Filol. Linguíst. Port., São Paulo, v.28, n.1, p.1-19, e-240890, 2026
https://dx.doi.org/10.11606/issn.2176-9419.v28i1e-240890

A utilização de LLMs para anotação de relações discursivas e seus meios de sinalização: um teste empírico

The use of large language models for the annotation of discourse relations and its means of signalling: an empirical test

Juliano Desiderato Antonio*
Universidade Estadual de Maringá

Resumo: Neste trabalho, investiga-se a utilização de modelos de linguagem de larga escala (LLMs) para a identificação de relações retóricas e dos meios de sinalização que permitem sua identificação. Fundamentado na *Rhetorical Structure Theory* (RST), que concebe o texto como uma rede de relações funcionais entre unidades discursivas, o estudo avalia a possibilidade da utilização de ferramentas de inteligência artificial na anotação automática de *corpora* discursivos. A metodologia consistiu em reaplicar ao *ChatGPT* dez excertos de um *corpus* de língua falada aplicado a professores universitários em pesquisa anterior, mantendo-se as mesmas perguntas sobre a identificação das relações de sentido e das pistas utilizadas. Os resultados revelam que, em sete dos dez casos, o *ChatGPT* produziu análises coincidentes com a da pesquisa anterior, utilizando rótulos próprios da RST e justificando-os com base em sinais semânticos, formais e pragmáticos. Nos três casos restantes, ainda que divergentes, as respostas do modelo mostraram-se plausíveis do ponto de vista teórico. Conclui-se que o *ChatGPT* apresenta desempenho satisfatório na identificação de relações discursivas, podendo ser considerado uma ferramenta promissora para a anotação automática de *corpora*, desde que suas análises sejam validadas por especialistas humanos.

Palavras-chave: RST. Modelos de Linguagem de Larga Escala. Relações discursivas. Anotação automática. ChatGPT.

Abstract: This study investigates the use of large language models (LLMs) for the identification of rhetorical relations and the signalling means that enable their recognition. Grounded in Rhetorical Structure Theory (RST), which conceives text as a network of functional relations between discourse units, the research evaluates the potential of artificial intelligence tools for the automatic annotation of discourse *corpora*. The methodology consisted of reapplying to *ChatGPT* ten excerpts from a spoken language *corpus* administered to university professors in a previous investigation, keeping the same questions regarding the identification of discourse relations and the cues that signal such relations. The results reveal that, in seven out of ten cases, *ChatGPT* produced analyses coinciding with those of the previous paper, employing labels established in RST and justifying them on the basis of semantic, formal, and pragmatic signals. In the remaining three cases, although divergent, the model's responses proved theoretically plausible. It is concluded that *ChatGPT* demonstrates satisfactory performance in the identification of discourse relations, presenting itself as a promising tool for the automatic annotation of *corpora*, provided that its analyses are validated by human experts.

Keywords: RST. Large Language Models. Discourse relations. Automatic annotation. ChatGPT.

1 Considerações iniciais

A coerência textual depende, dentre outros fatores, das relações estabelecidas entre duas ou mais porções de texto (Sanders; Spooren; Noordman, 1992). Um modelo

* Docente, Universidade Estadual de Maringá, PR, Brasil; e-mail: jdantonio@uem.br ; ORCID: https://orcid.org/0000-0002-9816-5852

que se destaca no estudo dessas relações é a '*Rhetorical Structure Theory*' (RST), uma teoria descritiva que tem por objeto o estudo da organização dos textos, caracterizando as relações que se estabelecem entre suas partes (Mann; Matthiessen; Thompson, 1992; Mann; Thompson, 1988; Matthiessen; Thompson, 1988). Essas relações têm recebido várias denominações nos estudos linguísticos: predicados retóricos (Grimes, 1975), proposições relacionais (Mann; Thompson, 1983), relações de coerência (Hobbs, 1985), relações retóricas (Mann; Thompson, 1988). Neste trabalho, será adotado este último rótulo por sua filiação à RST.

Na perspectiva da RST, as relações retóricas são de sentido e não de forma (Mann; Thompson, 1983), ou seja, as relações são estabelecidas e interpretadas independentemente de serem marcadas explicitamente por conectivos. Diversos estudos foram realizados para identificar outros meios de sinalização das relações além de conectivos e marcadores discursivos. Para Gómez-González e Taboada (2005), além dos marcadores discursivos (itens como 'porque', 'mas', 'embora'), que atuam como instruções de processamento, outros meios gramaticais também servem como sinalização (por exemplo: ordem dos constituintes, categorias verbais como modo e aspecto, entonação *etc*.). Com base na análise de *corpus*, Taboada (2006) demonstra que várias relações ocorrem com maior frequência sem nenhum marcador explícito (por exemplo, elaboração), ao passo que outras relações são sinalizadas com maior frequência por meio de marcadores discursivos (por exemplo, contraste e condição). Em seu texto de 2009, Taboada propõe uma taxonomia dos sinalizadores das relações, que vai dos marcadores discursivos aos sinais gráficos, passando pelos sinais morfossintáticos e pelos sinais semânticos. Antonio (2017) investigou se professores de curso superior seriam capazes de identificar as relações retóricas não sinalizadas por meio de conectivos em excertos de aulas. Em sua grande maioria, as relações retóricas foram identificadas pelos informantes, que indicaram as pistas utilizadas na identificação das relações. As pistas eram formais, fonológicas, morfossintáticas, semânticas, textuais e cognitivas. Das e Taboada (2018) analisaram o RST '*Discourse Treebank*' e propuseram um inventário exaustivo de elementos que sinalizam as relações de coerência.

Com os recentes avanços dos chamados Modelos de Linguagem de Larga Escala (LLMs, em inglês), muitas atividades linguísticas complexas (tradução automática, sumarização automática *etc*.) podem ser realizadas por interfaces como o *ChatGPT*, por exemplo. Diante disso, investiga-se, neste trabalho, se um modelo de Inteligência Artificial (IA) consegue identificar as relações retóricas que se estabelecem entre partes do texto, bem como indicar os sinais utilizados na identificação. Para isso, realizou-se um teste com o *ChatGPT* utilizando-se dez excertos de um *corpus* de língua falada que foi aplicado anteriormente a um grupo de professores universitários (Antonio, 2017). O bom desempenho desse modelo de IA na realização dessas tarefas poderia indicar, por exemplo, a possibilidade de utilizá-lo na anotação discursiva automática de um *corpus*. De acordo com Paes, Vianna e Rodrigues (2024), os LLMs são modelos de linguagem neurais que se diferenciam de outros modelos pré-treinados pelo fato de, dentre outros fatores: a) trabalharem com uma quantidade enorme de parâmetros (estima-se que o *ChatGPT*-4, por exemplo, tenha quase 2 trilhões de parâmetros); b) atuarem como métodos de IA Generativa, cuja função primária é a geração de conteúdo na forma de texto; c) terem capacidade de aprender em contexto, sem treinamento adicional que

atualize seus parâmetros.

> Esses parâmetros podem ser entendidos como pesos de conexões entre neurônios artificiais que são ajustados durante a fase de treinamento. Esses modelos são treinados em grandes conjuntos de dados para aprender uma variedade de padrões e fatos, e são capazes de aplicar esse aprendizado para responder perguntas e realizar outras tarefas de PLN (Cortes; Vieira; Barone, 2023, p. 423).

Em termos de sua estrutura, além destas considerações iniciais, este trabalho apresenta outras três seções. Na fundamentação teórica, apresentam-se os pressupostos básicos da RST. Na seção de metodologia, apresentam-se os procedimentos adotados na realização da pesquisa para que os objetivos fossem alcançados. Os resultados são apresentados e discutidos na seção de número quatro. Por fim, encerra-se o trabalho com as considerações finais.

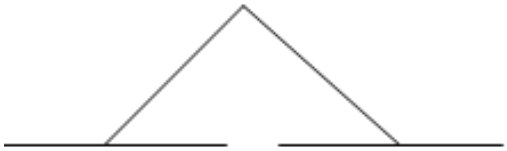
2 Fundamentação teórica

A '*Rhetorical Structure Theory*' (RST), proposta por Mann e Thompson (1988), concebe o texto como uma rede de relações funcionais entre segmentos discursivos. Esses segmentos, denominados '*Elementary Discourse Units*' (EDUs) (Carlson; Marcu, 2001), organizam-se em estruturas hierárquicas representadas por diagramas arbóreos, nos quais cada relação retórica envolve papéis distintos para os constituintes: o núcleo, portador da informação central, e o satélite, que serve de subsídio para o núcleo. Esse tipo de relação, chamado núcleo-satélite, é representado na Figura 1, em que um arco vai do satélite em direção ao núcleo (*cf.* Diagrama 1 da seção 3 deste trabalho). Também há relações multinucleares, nas quais uma porção do texto não é ancilar da outra, sendo cada porção um núcleo distinto, como na Figura 2 (*cf.* Diagrama 2 da seção 3 deste trabalho).

Figura 1 – Relação núcleo-satélite

Fonte: O autor.

Figura 2 – Relação multinuclear



Fonte: O autor.

No texto-fundador da teoria (Mann; Thompson, 1988), os autores propuseram uma lista de vinte e quatro relações após a análise de centenas de textos utilizando a RST. Posteriormente, o *website* da teoria divulgou uma lista com 30 relações (www.sfu.ca/rst). Mann e Thompson (1988) não consideram que essa lista deva ser fechada e outras relações podem ser definidas de acordo com o tipo ou gênero de texto investigado. Dessa forma, outros autores, como Carlson e Marcu (2001), propõem uma lista com 136 relações.

Embora tenha surgido no contexto da linguística computacional (Matthiessen, 2005), com o objetivo inicial de fornecer suporte para a geração automática de textos, a RST rapidamente expandiu seu escopo e consolidou-se como uma abordagem interdisciplinar, encontrando aplicações em áreas como linguística do texto, tradução, ensino de línguas e estudos de aquisição. Souza, Cardoso e Rodrigues (2024), em

revisão sistemática de trabalhos publicados entre 2010 e 2022 que têm como base teórico-metodológica a RST, destacaram três eixos principais:

1. Teoria e descrição: estudos que caracterizam e identificam as relações da RST, além de pesquisas que recuperam o que a literatura propõe como uma teoria para o modelo. Das e Taboada (2013), por exemplo, investigam marcas além de marcadores discursivos que sinalizam as relações retóricas (morfológicas, lexicais, sintáticas, semânticas, gráficas *etc*.). Antonio e Santos (2014) descrevem a estrutura do gênero resposta argumentativa utilizando a RST;
2. Linguística de *Corpus*: trabalhos que exploram, compilam e/ou anotam *corpora* linguísticos de acordo com a RST. Existem diversos *corpora* anotados discursivamente, dentre os quais podem ser citados '*RST Discourse Treebank*' (Carlson *et al.*, 2001), '*RST Spanish Treebank*' (da Cunha *et al.*, 2011), '*Dutch Discourse Treebank*' (Redeker *et al.*, 2012), '*Basque RST Treebank*' (Iruskieta *et al.*, 2013), '*Potsdam Commentary Corpus*' (Stede; Neumann, 2014), '*Georgetown University Multilayer Corpus*' (Zeldes, 2017), '*Russian RST Treebank*' (Toldova *et al.*, 2017); '*Persian RST Corpus*' (Shahmohammadi *et al.*, 2021). Para o português, existe o *corpus* '*CST News Corpus*' (Cardoso *et al.*, 2011) para análise discursiva multidocumento;
3. PLNP: estudos que abordam a RST a partir de aplicações computacionais. Encontram-se, nesse eixo, trabalhos relacionados à análise discursiva automática, sumarização automática, parsing, análise de sentimentos, tradução automática, detecção automática de *fake news*, dentre outras possibilidades.

3 Metodologia

Antonio (2017) realizou uma pesquisa com professores de ensino superior de diversas áreas apresentando trechos de aulas e perguntando que relação emergia da combinação entre porções textuais, bem como os sinais que auxiliaram na identificação da relação. Os excertos foram retirados de um *corpus* formado por aulas de curso superior, por aulas de curso preparatório para o vestibular e por entrevistas com pesquisadores. Na referida pesquisa, a escolha de professores de curso superior como informantes se justifica pelo fato de as elocuções formais do *corpus* terem esse público como produtor dos textos. Tentou-se, dessa forma, evitar que a falta de informação pragmática fosse um fator que atrapalhasse a compreensão das porções de texto apresentadas aos informantes.

Nos dez excertos apresentados aos informantes, as relações retóricas não eram sinalizadas por conectores, mas por outros meios, como pontuação, correlação modo-temporal, conteúdo das porções textuais, paralelismo sintático, paráfrase, inserção parentética, repetição, apresentação de evidências *etc*., com a finalidade de verificar se os informantes reconheceriam essas relações por meio dessas outras pistas.

Procurou-se verificar se a falta de sinalização por meio de um conector impediria ou dificultaria a identificação da relação e também se os informantes saberiam explicitar outros meios além dos conectivos que servem como pista para identificação da relação. A pergunta feita aos informantes foi "Que relação de sentido há entre os enunciados (1) e (2)?". Uma outra pergunta também foi feita para tentar verificar que meios foram

mobilizados pelos informantes para identificação da relação. A pergunta foi a seguinte: "O que ajudou você a identificar essa relação?".

Para este trabalho, utilizaram-se os mesmos dez excertos e as mesmas perguntas, mas, desta vez, as perguntas foram dirigidas ao *ChatGPT*, em sua versão GPT-5. Apenas uma modificação foi feita na primeira pergunta, à qual se acrescentou o trecho em itálico "*Considerando a Rhetorical Structure Theory*, que relação de sentido há entre os enunciados (1) e (2)?". Na próxima seção, analisa-se a performance da ferramenta de IA na identificação das relações e dos sinais que auxiliam em sua identificação.

Törnberg (2024) propõe um conjunto de boas práticas fundamentais para garantir que o uso de LMs na anotação de textos seja confiável, reproduzível e ético. Dentre essas boas práticas estão a escolha de um modelo adequado, um procedimento sistemático de codificação (em que se examinam pontos de discordância com anotadores humanos e se pede que o modelo explique suas respostas), uma análise de estabilidade, verificando se o modelo retorna o mesmo resultado para o mesmo texto em múltiplas execuções ou se pequenas variações no *prompt* alteram significativamente a resposta, dentre outros.

4 Resultados e discussão

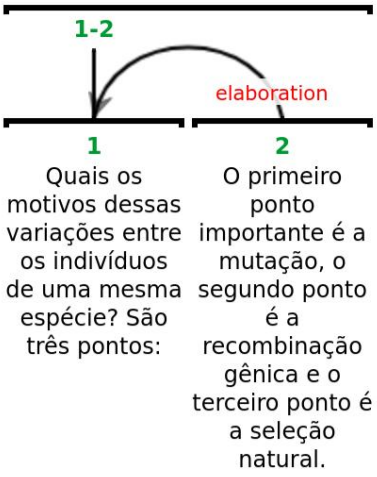
No primeiro excerto, reproduzido abaixo, a relação que emerge da combinação entre as porções de texto (1) e (2) é a de 'elaboração' (*elaboration*), como pode ser observado no diagrama 1 da Figura 3. Utiliza-se a ferramenta *rstWeb* (Zeldes, 2016) para a criação dos diagramas. Como a ferramenta é disponibilizada na língua inglesa, os rótulos da relação aparecem nessa língua. Na pesquisa anterior, todos os informantes afirmaram que o satélite – unidade (2) – acrescenta informações ao conteúdo do núcleo – unidade (1). Essa é justamente a definição da relação de elaboração.

Excerto 1

(1) Quais os motivos dessas variações entre os indivíduos de uma mesma espécie? São três pontos:

(2) O primeiro ponto importante é a mutação, o segundo ponto é a recombinação gênica e o terceiro ponto é a seleção natural.

Figura 3 – Diagrama 1: Relação de elaboração



Fonte: O autor.

Com relação às pistas utilizadas na identificação da relação, os informantes apontaram a sinalização gráfica (o sinal de dois pontos), a estratégia comunicativa (a pergunta retórica e o 'anúncio' de que os 'três pontos' seriam enumerados), a prosódia (um informante leu o trecho em voz alta e se baseou na sua entonação para realizar a identificação).

O modelo respondeu mencionando o rótulo da relação, ou seja, elaboração. No que diz respeito às pistas, o modelo informou que identificou a relação pelo fato de a primeira unidade anunciar que seriam três pontos e a segunda unidade efetivamente apresentar esses três pontos. Outra pista, segundo o modelo, é a dependência informacional de uma unidade em relação à outra. Embora pareça tentador tratar as respostas do modelo como evidência suficiente de interpretabilidade, é importante fazer uma ressalva com relação à computação realizada pelo modelo, conhecida como *Chain-of-Thougt* (CoT). Para Barez *et al.* (2025), o CoT é uma técnica que permite aos LLMs gerar uma sequência de etapas intermediárias de raciocínio antes de chegarem a uma resposta final. Embora o CoT melhore significativamente o desempenho em tarefas complexas, como matemática e raciocínio de senso comum, Barez *et al.* (2025) argumentam que ele oferece apenas uma impressão de transparência, não sendo uma garantia de explicabilidade real.

No segundo excerto, emerge da combinação entre as porções textuais (1) e (2) a relação de 'contraste' (*contrast*), utilizada para contrapor as diferenças entre as teorias heterotrófica e autotrófica. Essa relação é multinuclear e não pode apresentar mais de dois núcleos (diagrama 2, apresentado na Figura 4).

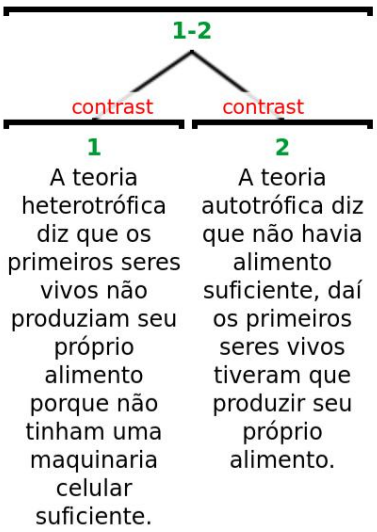
Excerto 2

(1) A teoria heterotrófica diz que os primeiros seres vivos não produziam seu próprio alimento porque não tinham uma maquinaria celular suficiente.

(2) A teoria autotrófica diz que não havia alimento suficiente, daí os primeiros seres

vivos tiveram que produzir seu próprio alimento.

Figura 4 – Diagrama 2: Relação de contraste



Fonte: O autor.

Na pesquisa anterior, a noção de oposição da relação de contraste foi identificada corretamente por todos os informantes, que se apoiaram, em sua maioria, em um critério semântico para a identificação da relação, a saber, o conteúdo antagônico das porções textuais. Outras pistas (mais formais) utilizadas foram o paralelismo das porções textuais que se contradizem e os prefixos **auto-** e **hetero-**, presentes nos vocábulos 'autotrófico' e 'heterotrófico'.

Por sua vez, o modelo utilizou o rótulo da RST para nomear a relação: contraste. Como pista de identificação, assim como os informantes, o modelo também mencionou ter utilizado o critério semântico (conteúdo proposicional oposto) e o critério estrutural (simetria estrutural entre as porções de texto).

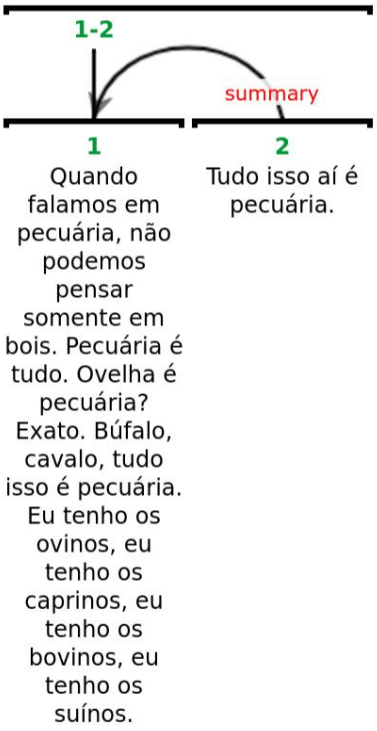
A relação que emerge da combinação entre as porções textuais do excerto 3 é a de 'resumo' (*summary*) (diagrama 3, apresentado na Figura 5).

Excerto 3

(1) Quando falamos em pecuária, não podemos pensar somente em bois. Pecuária é tudo. Ovelha é pecuária? Exato. Búfalo, cavalo, tudo isso é pecuária. Eu tenho os ovinos, eu tenho os caprinos, eu tenho os bovinos, eu tenho os suínos.

(2) Tudo isso aí é pecuária.

Figura 5 – Diagrama 3: Relação de resumo



Fonte: O autor.

Na pesquisa anterior, os informantes responderam que, na unidade (2), havia alguma informação a ser ressaltada em relação ao conteúdo do núcleo. Alguns disseram que era resumo, outros, conclusão. As pistas indicadas pelos informantes foram textuais (a expressão referencial anafórica "tudo isso aí") e cognitivas (o campo semântico da pecuária).

A resposta do modelo foi a relação de resumo. E as pistas que o modelo afirmou ter utilizado utilizadas para a identificação, segundo o GPT-5, foram a repetição condensada do conteúdo da porção textual (1) pela porção (2) (essa pista é compatível com a relação de resumo); a dependência informacional, por se tratar de uma relação núcleo-satélite; e, por fim, o modelo menciona uma pista típica da língua falada, o fechamento do tópico discursivo.

No excerto 4, reproduzido abaixo, a análise apresentada pelo modelo foi diferente da análise de Antonio (2017). Isso não é um problema, porque a identificação das relações pelo analista é guiada por julgamentos funcionais e semânticos, que buscam identificar a função de cada porção de texto e verificar como o texto produz o efeito desejado em seu possível receptor. Esses julgamentos são de plausibilidade, pois o analista tem acesso ao texto, tem conhecimento do contexto em que o texto foi produzido e das convenções culturais do produtor do texto e de seus possíveis receptores, mas não tem acesso direto ao produtor do texto ou aos seus possíveis receptores, de forma que não pode afirmar com certeza que esta ou aquela análise é a correta, mas pode sugerir uma análise plausível (Mann; Thompson, 1988).

Excerto 4

Pegam uma sementinha da soja,

(1) começam a manipular,

(2) manipular,

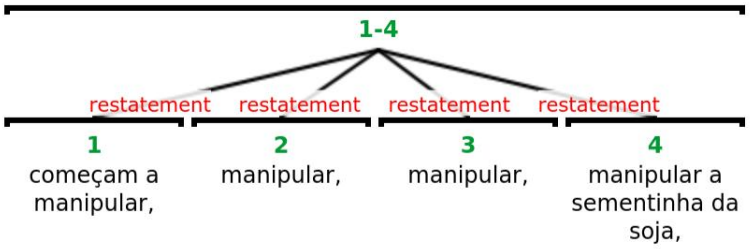
(3) manipular,

(4) manipular a sementinha da soja,

e começam a plantar soja em regiões de clima hostil.

Na análise de Antonio (2017), a relação que se estabelece entre as porções de (1) a (4) é de 'reafirmação nuclear' ('*multinuclear restatement*') (diagrama 4a, Figura 6). A realização dessa relação se dá por meio da repetição que, segundo Marcuschi (2006), é um importante fenômeno da língua falada que auxilia na coesão e na continuidade tópica. Especificamente nesse excerto, a repetição tem motivação icônica, pois a repetição do verbo 'manipular' indica que se tratou de um processo longo.

Figura 6 – Diagrama 4a: Relação de reafirmação multinuclear



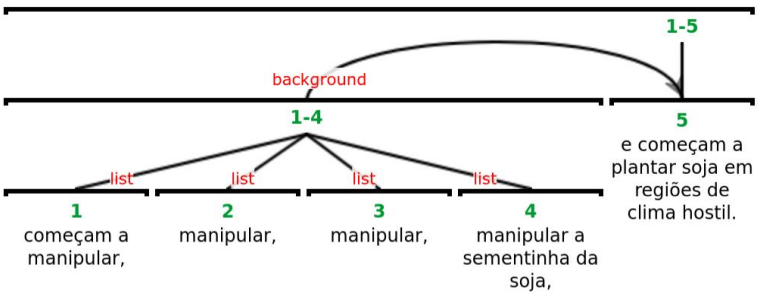
Fonte: O autor.

Na pesquisa anterior, dos dez informantes, nove responderam que o produtor do texto utilizou a repetição para destacar a duração do processo de manipulação das sementes de soja, ao passo que um informante não conseguiu explicar a função da repetição no trecho em análise.

Já na análise do modelo, estabelece-se entre as unidades de (1) a (4) uma relação multinuclear de 'lista' (*list*) (diagrama 4b, Figura 7). Essas quatro unidades formam, então, uma porção textual que funciona como satélite de fundo (*background*) para a última unidade do trecho, a saber, "e começam a plantar soja em regiões de clima hostil". Na relação de fundo, o satélite traz informações sem as quais ficaria difícil para o destinatário compreender as informações contidas no núcleo. O modelo respondeu que utilizou como pistas o paralelismo sintático (para a relação de lista) e a mudança de função informativa da porção textual formada pelas unidades de (1) a (4) em relação à

última unidade.

Figura 7 – Diagrama 4b: Relações de lista e fundo



Fonte: O autor.

Duas análises são plausíveis para o excerto (5): 'condição' (*condition*) ou 'causa-consequência'. Segundo Neves (2000), as construções causais e as construções condicionais fazem parte de um mesmo contínuo semântico juntamente com as construções concessivas. Há um extremo em que a relação de causa é afirmada (construções causais), um extremo em que o vínculo causal entre as orações é negado (construções concessivas) e um espaço intermediário em que a relação de causa entre as orações é hipotetizada (condicionais). Na pesquisa anterior, as respostas dos informantes contemplaram essas duas relações possíveis, o mesmo ocorrendo com o modelo.

Excerto 5

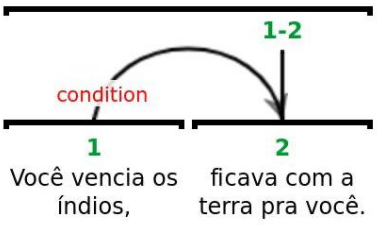
Antigamente no Brasil era assim: você chegava e pegava a terra, não tinha lei.

(1) Você vencia os índios,

(2) ficava com a terra pra você.

As pistas apontadas pelos informantes foram a interdependência entre os eventos e a ordem das orações. O tempo verbal (pretérito imperfeito) foi apontado por um informante como uma pista que o ajudou na identificação da relação, e um outro informante parafraseou a construção iniciando-a com o conectivo 'se'. O modelo respondeu que utilizou as mesmas pistas foram utilizadas pelo GPT-5: ordem sequencial, lógica pragmática/conhecimento de mundo, potencial para paráfrase condicional.

Figura 8 – Diagrama 5: Relações de lista e fundo



Fonte: O autor.

Os excertos 6 e 7 trazem ocorrências de orações adverbiais com formas verbais

não finitas (orações adverbiais reduzidas de gerúndio, utilizando-se os termos da Gramática Tradicional). Braga (2002) realizou um estudo a respeito das orações de gerúndio no português falado no Brasil, que encontrou, no *corpus* investigado, relações aditivas, adjetivas, causais, concessivas, condicionais, consequenciais, modais, temporais e temporais-condicionais. No entanto, segundo a autora, "a identificação da relação semântica codificada pelas orações de gerúndio é muitas vezes problemática, já que elas tendem a favorecer a superposição de relações proposicionais" (p. 242).

Excerto 6

(1) Aí eu posso interferir então naquela fórmula

(2) diminuindo a velocidade de sedimentação.

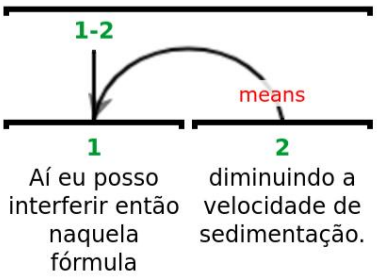
Excerto 7

(1) Essa célula cancerígena perde o controle, o organismo não tem mais controle sobre ela e ela começa a se proliferar de forma desordenada,

(2) formando o melanoma.

No caso do excerto 6, a relação que emerge da combinação entre as porções (1) e (2) é a de 'meio' (*means*), em que o satélite apresenta um método ou instrumento que tende a fazer a realização do núcleo mais provável (diagrama 6 apresentado na Figura 9).

Figura 9 – Diagrama 6: Relações de meio



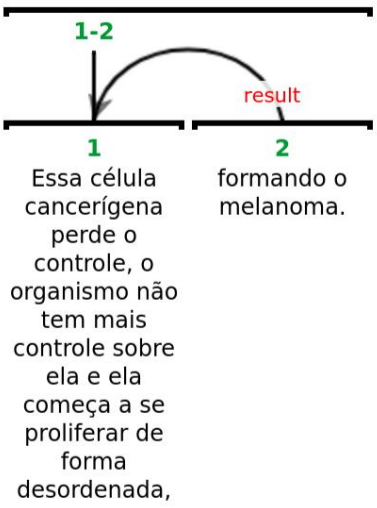
Fonte: O autor.

Na pesquisa anterior, apenas dois informantes identificaram a relação de forma adequada utilizando os rótulos 'maneira' ou 'modo'. Os demais apontaram relações que não parecem ser plausíveis para a análise do excerto: causa-consequência, explicação, solução e condição. Provavelmente isso ocorre, como afirma Braga, pela superposição de relações, tornando a identificação problemática. Já o modelo identificou corretamente a relação de meio e apontou ter utilizado como pistas a estrutura sintática (gerúndio), a relação de dependência entre as orações e potencial de paráfrase utilizando-se a expressão 'por meio de'.

No excerto 7, a relação que emerge da combinação entre as unidades (1) e (2) é a de 'resultado' (*result*), segundo a qual o evento do satélite é causado pelo evento do núcleo (diagrama 7, Figura 10). Ou seja, trata-se de uma relação pertencente ao domínio

semântico da causa. A diferença é que nesta última o evento do núcleo é que causa o evento do satélite.

Figura 10 – Diagrama 7: Relação de resultado



Fonte: O autor.

Na pesquisa anterior, sete informantes apontaram a relação de resultado, rotulando-a como causa-consequência. Os outros três informantes disseram que a relação era de explicação, não plausível no caso do excerto analisado. A pista utilizada para identificação da relação, segundo os informantes, foi a interdependência entre os eventos codificados pelas orações.

O modelo respondeu corretamente que a relação é a de resultado. As pistas que o modelo respondeu ter utilizado para a identificação foram a forma verbal (gerúndio), a progressão lógica do processo, o conhecimento de mundo e a possibilidade de paráfrase por expressões resultativas como 'e', 'como resultado'.

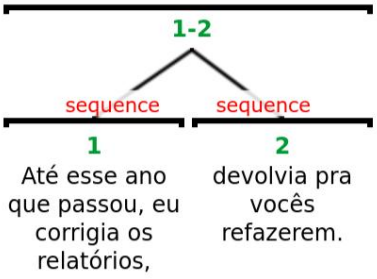
No excerto 8, a relação que emerge da combinação entre as porções de texto é a de 'sequência' (*sequence*), uma relação multinuclear na qual há sucessão temporal entre os eventos dos núcleos (diagrama 8, representado na Figura 11).

Excerto 8

(1) Até esse ano que passou, eu corrigia os relatórios,

(2) devolvia pra vocês refazerem.

Figura 11 – Diagrama 8: Relação de sequência



Fonte: O autor.

Na pesquisa anterior, a relação foi identificada corretamente por seis dos dez informantes. Quatro informantes mencionaram relações não plausíveis para o excerto (condição, consequência, correção). As pistas indicadas pelos informantes foram o conteúdo do próprio texto (uma vez que um evento era anterior ao outro), a expressão temporal "até esse ano que passou" e o tempo verbal pretérito perfeito ('passou').

O modelo identificou corretamente a relação de sequência e respondeu ter utilizado os seguintes sinais como pistas: mesma forma verbal no pretérito imperfeito ('corrigia', 'devolvia'), ordem natural das ações, ausência de conectores subordinativos (descartando uma relação do tipo núcleo-satélite), possibilidade de paráfrase utilizando-se marcador de sequência ('depois').

No excerto 9, a relação que emerge da combinação entre as porções textuais é a de 'reformulação núcleo-satélite' (*N-S restatement*), na qual o falante faz uma reformulação da porção textual anterior (diagrama 9a, Figura 12). Nos termos da 'Perspectiva Textual Interativa', é um caso de parafraseamento, uma estratégia da língua falada na qual o falante usa o discurso anterior como matriz para o novo enunciado (Fávero; Andrade; Aquino, 2006). Segundo Hilgert (2006), o parafraseamento ajuda o falante a encontrar o vocábulo ou expressão mais apropriados para o conteúdo que deseja veicular.

Excerto 9

(1) O organismo heterotrófico não produz seu próprio alimento,

(2) Tem que obter esse alimento do meio.

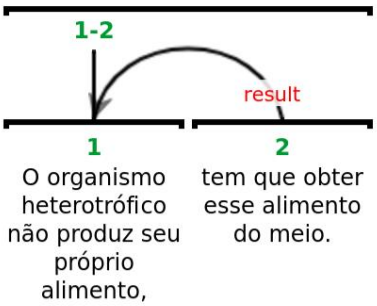
Figura 12 – Diagrama 9a: Relação de reformulação núcleo-satélite

Fonte: O autor.

A função de reformulação da porção textual que funciona como satélite foi identificada por oito dos dez informantes da pesquisa anterior. As pistas utilizadas foram, segundo eles: a) o conteúdo das porções textuais, uma vez que o satélite explica o conteúdo do núcleo; b) a perífrase modal deôntica 'ter que'. Os outros dois informantes consideraram que se tratava de uma relação de causa-consequência. Essa análise é plausível para o excerto em tela, uma vez que, no entendimento desses informantes, ter que obter alimento do meio é consequência de não se produzir seu próprio alimento.

Assim como esses dois informantes, o modelo também identificou a relação como sendo de 'resultado' (*result*), uma vez que o evento do satélite ("tem que obter esse alimento do meio") é consequência do evento do núcleo (o fato de o organismo heterotrófico não produzir seu próprio alimento) (diagrama 9b, Figura 13). O modelo respondeu que utilizou as seguintes pistas para a identificação da relação de resultado nesse excerto: progressão lógica causa-efeito, parafraseabilidade com conectores de consequência (como 'portanto', por exemplo), ausência de gerúndio ou marcadores de condição (indica que não é meio nem condição, mas consequência direta).

Figura 13 – Diagrama 9b: Relação de resultado



Fonte: O autor.

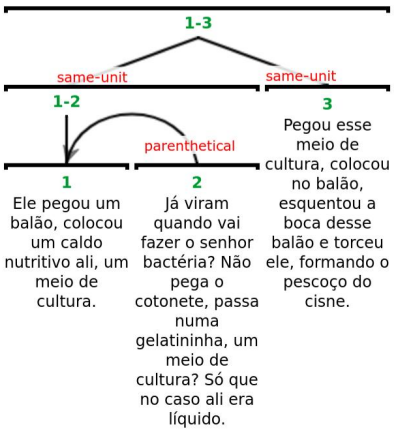
Por fim, no excerto 10, a relação 'parentética' (*parenthetical*) emerge da combinação entre as porções textuais (diagrama 10a, ver Figura 14). Nos termos da 'Perspectiva Textual Interativa', as inserções parentéticas são desvios do tópico discursivo que trazem informação paralela sobre o conteúdo do tópico discursivo,

sobre a expressão linguística do tópico discursivo ou sobre o contexto comunicativo (Jubran, 2006). Essa relação não faz parte do rol de relações clássicas da RST, mas foi definida posteriormente por Carlson e Marcu (2001), que consideram que o efeito dessa relação é que o destinatário reconheça que o satélite apresenta informação extra referente ao núcleo, complementando o núcleo. No caso do excerto 10, o professor utiliza a inserção parentética para exemplificar para os alunos como Louis Pasteur realizou uma experiência. Para isso, menciona o "senhor bactéria", que ele pressupõe que os alunos conheçam dos programas de televisão. Terminado o conteúdo parentético, o professor retoma o fluxo do que vinha explicando por meio da relação de retomada (*same unit*).

Excerto 10

> Louis Pasteur é um grande laboratorista, tem um monte de instrumentos de laboratório que tem o nome dele. Então ele sabia dominar essa técnica de vidraria.
>
> O que ele fez?
>
> (1) Ele pegou um balão, colocou um caldo nutritivo ali, um meio de cultura.
>
> (2) Já viram quando vai fazer o senhor bactéria? Não pega o cotonete, passa numa gelatininha, um meio de cultura? Só que no caso ali era líquido.
>
> Pegou esse meio de cultura, colocou no balão, esquentou a boca desse balão e torceu ele, formando o pescoço do cisne.

Figura 14 – Diagrama 10a – Relação de parentética

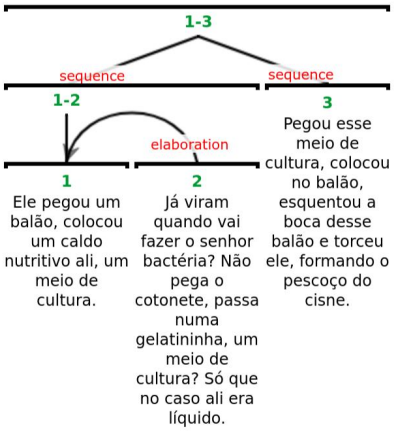


Fonte: O autor.

Nove dos dez informantes da pesquisa anterior demonstraram ter compreendido a função da porção textual que funciona como satélite parentético, mencionando que aquela porção textual servia para introduzir uma explicação mais próxima do conhecimento dos alunos ou para introduzir um exemplo. Um informante forneceu uma resposta não plausível, a relação de resultado. A maioria dos informantes afirma ter se baseado no conteúdo proposicional dos enunciados para realizar a identificação e um informante informou que se pautou pela pontuação, ao passo que três informantes não souberam informar como realizaram a identificação da relação.

Assim como os nove informantes que compreenderam a função do satélite, o modelo apontou a relação de 'elaboração' (*elaboration*) exercendo a função de apresentar um exemplo (diagrama 10b, *cf.* Figura 15). Talvez isso tenha acontecido pelo fato de o modelo ter se baseado apenas no rol de relações clássicas. E as pistas que o modelo indicou ter utilizado foram as seguintes: a) marcadores discursivos de explicação ou exemplificação (pergunta retórica "já viram quando vai fazer..."); b) função de analogia/exemplo; c) dependência semântica; d) parafraseabilidade por conectores como 'quando'.

Figura 15 – Diagrama 10b: Relação de elaboração



Fonte: O autor.

5 Considerações finais

Este trabalho teve como objetivo verificar se um modelo de linguagem de larga escala é capaz de identificar relações retóricas entre porções textuais e indicar os meios de sinalização utilizados na identificação. A partir da comparação entre as respostas do modelo e as de professores universitários que serviram de informantes em uma pesquisa realizada anteriormente, foi possível observar que o modelo apresentou desempenho bastante satisfatório. Na maioria dos casos, utilizou rótulos da própria RST para nomear as relações e indicou pistas formais, semânticas e pragmáticas como responsáveis pela identificação das relações.

Em sete dos dez excertos, as análises fornecidas pelo modelo coincidiram com as análises do trabalho anterior (Antonio, 2017). Nas três situações em que o modelo apresentou análises diferentes, as interpretações eram plausíveis do ponto de vista da RST, reforçando a natureza interpretativa e não determinística da anotação discursiva.

Os resultados obtidos demonstram que é possível utilizar os LLMs para anotação discursiva de *corpora*, permitindo uma grande economia de tempo. Obviamente, cabe aos pesquisadores, ao final, validarem (ou não) a anotação realizada pelas ferramentas de IA. Dessa forma, em um próximo trabalho, pretende-se verificar o desempenho de modelos de IA na anotação de um *corpus* mais extenso, uma vez que, nesta pesquisa, os trechos anotados eram curtos.

Declaração de uso de inteligência artificial

O autor declara que não foi feito uso de IA na redação do texto nem na criação de figuras.

Recebido em setembro de 2025
Publicado em maio de 2026

Referências

ANTONIO, J. D. Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores. **DELTA**: Documentação de Estudos em Lingüística Teórica e Aplicada, v. 33, n. 1, p. 79-108, 2017.

ANTONIO, J. D.; SANTOS, J. A. A estrutura retórica do gênero resposta argumentativa. **Signum: Estudos da Linguagem**, v. 17, n. 2, p. 193-223, 2014.

BAREZ, F. *et al*. Chain-of-thought is not explainability. Preprint, **alphaXiv**, 2025. Disponível em https://www.alphaxiv.org/overview/2025.02v1. Acesso em: 18 mar. 2026.

BRAGA, M. L. Processos de redução: o caso das orações de gerúndio. *In*: KOCH, I. G. V. (org.). **Gramática do português falado**: desenvolvimentos. Campinas: Ed. da Unicamp, 2002. v. 6, p. 239-258.

CARDOSO, P. C. F. *et al*. CSTNews: a discourse-annotated corpus for single and multi-document summarization in Brazilian Portuguese. **NILC Technical Reports**, University of São Paulo, 2011.

CARLSON, L.; MARCU, D. **Discourse tagging reference manual**. Los Angeles: University of Southern California, 2001.

CARLSON, L.; MARCU, D.; OKUROWSKI, M. E. RST Discourse Treebank. Philadelphia: **Linguistic Data Consortium**, 2001.

CORTES, E.; VIEIRA, R.; BARONE, D. Perguntas e respostas. *In*: CASELI, H. M.; NUNES, M. G. V. (org.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 3. ed. São Carlos: BPLN, 2023. p. 416-439.

CUNHA, I. da; TORRES-MORENO, J.-M.; SIERRA, G. On the development of the RST Spanish treebank. *In*: LINGUISTIC ANNOTATION WORKSHOP, 5., 2011. **Proceedings** [...]. Stroudsburg: Association for Computational Linguistics, 2011. p. 1-10.

DAS, D.; TABOADA, M. RST signalling corpus: a corpus of signals of coherence relations. **Language Resources and Evaluation**, Dordrecht, v. 52, n. 1, p. 149-184, 2018.

FÁVERO, L. L.; ANDRADE, M. L. C. V. O.; AQUINO, Z. G. O. O par dialógico pergunta–resposta. *In*: JUBRAN, C. C. A. S.; KOCH, I. G. V. (org.). **Gramática do português culto falado no Brasil**: construção do texto falado. Campinas: Ed. da

Unicamp, 2006. v. 1, p. 133-166.

GÓMEZ-GONZÁLEZ, M. A.; TABOADA, M. Coherence relations in functional discourse grammar. *In*: MACKENZIE, J. L.; GÓMEZ-GONZÁLEZ, M. A. (ed.). **Studies in functional discourse grammar**. Berne: Peter Lang, 2005. p. 227-259.

GRIMES, J. **The thread of discourse**. The Hague: Mouton, 1975.

HILGERT, J. G. Parafraseamento. *In*: JUBRAN, C. C. A. S.; KOCH, I. G. V. (org.). **Gramática do português culto falado no Brasil**: construção do texto falado. Campinas: Ed. da Unicamp, 2006. v. 1, p. 255-273.

HOBBS, J. R. **On the coherence and structure of discourse**. Stanford: CSLI, 1985. (Report n. 35-87).

IRUSKIETA, M. *et al*. The RST Basque TreeBank: an online search interface to check rhetorical relations. *In*: RST AND DISCOURSE STUDIES WORKSHOP, 4., 2013. **Proceedings** [...]. s.l.: s.n., 2013. p. 40-49.

JUBRAN, C. C. A. S. Parentetização. *In*: JUBRAN, C. C. A. S.; KOCH, I. G. V. (org.). **Gramática do português culto falado no Brasil**: construção do texto falado. Campinas: Ed. da Unicamp, 2006. v. 1, p. 301-357.

MANN, W. C.; THOMPSON, S. A. **Relational propositions in discourse**. Marina del Rey: ISI, 1983. (ISI/RR-83-115).

MANN, W. C.; THOMPSON, S. A. Rhetorical structure theory: toward a functional theory of text organization. **Text**, v. 8, n. 3, p. 243-281, 1988.

MANN, W. C.; MATTHIESSEN, C. M. I. M.; THOMPSON, S. A. Rhetorical structure theory and text analysis. *In*: MANN, W. C.; THOMPSON, S. A. (ed.). **Discourse description**: diverse linguistic analyses of a fund-raising text. Amsterdam: John Benjamins, 1992. p. 39-77.

MATTHIESSEN, C.; THOMPSON, S. The structure of discourse and 'subordination'. *In*: HAIMAN, J.; THOMPSON, S. (ed.). **Clause combining in grammar and discourse**. Amsterdam: John Benjamins, 1988. p. 275-329.

MARCUSCHI, L. A. Repetição. *In*: JUBRAN, C. C. A. S.; KOCH, I. G. V. (org.). **Gramática do português culto falado no Brasil**: construção do texto falado. Campinas: Ed. da Unicamp, 2006. v. 1, p. 219-254.

NEVES, M. H. M. **Gramática de usos do português**. São Paulo: Ed. da Unesp, 2000.

PAES, A.; VIANNA, D.; RODRIGUES, J. Modelos de linguagem. *In*: CASELI, H. M.; NUNES, M. G. V. (org.). **Processamento de linguagem natural**: conceitos, técnicas e aplicações em português. 3. ed. São Carlos: BPLN, 2023. p. 385-414.

REDEKER, G. *et al*. **Multi-layer discourse annotation of a Dutch text corpus**. Paris: ELRA, 2012.

SANDERS, T. J. M.; SPOOREN, W. P. M.; NOORDMAN, L. G. M. Toward a taxonomy of coherence relations. **Discourse Processes**, v. 15, n. 1, p. 1-35, 1992.

SOUZA, J. W. C.; CARDOSO, P. C. F.; RODRIGUES, R. Systematic review of studies on rhetorical structure theory (RST). **Revista de Estudos da Linguagem**, v. 31, n. 3, p.

1643-1675, 2024.

TABOADA, M. Discourse markers as signals (or not) of rhetorical relations. **Journal of Pragmatics**, v. 38, n. 4, p. 567-592, 2006.

TABOADA, M. Implicit and explicit coherence relations. *In*: RENKEMA, J. (ed.). **Discourse, of course**. Amsterdam: John Benjamins, 2009. p. 127-140.

TABOADA, M.; DAS, D. Annotation upon annotation: adding signalling information to a corpus of discourse relations. **Dialogue & Discourse**, v. 4, n. 2, p. 249-281, 2013.

SHAHMOHAMMADI, M. *et al*. PrunedRST: a large-scale RST treebank for Persian with an optimized annotation scheme. **arXiv**, Ithaca, 2021. Preprint. Disponível em: https://arxiv.org/abs/2102.03003. Acesso em: 18 mar. 2026.

STEDE, M.; NEUMANN, A. Potsdam Commentary Corpus 2.0: annotation for discourse research. *In*: LANGUAGE RESOURCES AND EVALUATION CONFERENCE, 9., 2014. **Proceedings** [...]. Paris: ELRA, 2014. p. 925-929.

TOLDOVA, S. *et al*. Rhetorical relations markers in Russian RST Treebank. *In*: RECENT ADVANCES IN RST AND RELATED FORMALISMS, 6., 2017. **Proceedings** [...]. s.l.: s.n., 2017. p. 29-33.

TÖRNBERG, P. Best Practices for Text Annotation with Large Language Models. **arXiv**, 2024. Disponível em: https://arxiv.org/abs/2402.05129. Acesso em: 18 mar. 2026.

ZELDES, A. rstWeb - A Browser-based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. *In*: NAACL-HLT 2016. **Proceedings** [...]. San Diego, 2016. San Diego. p. 1-5.

ZELDES, A. The GUM corpus: creating multilayer resources in the classroom. **Language Resources and Evaluatio**n, v. 51, n. 3, p. 581-612, 2017.