploring the potential of CHATGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In: **Findings of the Association for Computational Linguistics: EACL 2024**. 2024. p. 684-721.

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.; RODRIGUES, R. **Manual de anotação de sinalizadores discursivos em textos jornalísticos.** Relatório Técnico do Núcleo Interinstitucional de Linguística Computacional, n. 447, 2024. Disponível em: https://repositorio.usp.br/item/003207370. Acesso em: 20 mar. 2025.

DAS, D.; TABOADA, M. RST Signalling Corpus: A corpus of signals of coherence relations. **Language Resources and Evaluation**, v. 52, p. 149-184, 2018. DOI: https://doi.org/10.1007/s10579-017-9383-x.

DAVIES, M. **The Corpus of Contemporary American English (COCA)**. 2008. Disponível em: https://www.english-corpora.org/coca/. Acesso em: 20 mar. 2025.

HUANG, Q.; TAO, M.; AN, Z.; ZHANG, C.; JIANG, C.; CHEN, Z.; WU, Z.; FENG, Y. Lawyer LLaMA technical report. **arXiv** 10.48550/arXiv.2303.08774, 2023. DOI: DOI: https://doi.org/10.48550/arXiv.2305.15062.

LAI, A.; TETREAULT, J. Discourse coherence in the wild: A dataset evaluation and methods. *In*: **Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue**, 2018. p. 214–223. DOI: https://aclanthology.org/W18-5023/.

MANN, W. C.; THOMPSON, S. A. **Rhetorical structure theory: A theory of text organization**. Los Angeles: University of Southern California, Information Sciences Institute, p. 87-190, 1987.

MIKKELSEN, L. F.; KINCH, O.; PEDERSEN, A. J.; LACROIX, O. Ddisco: A discourse coherence dataset for danish. *In*: **Proceedings of the 13th Language Resources and Evaluation Conference (LREC)**, 2022. p. 1234–1243.

MUERMANS, T. C.; KOSSEIM, L. A BERT-Based approach for multilingual discourse connective detection. *In*: **International Conference on Applications of Natural Language to Information Systems**. Cham: Springer International Publishing, 2022. p. 449-460. https://doi.org/10.1007/978-3-031-08473-7_41.

NASUTION, A. H.; ONAN, A. ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. **IEEE Access**, v. 12, p. 71876-71900, 2024. DOI: https://doi.org/10.1109/ACCESS.2024.3402809.

PARDO, T. A. S. **Métodos para análise discursiva automática**. Tese (Doutorado em Ciência da Computação) – Universidade de São Paulo, São Carlos, 2005. DOI: https://doi.org/10.11606/T.55.2005.tde-29082005-172336.

SIVARAJKUMAR, S.; KELLEY, M.; SAMOLYK-MAZZANTI, A.; VISWESWARAN, S.; WANG, Y. An empirical evaluation of prompting strategies for large language models in zero-shot clinical natural language processing: algorithm development and validation study. **JMIR Medical Informatics**, v. 12, p. e55318, 2024. DOI: https://doi.org/10.2196/55318.

SOUZA, J. W. C.; CARDOSO, P. C. F.; RODRIGUES, R. Taxonomy of Discourse Signals for RST Relations: A Study in a News Corpus. **Revista Brasileira de Linguística Aplicada**, 2025. DOI: https://doi.org/10.1590/1984-6398202549793.

WANG, X.; KIM, H.; RAHMAN, S.; MITRA, K.; MIAO, Z. Human-LLM collaborative annotation through effective verification of LLM labels. *In*: **Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems**. 2024. p. 1-21. DOI: https://doi.org/10.1145/3613904.3641960.

ZELDES, A. 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**. 2016. p. 1-5.

ZELDES, A.; AOYAMA, T.; LIU, Y. J.; PENG, S.; DAS, D.; GESSLER, L. eRST: A Signaled Graph Theory of Discourse Relations and Organization. **Computational Linguistics**, v. 51, n. 1, p. 23-72, 2024. DOI: https://doi.org/10.1162/coli_a_00538.

Recebido em: 08.07.2025 Aprovado em: 09.01.2026