

# Taxonomy of Discourse Signals for RST Relations: A Study in a News Corpus

## *Taxonomia de sinalizadores discursivos de relações RST: um estudo em um corpus jornalístico*

**Jackson Wilke da Cruz Souza\***

\*Universidade Federal da Bahia (UFBA)  
Salvador | BA | BR  
jackcruzsouza@gmail.com  
<https://orcid.org/0000-0003-1881-6780>

**Paula Christina Figueira  
Cardoso\*\***

\*\*Universidade Federal do Pará (UFPA)  
Belém | PA | BR  
paulastm@gmail.com  
<https://orcid.org/0000-0003-3621-8960>

**Roana Rodrigues\*\*\***

\*\*\*Universidade Federal de Sergipe (UFS)  
São Cristóvão | SE | BR  
roana@academico.ufs.br  
<https://orcid.org/0000-0002-7748-8716>

**Abstract:** We have observed that the relationship between two or more textual propositions is not always signaled by prepositions or conjunctions, commonly referred to as Discourse Markers (DMs). Recent studies in English, grounded in the Rhetorical Structure Theory (RST), have compiled a list of linguistic and structural cues that facilitate the manual and automatic identification of coherence relations. However, there is a scarcity of similar research for the Portuguese language. Through an analysis of the CSTNews *corpus*, we identified discourse signals (DSs) of RST relations that go beyond DMs. In this study, we propose a classification of these DSs into five categories: *graphical*, *syntactic*, *morphological*, and *discourse markers*, and we describe the occurrence of each category according to the identified relationships.

**Keywords:** Discourse signals; Rhetorical Structure Theory; Coherence relations.

**Resumo:** Observamos que a relação entre duas ou mais proposições textuais nem sempre é marcada por preposições ou conjunções, comumente denominadas por Marcadores Discursivos (MDs). Estudos recentes sobre a língua inglesa, baseados na *Rhetorical Structure Theory* (RST), elaboraram uma lista de pistas linguísticas e estruturais que facilitam a identificação manual e automática de relações de coerência. No entanto, há uma escassez de pesquisas similares para a língua portuguesa. A partir de um estudo no *corpus* jornalístico CSTNews, identificamos sinalizadores discursivos (SDs) das relações RST que transcendem os MDs. Assim, neste trabalho apresentamos a primeira proposta de

taxonomia desses SDs, organizados nos tipos *gráfico*, *semântico*, *sintático*, *morfológico* e *marcador discursivo*, além de descrevermos a ocorrência de cada um deles conforme as relações identificadas.

**Palavras-chave:** Sinalizadores discursivos; Rhetorical Structure Theory; Relações de coerência.

## 1 Introduction

Among the various linguistic levels, Discourse is one of the most challenging for Natural Language Processing (NLP), also known as Computational Linguistics. In NLP, Discourse can be defined as any connected segment of text or speech, comprising one or more sentences (Sidner, 1978). This definition seems pertinent when considering different works that computationally process various textual genres, linguistic modalities, and discourse sizes.

Mitkov (2022) argues that discourse is not a random selection of symbols or words but should be regarded as a linguistic level that deals with related and significant elements possessing specific communicative objectives. This authorial intentionality concerning discourse in each communicative situation highlights the complexity of automatic processing. There is a confluence and influence of other linguistic levels and phenomena in discourse: how was a particular morphological choice arrived at? Why did the author decide to construct a certain sentence using the active voice? How can one explain the presence of optional punctuation in a construction? Is it possible to determine which presuppositions and implied meanings are activated in the production and circulation of a text?

Such questions are not exclusive to NLP but originate in linguistic studies. In this sense, there is a common concern between these fields of knowledge: the way relationships between the elements of a text are constructed, given that the production of a text presupposes a process of interaction and intentions among the subjects involved in each communicative situation (Rodrigues; Souza; Cardoso, 2023).

One of the theoretical approaches for modeling and processing discourse, which has significant computational adherence, is Rhetorical Structure Theory (RST) (Mann; Thompson, 1988). In the 1980s at the University of Southern California, RST was developed by a group of researchers, primarily William Mann and Sandra Thompson, who were interested in natural language generation. According to Taboada and Mann (2006), RST's authors initially aimed to create a model to assist in automating text generation. However, it has since been adopted by researchers in various fields for diverse purposes, including teaching, description, and NLP, aiding in understanding text and proposing a conceptual framework for coherence relations.

In RST, the minimum unit of analysis is *Elementary Discourse Unit* (EDU), which is like the concept of a clause in traditional Brazilian Portuguese (BP) grammars. These units consist of a *nucleus*, the most important part, and a *satellite*, a secondary part that supports the nucleus and is incomprehensible without it. These EDUs are connected through *rhetorical relations* (also known as *coherence* or *discourse relations*), forming a fully connected discourse structure typically represented as a tree. To determine which rhetorical relations apply between two or

more propositions, it is necessary to consider the four constraints of the nucleus, the satellite, the nucleus-satellite relationship, and the effect generated, as we exemplify in (1). Each constraint specifies judgments that the text analyst must take in building the RST structure.

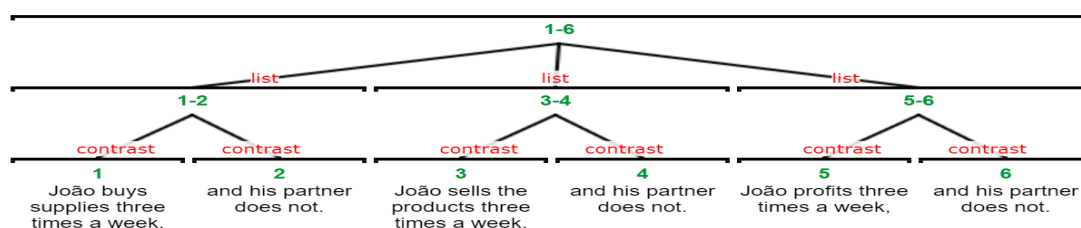
- (1) [João buys supplies three times a week,]<sup>A</sup> [and his partner does not.]<sup>B</sup>  
 [João sells the products three times a week,]<sup>C</sup> [and his partner does not.]<sup>D</sup>  
 [João profits three times a week,]<sup>E</sup> [and his partner does not.]<sup>F</sup>

In (1), there are six units of analysis that can relate in diverse ways. In one interpretation, between (1A) - (1B), (1C) - (1D), and (1E) - (1F), there are contrasts structured in a list format. In another interpretation, the contrastive relations remain, but the relationship between (1E) - (1F) could be a result concerning the other propositions, suggesting that João is profiting secretly from his partner. These interpretative possibilities show that propositions can be understood in diverse ways: in the first interpretation, (1C) - (1D) and (1E) - (1F) are nuclei; in the second, (1E) - (1F) becomes a satellite of (1C) - (1D), for instance.

RST proposes a set of rhetorical relations not fixed due to language differences. For English, there are 24 relations (Mann; Thompson, 1988), whereas for Portuguese, there are 32 (Pardo, 2005). While the theoretical model allows for some flexibility in proposing new relations, Taboada and Mann (2006) caution against increasing the number of classes excessively, as a higher number of classification possibilities makes manual text analysis more difficult.

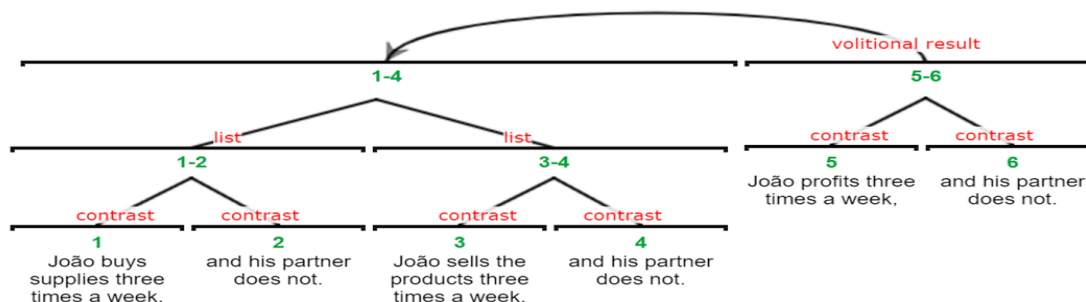
Therefore, in (1), depending on the relations chosen by the analyst, we could arrive at the representations depicted below, attempting to illustrate how discourse is rhetorically structured: sometimes interpreting the contrastive propositions as components of a list of events (as in Figure 1), other times interpreting the last contrastive proposition as the intentional result arising from the first two (as in Figure 2).

Figure 1 – First rhetorical interpretation of Example (1)



Source: The authors.

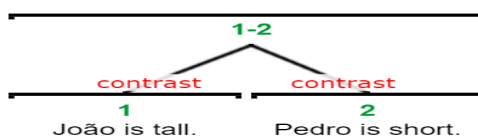
Figure 2 – Second rhetorical interpretation of Example (1)



Source: The authors.

Mostly, RST relations have been identified through Discourse Markers (DMs or connectives) that characterize rhetorical relations (such as “but” for Contrast and “if” for Condition), primarily belonging to the grammatical classes of prepositions and conjunctions. According to Marcu (2000), DMs play a fundamental role in segmenting texts into elementary units of discourse and signaling the rhetorical relationships that occur between segments of discourse. However, depending on the textual genre and linguistic modality comprising the *corpus* of analysis, the presence of DMs is not always guaranteed, despite the identification of the RST relation, as shown in Figure 3.

Figure 3 – RST relation without the presence of discourse markers



Source: Prepared by the authors based on Das and Taboada (2018).

In Figure 3, it is possible to analyze that there is a contrast between EDUs 1 and 2, despite the absence of DM. This relationship could be evidenced, for instance, between antonymous adjectives such as “tall” and “short”. Taboada and Das (2013), analyzing the RST Discourse Treebank (RST-DT) *corpus* (Carlson; Marcu; Okurowski, 2002), composed of English news texts, found that approximately 86% of RST relations were signaled. Within this sample, about 77% of RST relations were not indicated by DMs, but by other signals.

Based on these findings and the scarcity of research in this area for BP, we aimed to investigate additional clues in the *corpus* that could indicate rhetorical relations in the same sentence. We conducted a pilot study (Rodrigues; Souza; Cardoso, 2023) using a *subcorpus* of annotated news texts with the RST model. From this study, we identified several linguistic and structural signals that potentially indicate RST relations, and have not yet been cataloged in Portuguese. Building on this initial exploration, we progressed to the study and description of larger *corpora*, culminating in this paper with new contributions. Here, we aim to demonstrate not only our typological proposal for organizing and understanding these signals but also how these clues behave in RST relations.

We organized this paper into five sections, in addition to this Introduction. In Section 2, we revisit related studies, specifically focusing on *discourse signals* (DSs). In Section 3 we explain the methodology employed in this work. In Section 4, we present the constructed taxonomy, delineating each category based on a *corpus* already annotated for coherence relations. Finally, in Section 5 we present final remarks and establish future work.

## 2 Related Works

Identifying RST relations through explicit markers in the text is not new, especially in NLP, as a strategy to promote discourse analysis. DMs are the surface-level clues in the text, and they have been extensively studied in literature. This is because, in many cases, DMs are

considered indicators capable of identifying, articulating, and characterizing coherence relations between propositions (Cunha *et al.*, 2012; Hernault *et al.*, 2010; Marcu, 2000; Pardo; Nunes, 2008; Liu; Zeldes, 2019; Rodrigues; Souza; Cardoso, 2023; Cardoso *et al.*, 2024).

When considering the possibility of discourse analysis through DMs, it is necessary to acknowledge that the author has a communicative and semantic intention regarding the content, which can be interpreted by the reader and/or analyst through the structured mapping facilitated by the identification of these DSs. In this regard, Taboada and Das (2013) argue that successful communication relies on an unambiguous interpretation of relations, for which clear signals are necessary. The authors hypothesize that coherence relations are cognitive entities; thus, we can explore how listeners and readers can identify them based on linguistic and structural cues.

According to Liu and Zeldes (2019), theoretical models for annotation of signals are of great interest as they provide insights in both Linguistics and NLP. The authors suggest that from a psycholinguistic perspective, we could understand how readers recognize relationships to derive identical interpretations from the same text. From a computational standpoint, understanding DS can aid in feature engineering for automatic discourse parsing, text summarization, and automated text correction tools. Thus, it is observed that both linguistic and computational approaches have relied on theoretical models to understand how relationships are processed.

One study illustrating the application of DMs in NLP is Marcu's (2000) work on the English language. The author argues that DMs are ambiguous with respect to the rhetorical relations they signal and the importance of textual spans they relate. For example, the occurrence of the DM "but" at the beginning of a sentence most often signals either a mononuclear relation of Antithesis or Concession between a satellite, a textual span that precedes the occurrence of "but", and a nucleus, a textual span that starts with "but"; or a multinuclear relation of Contrast between two nuclei: a textual span that precedes the occurrence of "but" and a textual span that starts with "but". To build a discourse parser, the author associated each DM with the potential rhetorical relationships signaled. For instance, the connective "although", depending on its position within the sentence, can indicate a Concession or Contrast relationship between the propositions expressed by the segments in which the DM is observed. In cases where no DMs appear between textual segments, the author applies simple heuristics to infer the rhetorical relationship. If a segment repeats words from the previous segment with no DMs between them, a Background relationship is established; otherwise, an Elaboration relationship is inferred, which is the most common and generic relationship in the RST framework.

Regarding the Portuguese language, Pardo (2005) was a pioneer in investigating the development of discourse parsers. Through manual analysis of a *corpus* of scientific texts annotated with RST, the author concluded that DMs are essential for automatic discourse analysis, as they are the most prominent superficial indicators of rhetorical relationships in the text. The same relationship can be signaled by various connectives – for example, the Concession relationship can be signaled by DMs such as 'no entanto'/'however', 'porém'/'though', among others –, and a single DM can signal multiple relationships – for example, the connective 'porque'/'because' can signal relationships such as Cause, Result (volitional or not), Justify, Explanation, among others.

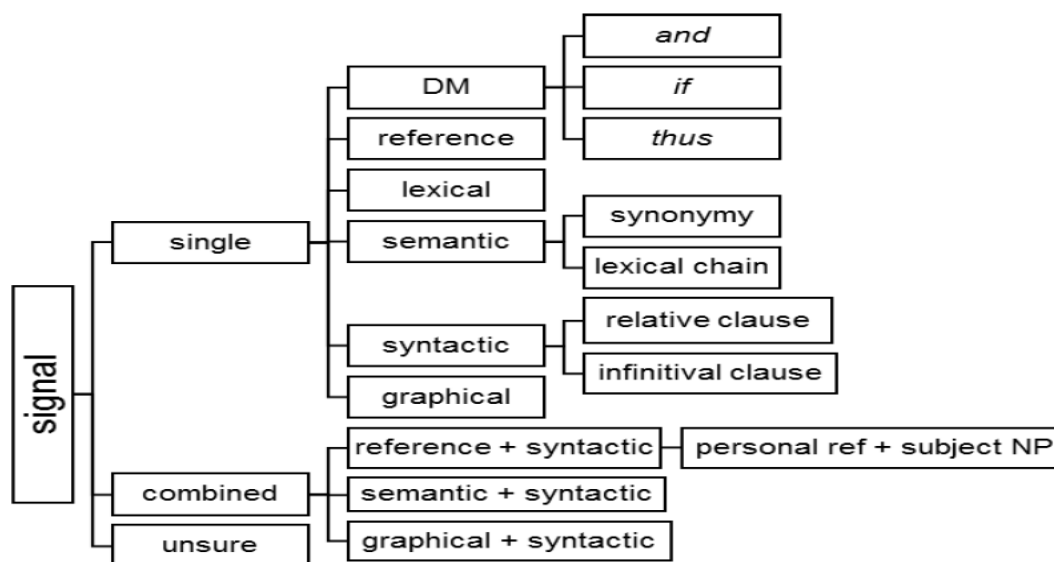
Further regarding Portuguese, Maziero (2016) investigated textual organization attributes, morphosyntax, syntax, semantics, and discourse to build a discourse parser based on RST. After several experiments with machine learning to identify discourse relations in the same sentence, the author concluded that attributes at the morphosyntactic level led to better results than the use of semantic and discourse-level attributes.

According to studies in linguistic *corpora*, especially those that encompass non-formal textual genres, there has been a recognized need to broaden the scope of understanding regarding the rhetorical structure of propositions in a text. The studies of Taboada and Das (2013), Das and Taboada (2018), and Liu and Zeldes (2019) report efforts to map all indicators establishing discourse relations in *corpora* annotated with RST.

Taboada and Das (2013) emphasize that understanding texts begins with constructing a representation of the information present in them, where part of this process involves reconstructing the possibilities of organizing propositions. The authors point out that identifying connectives indicating relations facilitates text processing, but the absence of a prototypical DM does not hinder its interpretation. Thus, they argue that coherence relations are cognitive entities, and therefore, text interpretation is possible. In this sense, the authors chose not to consider unsigned relations in their analyses.

Based on the RST Discourse Treebank, a *corpus* already annotated with rhetorical relations, Das and Taboada (2018) conducted a detailed annotation of DMs and signals. As a result, the RST Signalling *corpus* (RST-SC) was obtained. For this purpose, the authors considered formal elements beyond DMs, organizing the taxonomy as presented in Figure 4.

Figure 4 – Fragment of English taxonomy of discourse signals



Source: Das and Taboada (2018).

The signals are organized hierarchically in three major levels: signal class, signal type and specific signal. The top level, signal class, has three classes: single, combined and unsure. The second level, signal type, has nine types (DMs, reference, lexical, semantic, morphological, syntactic, graphical, genre and numerical features). In the third level, specific signal, each type further breaks down into specific signals: personal reference, demonstrative reference,

comparative reference and propositional reference. This hierarchical structure organizes signals into broader classes, further categorizing them into types, and finally specifying them into more granular specific signals. This taxonomy helps in systematically organizing and classifying different types of signals within its domain of study or analysis.

In Figure 4, the authors note that a particular relation can be annotated by a single signal (such as DMs, personal reference, relative clause, and colon) or a combined signal (comma + past participle clause; parallel syntactic construction + lexical chain, for example). For annotation, they performed a sequence of three tasks: 1) they examined each relation in the RST-DT *corpus*; 2) they assumed that the relational annotation was correct and searched for signals that indicated that such relation was present; and 3) they added to those relations a new layer of annotation for signaling information. Additionally, during the annotation process, the authors report cases annotated as uncertain, where it was not possible to establish with precision the signal that denotes a specific relation. When encountering a new instance of a specific type of relation, annotators consulted the taxonomy (Figure 4) and attempted to identify the appropriate signal(s) that could best serve as indicator(s) for that instance of relation. In such cases, the annotators followed the following protocol:

- (a) If the search allowed them to find an appropriate signal (or more than one appropriate signal) for that relation, then they succeeded in identifying the signals for that relation.
- (b) If the search did not match any of the signals in the taxonomy, they examined the context (including spans) to discover any potential new signals.
- (c) If a new signal was identified, it would be included in the appropriate category in the existing taxonomy.

The authors noticed that after approximately 50 files, or 2000 relations, they had added very few new signals to the taxonomy. In total, 385 documents were annotated, divided into 347 for training and 38 for testing, containing a total of 20,123 relations.

Based on this typological proposal of signals, Liu and Zeldes (2019) described the annotation of DSs at all levels of RST trees (from EDUs to the highest levels of the tree). The annotation was conducted on a sample of eleven documents extracted from the RST-SC *corpus* (Das; Taboada, 2018). Two annotators developed the annotation scheme as they annotated the documents and judged questionable cases. Thus, for each instance of a signal, the following categories were considered:

- (a) Type: RST-SC, e.g., syntactic, genre
- (b) Signal: subcategories of RST-SC types such as relative clause, verb tense, etc.
- (c) Anchor: token's correspondence to signal
- (d) Source/Destination: the related discourse units
- (e) Relname: the name relation being signaled
- (f) Discontinuous: a co-index for signals anchored in intervals of discontinuous tokens

The results showed that more than 92% of DMs could be anchored to specific tokens in the text, with the semantic type representing most cases (41.7% of the signaling anchors). Most of these cases correspond to the use of lexical chains and coreferential expressions, including pronominal anaphora. In contrast, discourse relations anchored by DMs account

for only about 8.5%. The most frequent types of signals identified were semantic, syntactic, and a combination of both semantic and syntactic signals.

In an analysis of transcribed oral texts in Portuguese, Antonio (2017) listed different cues that function as indicators of RST relations. According to the author, coherence relations can be interpreted even without the presence of a connective, as they are meaning relations. In his proposal, Antonio points out that punctuation marks, clause mood, phonological, morphosyntactic, semantic, and cognitive aspects can signal specific relations in the rhetorical structure.

Based on these proposals, we conducted a pilot study (Rodrigues; Souza; Cardoso, 2023) aiming to identify DMs in the CSTNews<sup>1</sup> *corpus* (Cardoso *et al.*, 2011) at the intra-sentential level. We performed a manual analysis, indicating any clues that could signal and/or justify the RST relation pre-annotated in the text. For this, we based our analysis on twenty-one documents from the *corpus*, covering diverse topics. As a result, we observed different signals occurring in various RST relations, such as DMs and syntactic information for the Attribution relation, for example. Additionally, we proposed an analysis that related the quantity of signal occurrences to the RST relation.

Until then, the study was conducted in an exploratory manner, so after annotation sessions, it was necessary to meet to align and decide on the label to use for a given signal (such as “communication verb” or “dicendi verb”; or “direct discourse” or “communication constructions”, for example). The ambiguity in labeling often leads to disagreement among annotators, hindering the human and automatic learning process of the task. Therefore, we decided to typologically organize the identified DS and advance the number of texts analyzed. Consequently, the signals were reorganized, resulting in (i) simplification or merging in certain categories, (ii) gain of subtypes in some categories, or (iii) detachment from the proposal, given the low occurrence or lack of understanding of their linguistic-structural behavior.

Thus, in the next section, we will present the methodology for developing the typological proposal of RST relation signals for Portuguese, the focus of this study.

### 3 Methodology

As previously mentioned, this taxonomy aims to expand upon the pilot study conducted by Rodrigues, Souza, and Cardoso (2023), in which the signals were identified from the CSTNews *corpus* (Cardoso *et al.*, 2011). Freitas (2024) defines a *corpus* as a collection of naturally produced linguistic data – i.e., data produced by humans – gathered for the purpose of exploring or investigating linguistic aspects. Furthermore, the author notes that for a dataset to qualify as a *corpus*, it must meet specific technical criteria related to *size*, *annotation type*, and *utility*.

Regarding CSTNews, it comprises news articles collected in 2007 from Brazilian newspapers in their online versions, including *Folha de S. Paulo*, *Estadão*, *Jornal do Brasil*, *Gazeta do Povo*, and *O Globo*. The *corpus* consists of 50 clusters representing different sections of the newspapers: 10 clusters of sports, 14 of world, 1 of money, 10 of politics, 1 of science, and 14 of daily life. In total, CSTNews contains 72,148 words distributed across 140 texts. In terms of utility, the *corpus* was originally designed for automatic summarization studies, supporting various

<sup>1</sup> CTS: Cross-document Structure Theory.

academic research endeavors in Descriptive Linguistics and NLP. Regarding annotation type, each *corpus* cluster includes (i) two or three news from different news sources, (ii) materials focused on summarization (such as human and automatic summaries), and (iii) various linguistic annotations, including morphosyntactic, semantic, and discursive annotations, the latter encompassing RST annotations.

Conforming to the methodology proposed by Taboada and Das (2013), it was necessary to start our study with a *corpus* already annotated with the RST model. This characteristic was already present in CSTNews, which is composed of the same textual genre as studied in English and has a similar distribution of rhetorical relations in the *corpus* for both languages.

Another crucial point highlighted by Taboada and Das (2013) was to remain “open-minded” about other signalers. The authors mention the Penn Discourse Treebank (Prasad; Webber; Joshi, 2017), which focused on DMs to indicate rhetorical relations. Thus, our first task was to examine the segments with RST relations already identified from CSTNews and attempt, as much as possible, to propose other discourse signals beyond DMs, without yet concerning ourselves with their linguistic nature (level of analysis, for example) or structural aspects (size and quantity, for instance).

It should be noted that we chose to analyze only the largest text in the collection for this study, based on the number of sentences and words. This decision was made because the two or three texts comprising each *corpus* set were on the same topic and presented a series of redundancies (Souza; Di Felippo; Pardo, 2013), despite varying sources. Therefore, the likelihood of the largest text containing all RST relations from the cluster is higher. Additionally, when conducting a quantitative analysis of the signalers, redundant data from the studied fifty texts in the *corpus* will not be represented.

The initial analysis in the previous studies (Rodrigues; Souza; Cardoso, 2023) was done manually without the aid of any computational tool. In that work, using .rs3 files (in Extensible Markup Language – XML format) from CSTNews, we opened them in a text editor and applied XML tags at the beginning and end in the identified signal. This process was slow, as there were occasional inaccuracies in placing opening and closing tags in the file, and it was challenging to visualize the RST relations and the propositions they related to. For this study, we utilized the rstWeb tool (Zeldes, 2016), which allowed us to annotate the signalers directly on the tree representations in the RST file.

In Figure 5, we loaded in the rstWeb tool a file annotated with an Attribution relation between the segments of the text. According to this annotation, “The spokesperson” serves as the *Source*, “stated” as a *Communication verb*, and “that” as a *Conjunction* linking these elements to the informational content. After the analyst saves the annotation changes, the rstWeb tool highlights the number of signals next to the RST relation and marks in the text those chosen during the process. As a result, the tool automatically adds to the .rs3 file the “signals” tag linking the chosen signals to the tokens identified by the analyst (Figure 6).

Figure 5 – Example of DS annotation using rstWeb

Source: The authors.

Figure 6 – Example of DS annotation in the .rs3 file

```
<signals>
  <signal source="1" type="SEMA"
    subtype="FONTE" tokens="1,2"/>
  <signal source="1" type="SEMA"
    subtype="VERB-COMUNI" tokens="3"/>
  <signal source="1" type="MD"
    subtype="CONJ" tokens="" />
</signals>
```

Source: The authors.

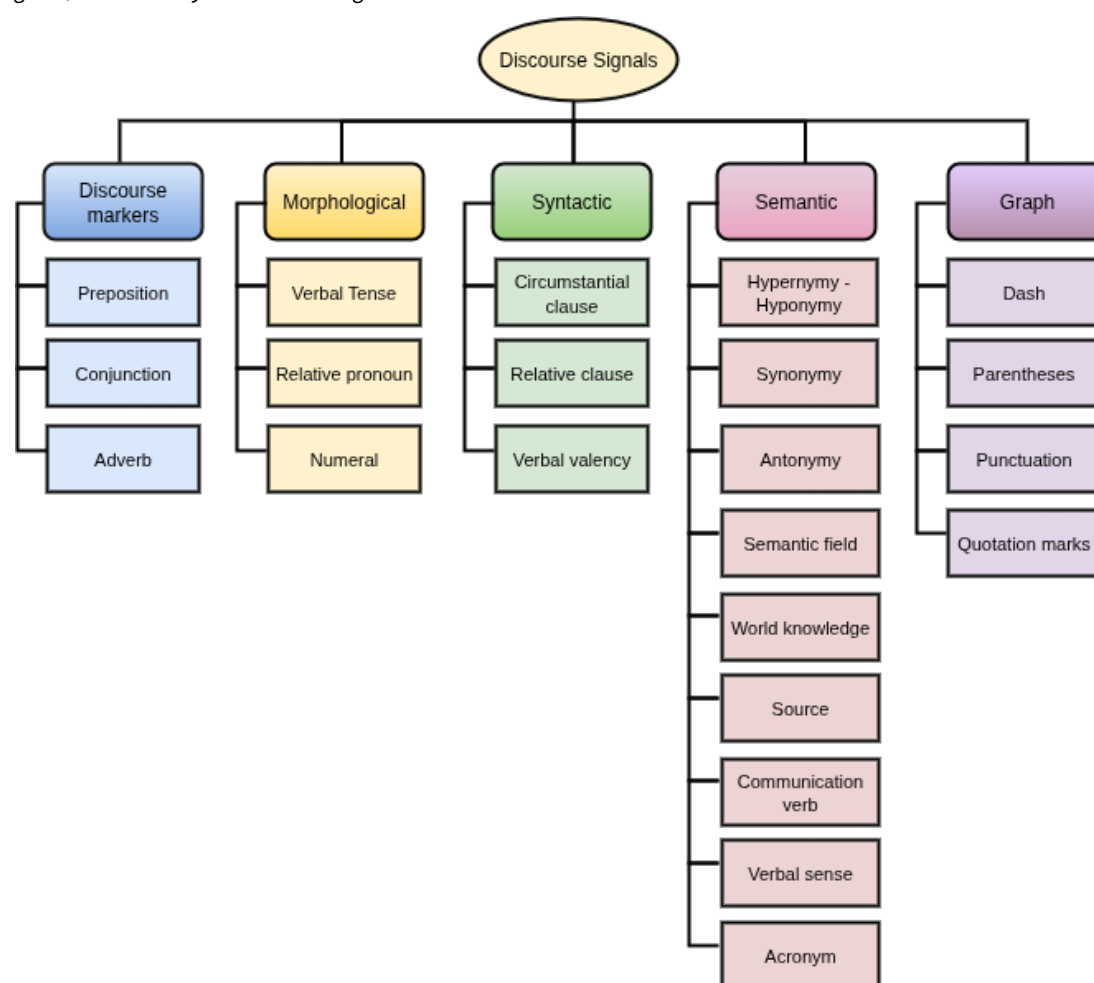
Based on these considerations, we present our taxonomy proposal in the next section.

## 4 Proposal for a Taxonomy of Discourse Signals

From the analysis of the CSTNews *corpus*, we identified a series of signals for RST relations in news texts produced in BP. Based on this, we developed a taxonomy organized into *types* and *subtypes* of signals, as shown in Figure 7. In this paper, we translate all the examples analyzed, with the original versions provided in footnotes.<sup>2</sup>

<sup>2</sup> It is important to emphasize that the examples presented in Section 4 are the same as those in our annotation manual (Dantas *et al.*, 2024). This occurred because the selected examples are highly prototypical of each case (of signal and rhetorical relation), and therefore, they satisfactorily illustrate the described phenomena. We also point out that the manual was developed based on this paper (which was still under evaluation by the journal at the time) and on the collaboration of other researcher-annotators, whose contributions are more thoroughly described in other works (Cardoso *et al.*, 2024; Cruz; Souza; Cardoso, 2024).

Figure 7 – Taxonomy of discourse signals



Source: Translated from Dantas *et al.* (2024).

In the following sections, we will present the main characteristics of each type: (i) Discourse markers type; (ii) Morphological type; (iii) Syntactic type; (iv) Semantic type; and (v) Graphic type. Then, in Section 4.6, we will describe the signals annotated in some rhetorical relations, the most abundant ones found in the *corpus*.

## 4.1 Discourse Markers Type

Based on Taboada and Das (2013), using Fraser's (2009) proposal, we understand that DMs: (i) signal a binary relation over a single discourse sequence (2a); (ii) can be positioned in different places of the sentence (at the beginning, middle, or end); (iii) do not create relations between paragraphs (2d), but only in the same sentence (2c); and (iv) do not imply an exclusive connection with a rhetorical relation; they only guide its interpretation. All examples in (2) came from CSTNews *corpus*.

- (2)
- (a) ✓ Lula fluctuates from 48% to 50%, **while** Alckmin drops from 39% to 36%.<sup>3</sup>
  - (b) ✓ **According** to airport sources, the crew members were of Russian nationality.<sup>4</sup>
  - (c) ✓ The shooting is one of the worst crimes of its kind on a university campus in the US **since** Charles Whitman opened fire.<sup>5</sup>
  - (d) ✗ Today's survey showed a variation in the spontaneous list [...]. **In this scenario**, Lula rises from 27% to 31% [...].<sup>6</sup>

In our taxonomy, we subcategorized DMs into 3 subtypes (Preposition, Conjunction and Adverb), as we can see in Table 1:

Table 1 – Examples of DMs as discourse signals

| SUBTYPE     | EXAMPLE                                                                                                                                                                 |
|-------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Preposition | <b>According to</b> airport sources, the crew members were of Russian nationality.<br>RST relation: Attribution                                                         |
| Conjunction | [...] there was a 110 km traffic jam throughout the city <b>while</b> the average for that time was 76 km. <sup>7</sup><br>RST relation: Comparison                     |
| Adverb      | <b>So far</b> , authorities in Sri Lanka have not confirmed the deaths or clarified what is happening in the city of Muttur. <sup>8</sup><br>RST relation: Circumstance |

Source: Translated from Dantas *et al.* (2024).

## 4.2 Morphological Type

The signals in this category contribute to the interpretation of rhetorical relations by indicating temporal information, through verbal inflection, or by belonging to a specific grammatical class, namely: *Relative pronoun* or *Numeral*. In Table 2, we illustrate each case:

<sup>3</sup> “Lula oscila de 48% para 50%, **enquanto** Alckmin cai de 39% para 36%.”

<sup>4</sup> “**Segundo** fontes aeroportuárias, os membros da tripulação eram de nacionalidade russa.”

<sup>5</sup> “O tiroteio é um dos piores crimes do tipo no campus de uma universidade nos EUA **desde que** Charles Whitman abriu fogo.”

<sup>6</sup> “A pesquisa de hoje apresentou uma variação na lista espontânea [...]. **Neste cenário**, Lula sobe de 27% para 31% [...].”

<sup>7</sup> “[...] havia 110 km de congestionamento em toda a cidade **enquanto** a média para o horário era de 76 km.”

<sup>8</sup> “**Até o momento**, as autoridades do Sri Lanka não confirmaram as mortes ou esclareceram o que acontece na cidade de Muttur.”

Table 2 – Examples of morphological discourse signals

| SUBTYPE          | EXAMPLE                                                                                                                                                               |
|------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Verbal tense     | On Friday, it <b>rained</b> 12 centimeters in some regions, and <b>there is</b> a forecast of more storms today. <sup>9</sup><br>RST relation: Sequence               |
| Relative pronoun | The temperature is expected to remain low due to the polar air mass accompanying the cold front <b>through*</b> the state. <sup>10</sup><br>RST relation: Elaboration |
| Numeral          | The <b>first</b> one started at 10a.m., and the <b>second</b> one is scheduled for 4p.m. <sup>11</sup><br>RST relation: List                                          |

Source: Translated from Dantas *et al.* (2024).

Portuguese is a morphologically rich language, presenting different morphemes (derivational and inflectional) in its composition; therefore, the subtype *Verb tense* is quite relevant. In Table 2, the verbal inflection helps in interpreting the Sequence relation, where one event (“rained”) occurs before another (“there is”).

In addition, in the Morphology Type, we also included word classes that proved important in the interpretation of certain relations, more specifically *pronouns* and *numerals*. There are, obviously, several types of pronouns in Portuguese; however, only the *Relative pronoun* is included in the taxonomy because we observed its productivity in interpreting Elaboration and Explanation relations. Finally, the *Numeral* subtype tends to indicate the order of information, as we can see in the List relation illustrated by the ordinal numbers (“first”; “second”).

### 4.3 Syntactic Type

In the Syntactic type, we included constructions and uses that aid in the interpretation of rhetorical relations. From the annotation, we observed that the following syntactic constructions are relevant and abundant in the corpus, namely: *Circumstantial clause*, *Relative clause*, and *Verbal valency*.

<sup>9</sup> “Na sexta-feira choveu 12 centímetros em algumas regiões, e há previsão de mais tempestades hoje.”

<sup>10</sup> “A temperatura deve permanecer baixa, por conta da massa de ar polar que acompanha a frente fria que passa pelo estado.”

<sup>11</sup> “A primeira iniciou às 10h, e a segunda está marcada para às 16h.”

Table 3 – Examples of syntactic discourse signals

| SUBTYPE               | EXAMPLE                                                                                                                                                                       |
|-----------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Circumstantial clause | [...] <b>since he was released</b> , Abadia has undergone four plastic surgeries [...] <sup>12</sup><br>RST relation: Sequence                                                |
| Relative clause       | The temperature is expected to remain low due to the polar air mass accompanying the cold front <b>passing through the state</b> . <sup>13</sup><br>RST relation: Elaboration |
| Verbal valency        | <b>River Thames</b> , which is at its limit, <b>it may overflow</b> during the next early morning. <sup>14</sup><br>RST relation: Same unit                                   |

Source: Translated from Dantas *et al.* (2024).

The *Circumstantial clause* signal indicates not only temporal information, as seen in Table 3, but also other satellite information in relation to a nucleus, such as *manner*, *place*, *condition*, etc. We opted for this type of generalization to facilitate the understanding of the phenomenon and the execution of human annotation. The *Relative clause* serves to specify or explain a core element. This subtype is the main characteristic of the Elaboration relation, as it provides details about the situation or element present in the nucleus of the EDU.

On the other hand, *Verbal valency* contributes to understanding the scope of the verbal core (its subject and complements) and, consequently, the relationships established between the elements of the sentence. In the example from Table 3, the subject (“River Thames”) is repeated in English by using a pronoun (“it may...”); however, in Portuguese, subject omission is quite common (“Ø pode transbordar...”), making the reference occur through verbal agreement.

#### 4.4 Semantic Type

This is the category with the most subtypes of signals (a total of 9), whose focus is on the meaning established between the elements of the sentence.

Prototypical semantic relations seem to be quite important in identifying different rhetorical relations, such as the concepts of: *Hypernymy* (when a word is more general – e.g., “river”) and *Hyponymy* (when a word is more specific – e.g., “River Severn”); *Synonymy* (words with similar meanings – e.g., “triumph”/ “victories”); *Antonymy* (words with opposite meanings – e.g., “dropped”/ “increased”); and *Semantic field* (words with related meanings that share a common theme – e.g., “rebellion”/ “prison”). Additionally, we have included *World Knowledge* in this category. Although it is not exactly a semantic concept, we understand that World knowledge accesses information external to the text to contribute to its interpretation, mobilizing the meaning attributed to the words in the text.

<sup>12</sup> “...desde que foi solto, Abadia fez quatro cirurgias plásticas [...]”

<sup>13</sup> “A temperatura deve permanecer baixa, por conta da massa de ar polar que acompanha a frente fria que passa pelo estado.”

<sup>14</sup> “Já o rio Tâmbisa, que está com seu leito no limite, pode transbordar durante a próxima madrugada.”

Table 4 – Examples of semantic discourse signals

| SUBTYPE                | EXAMPLE                                                                                                                                                                                                                                   |
|------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Hypernymy and hyponymy | The melting of the snow also affects the rising levels of the <b>rivers</b> . <b>River Severn</b> , the longest in the country, is five meters above the normal summer level. <sup>15</sup><br>RST relation: Explanation                  |
| Synonymy               | This was the Brazilians' seventh consecutive <b>triumph</b> in the competition — previously, the country achieved four <b>victories</b> against the Argentine team and two against Portugal. <sup>16</sup><br>RST relation: Parenthetical |
| Antonymy               | In the second round, President Lula's voting intentions <b>dropped</b> from 53% in June to 50% in July, while candidate Alckmin <b>increased</b> from 29% to 36%. <sup>17</sup><br>RST relation: Contrast                                 |
| Semantic field         | At the beginning of the <b>rebellion</b> , four people were injured, including a nurse assistant and a police officer who work at the <b>prison</b> . <sup>18</sup><br>RST relation: Elaboration                                          |
| World knowledge        | A total of 549 people died, 3043 were injured, and 295 are missing due to the <b>floods</b> . <sup>19</sup><br>RST relation: Non-volitional cause                                                                                         |
| Source                 | "It's an admirable pair of siblings, each with about 1% of the mass of the Sun," said <b>Jayawardhana</b> . <sup>20</sup><br>RST relation: Attribution                                                                                    |
| Communication verb     | "It's an admirable pair of siblings, each with about 1% of the mass of the Sun," <b>said Jayawardhana</b> .<br>RST relation: Attribution                                                                                                  |
| Verbal sense           | The rain <b>started</b> on Sunday night and intensified between 6:00 AM and 7:30 AM on Monday. <sup>21</sup><br>RST relation: Sequence                                                                                                    |
| Acronym                | The information collected by the Federal Police during the investigations was sent to the <b>CJ (Court of Justice)</b> of Rondônia and to the <b>SCJ (Superior Court of Justice)</b> . <sup>22</sup><br>RST relation: Parenthetical       |

Source: Translated from Dantas *et al.* (2024).

The signals *Source* (that marks the authorship of the reported content – e.g., “Jayawardhana”), *Communicative verb* (verbal construction that determines how informational content is reported – e.g., “said”), and *Verbal sense* (e.g., “started”) relate to semantic roles that

<sup>15</sup> “O degelo da neve também influi no aumento do nível dos rios. O rio Severn, o maior do país, está cinco metros acima do nível normal de verão.”

<sup>16</sup> “Este foi o sétimo **triunfo** consecutivo dos brasileiros na competição — antes, o país conquistou quatro **vitórias** contra a seleção argentina e duas diante de Portugal.”

<sup>17</sup> “No segundo turno, as intenções de voto do presidente Lula **caíram** de 53% em junho para 50% em julho, enquanto o candidato Alckmin **subiu** de 29% para 36%.”

<sup>18</sup> “No começo da **rebelião** quatro pessoas ficaram feridas, entre elas uma auxiliar de enfermagem e um agente de polícia que trabalham no **presídio**.”

<sup>19</sup> “Um total de 549 pessoas morreram, 3043 ficaram feridas e 295 estão desaparecidas em razão das **enchentes**.”

<sup>20</sup> “É um par de irmãos admirável, cada um com cerca de 1% da massa do Sol”, disse **Jayawardhana**.”

<sup>21</sup> “A chuva **começou** na noite de domingo e ficou mais forte entre 6h e 7h30 desta segunda.”

<sup>22</sup> “As informações coletadas pela PF durante as investigações foram enviadas ao **TJ (Tribunal de Justiça)** do Estado de Rondônia e ao **STJ (Superior Tribunal de Justiça)**.”

detail the meaning of grammatical elements that compose the sentence. We know that *Communicative verb* is a subtype of *Verbal sense*; however, it was considered a specific signal due to its high recurrence in the corpus, particularly in the Attribution relation.

Finally, in this work, *Acronym* also encompasses cases of Abbreviations and Initialisms, referring to the phenomenon of forming a new word by combining initials or parts of a term (e.g., “CJ – Court of Justice”). Acronyms are typically studied in Morphology; however, we associate them with Semantic type because their annotation contributes to the interpretation of rhetorical relations when the text provides the specification (or explanation) of the abbreviated term, which usually appears in parentheses.

### 4.5 Graph Type

As seen in Table 5, there are 4 subtypes of Graph signals that contribute to the identification of rhetorical relations. In general, these signals are used to indicate additional, complementary, or explanatory information to the text, as illustrated in the examples using *Dashes*, *Parentheses*, and *Punctuation*. Although less productive, graph signals can serve other functions that aid in discursive interpretation, as seen in the example with *Quotation marks*, which indicate that the fragment refers to a reported direct speech.

Table 5 – Examples of graph discourse signals

| SUBTYPE         | EXAMPLE                                                                                                                                                                                    |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Dash            | The failure of the thrust reverser – a mechanism that helps the airplane to brake – [...] <sup>23</sup><br>RST relation: Parenthetical                                                     |
| Parentheses     | [...] because this is the first CNI/Ibope poll with the official list of candidates from the SEC ( Superior Electoral Court ) <sup>24</sup><br>RST relation: Parenthetical                 |
| Punctuation     | One of the three positions will be filled by Major-Brigadier Allemander Jesus Pereira Filho, nominated to replace Jorge Luiz Brito Velozo [...] <sup>25</sup><br>RST relation: Explanation |
| Quotation marks | “It’s an admirable pair of siblings, each with about 1% of the mass of the Sun,” said Jayawardhana.<br>RST relation: Attribution                                                           |

Source: Translated from Dantas *et al.* (2024).

This proposed taxonomy is based on the first round of annotation of discourse signals in the BP news *corpus*. Considering the available data, we will present, in the next subsection, the signals found for the 13 most abundant rhetorical relations in the *corpus*.

<sup>23</sup> “A falha no reversor – mecanismo que ajuda o avião a frear – [...]”.

<sup>24</sup> “...porque esta é a primeira CNI/Ibope com a lista oficial dos candidatos do TSE (Tribunal Superior Eleitoral)”.

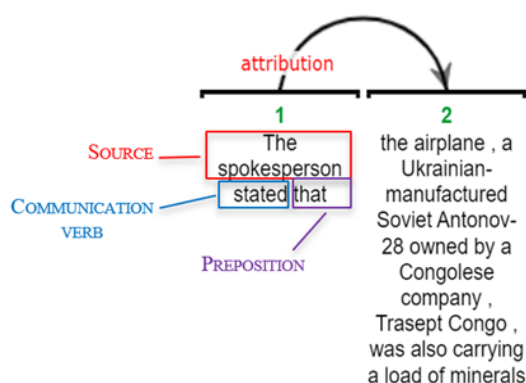
<sup>25</sup> “Uma das três vagas será ocupada pelo major-brigadeiro Allemander Jesus Pereira Filho, indicado para exercer o cargo em substituição a Jorge Luiz Brito Velozo...”

## 4.6 Analysis of Some RST Relations Based on the Signals

In this subsection, we present 13 rhetorical relations annotated with our proposed taxonomy.

- (a) Attribution: presents a statement and its source (author), in a direct or indirect discourse in the text. This relation is one of the most frequent occurrences in our news *corpus* and is indicated through the following signals: *Source*, *Communication verb*, and *Conjunction* (Figure 8).

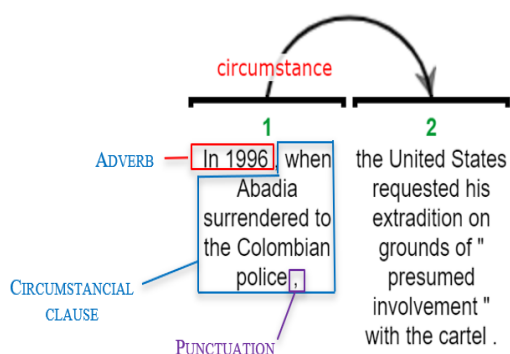
Figure 8 – Example of Attribution relation on rhetorical tree<sup>26</sup>



Source: Translated from Dantas *et al.* (2024).

- (b) Circumstance: has a satellite (EDU 1, in Figure 9) that contextualizes the situation presented in the nucleus (EDU 2, in Figure 9). In our annotation, this relation is signaled by: *Adverb*, *Circumstantial clause*, and *Punctuation*.

Figure 9 – Example of Circumstance relation on rhetorical tree<sup>27</sup>

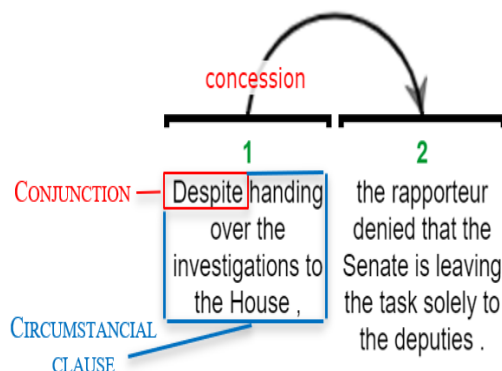


Source: Translated from Dantas *et al.* (2024).

<sup>26</sup> "O porta-voz informou que o avião, um Soviet Antonov-28 de fabricação ucraniana e propriedade de uma companhia congoleza, a Transept Congo, também levava uma carga de minerais."

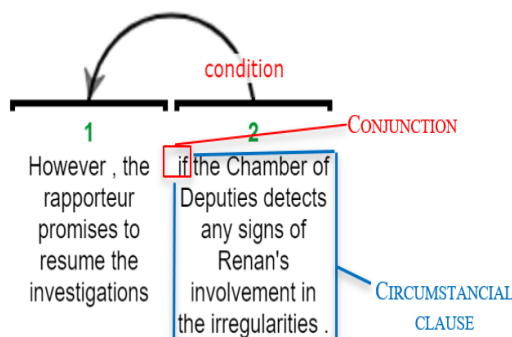
<sup>27</sup> "Em 1996, quando Abadia se entregou à polícia colombiana, os Estados Unidos pediram a extradição dele pelo 'presumível envolvimento' com o cartel."

- (c) Concession: highlights a concessive proposition, that is, the satellite presents an idea contrasting with the nucleus, but without undermining its validity. In the following example, *Conjunction* and *Circumstantial clauses* were identified as signals.

Figure 10 – Example of Concession relation on rhetorical tree<sup>28</sup>

Source: Translated from Dantas *et al.* (2024).

- (d) Condition: has a satellite (EDU 2, in Figure 11) that presents a condition for the situation described in the nucleus (EDU 1, in Figure 11) to be realized. In the following example, we identified *Conjunction* and *Circumstantial clauses*.

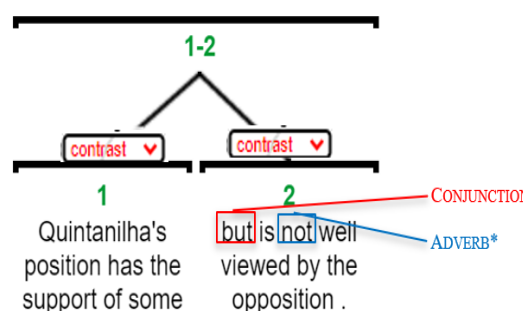
Figure 11 – Example of Condition relation on rhetorical tree<sup>29</sup>

Source: Translated from Dantas *et al.* (2024).

- (e) Contrast: is a multinuclear and a binary relation, meaning it always occurs between two nuclei (such as EDUs 1 and 2 in Figure 12), which can be understood in some respects and compared regarding their differences. In Figure 12, we identified two signals of the DM type, namely *Conjunction* and *Adverb*.

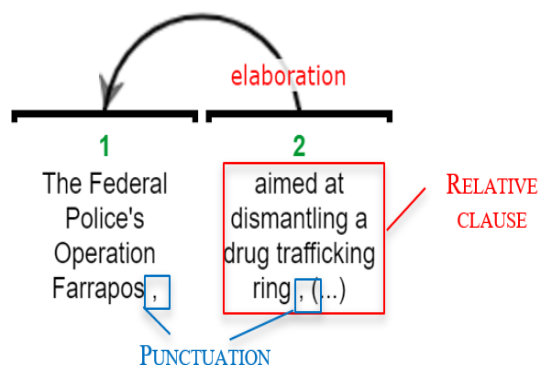
<sup>28</sup> “Apesar de deixar as investigações para a Câmara, o relator negou que o Senado esteja deixando a tarefa apenas para os deputados.”

<sup>29</sup> “O relator promete, no entanto, retomar as investigações se a Câmara perceber indícios do envolvimento de Renan nas irregularidades.”

Figure 12 – Example of Contrast relation on rhetorical tree<sup>30</sup>

Source: Translated from Dantas *et al.* (2024).

- (f) Elaboration: is an abundant relation in the CSTNews *corpus*. In this relation, the satellite provides an explanation or additional detail to what is presented in the nucleus. In the *corpus*, the Elaboration is indicated through the following signals: *Punctuation*, *Relative clause*, and *Relative pronoun*.

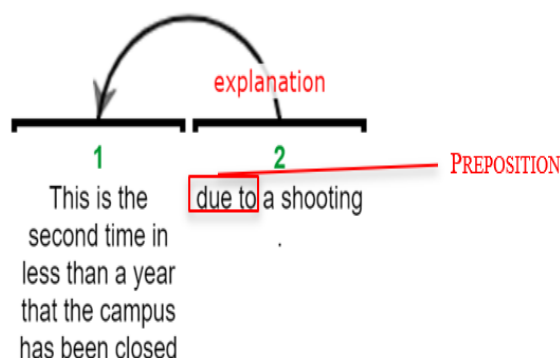
Figure 13 – Example of Elaboration relation on rhetorical tree<sup>31</sup>

Source: Translated from Dantas *et al.* (2024).

- (g) Explanation: focuses on the explanation or justification in the satellite (such as EDU 2, Figure 14) for the idea presented in the nucleus (such as EDU 1, Figure 14). In the analyzed example, the *Preposition* DM signals this relation.

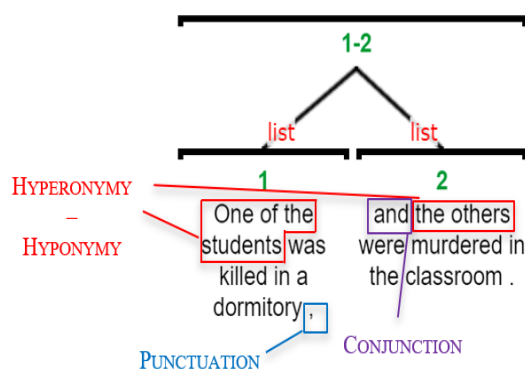
<sup>30</sup> “A posição de Quintanilha tem o apoio de alguns senadores, mas não é bem-vista pela oposição.”

<sup>31</sup> “A Operação Farrapos da Polícia Federal, que tem como objetivo desarticular uma quadrilha internacional de drogas, [...]”

Figure 14 – Example of Explanation relation on rhetorical tree<sup>32</sup>

Source: Translated from Dantas *et al.* (2024).

- (h) List: is a multinuclear relation, where the nucleus (such as in EDUs 1 and 2, Figure 15) present items that are comparable, with no hierarchy between them. In our *corpus*, we identified the following signals: *Hypernymy-Hyponymy*, *Punctuation*, and *Conjunction*.

Figure 15 – Example of List relation on rhetorical tree<sup>33</sup>

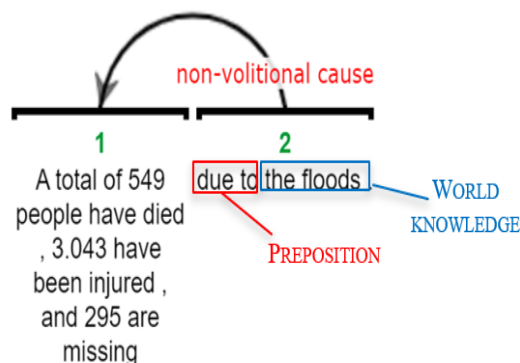
Source: Translated from Dantas *et al.* (2024).

- (i) Non-volitional cause: is characterized by presenting the unintentional cause in the satellite (such as in EDU 2, Figure 16) in relation to the main situation presented in the nucleus (such as in EDU 1, Figure 16). In the analyzed example, we identified the signals *Preposition* and *World knowledge* for this relation.

<sup>32</sup> “Esta é a segunda vez em menos de um ano que o campus é fechado devido a um tiroteio.”

<sup>33</sup> “Um dos estudantes foi morto em um dormitório, e os outros foram assassinados na sala de aula.”

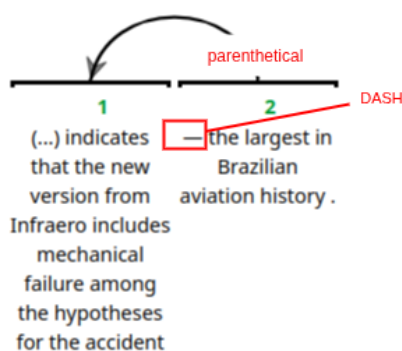
Figure 16 – Example of Non-volitional cause relation on rhetorical tree<sup>34</sup>



Source: Translated from Dantas *et al.* (2024).

- (j) Parenthetical: its satellite (such as in EDU 2, Figure 17) is responsible for presenting additional information related to the nucleus (such as in EDU 1, Figure 17), always having some graphical delimiter for it. In this sense, we identified *Dashes* as a signal in the analyzed example below.

Figure 17 – Example of Parenthetical relation on rhetorical tree<sup>35</sup>

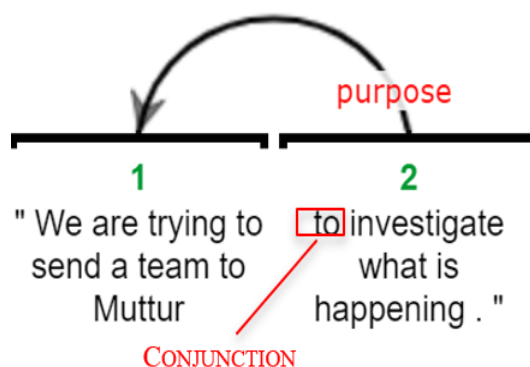


Source: Translated from Dantas *et al.* (2024).

- (k) Purpose: has a satellite (such as in EDU 2, Figure 18) that can realize the action presented in the nucleus (such as in EDU 1, Figure 18). In the study *corpus*, we identified the *Conjunction* as a signal.

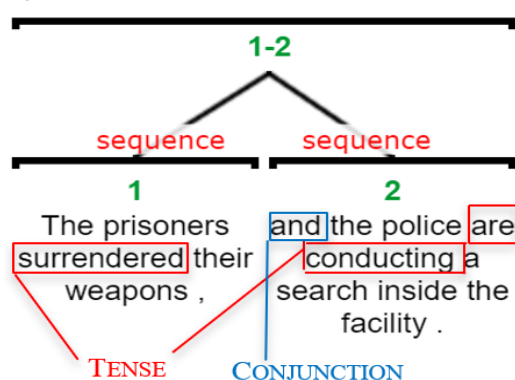
<sup>34</sup> “Um total de 549 pessoas morreram, 3043 ficaram feridas e 295 estão desaparecidas em razão das enchentes.”

<sup>35</sup> “[...] mostra que a nova versão da Infraero inclui falha mecânica entre as hipóteses para o acidente — maior da aviação brasileira.”

Figure 18 – Example of Purpose relation on rhetorical tree<sup>36</sup>

Source: Translated from Dantas *et al.* (2024).

- (l) Sequence: is a multinuclear relation that establishes a chronological connection between the events in the nucleus. *Verb tense* and *Conjunction* were identified as signals for this relation.

Figure 19 – Example of Sequence relation on rhetorical tree<sup>37</sup>

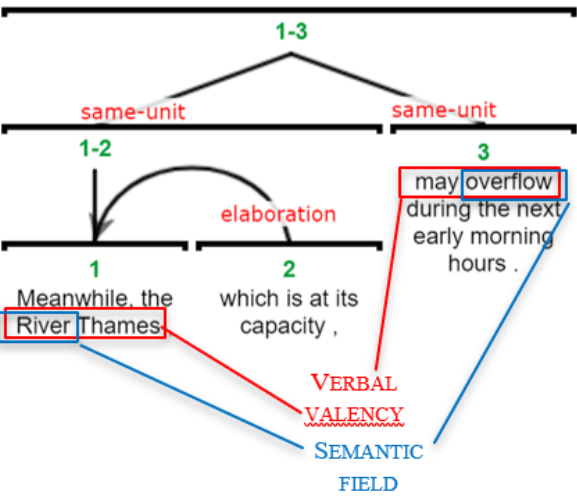
Source: Translated from Dantas *et al.* (2024).

- (m) Same-unit: is also a multinuclear relation, where its nucleus (EDUs 1 and 3, in Figure 20) constitute the same proposition that is segmented, most of the time, by another RST relation (such as EDU 2, in Figure 20). *Semantic field* and *Verbal valency* were signals that we have annotated.

<sup>36</sup> “Tentamos enviar uma equipe a Muttur para averiguar o que está acontecendo.”

<sup>37</sup> “Os presos entregaram as armas e a polícia faz uma revista dentro da unidade.”

Figure 20 – Example of Same-unit relation on rhetorical tree<sup>38</sup>



Source: Translated from Dantas *et al.* (2024).

In Table 6, we provide a summary of the DSs associated with each relationship identified in the study *corpus*.

Table 6 – Recurring signals by rhetorical relation

| Relation             | Discourse signals                              |
|----------------------|------------------------------------------------|
| Attribution          | Source, Communication verb, Conjunction        |
| Circumstance         | Adverb, Circumstantial clause, Punctuation     |
| Concession           | Conjunction, Circumstantial clauses            |
| Condition            | Conjunction, Circumstantial clauses            |
| Contrast             | Conjunction, Adverb                            |
| Elaboration          | Punctuation, Relative clause, Relative pronoun |
| Explanation          | Preposition                                    |
| List                 | Hypernymy-Hyponymy, Punctuation, Conjunction   |
| Non-volitional cause | Preposition, World knowledge                   |
| Parenthetical        | Parentheses, Dashes                            |
| Purpose              | Conjunction                                    |
| Sequence             | Conjunction, Verb tense                        |
| Same-unit            | Semantic field, Verbal valency                 |

Source: The authors.

As seen in Table 6, most relations are identified by two or more signals, except for Explanation, Parenthetical, and Purpose relations, which have unique signals. These cases need to be revisited with a more detailed annotation, as they seem too broad and, therefore, have gaps: Explanation and Purpose, for example, even share the same type of signal (a *Syntactic* one). In the combined signals, some relations appear to be more clearly identified,

<sup>38</sup> “Já o rio Tâmisá, que está com seu leito no limite, pode transbordar durante a próxima madrugada.”

such as Attribution, marked by both *Semantic* and *Syntactic* type signals, or Elaboration, marked by *Graph*, *Morphological* and *Syntactic* type signals. On the other hand, we highlight the Concession and Condition relations, which have clearly distinct interpretations but were annotated with identical combined signals; this poses a problem for both human and automatic annotation.

In the next subsection we will discuss in more detail some relevant aspects of the proposed taxonomy.

## 4.7 Questions and Lessons Learned

During the study and description of the CSTNews *corpus*, we encountered a series of questions. Some of them have already been answered, providing us with valuable lessons learned; however, others persist, motivating us to continue our research.

The first challenge we encountered relates to the methodology adopted in this work. As seen, we replicated the methodology of Taboada and Das (2013), which led us to analyze new DMs based on a *corpus* pre-annotated with the RST model. It is important to note that this decision had both limitations and advantages.

Since the set of texts was already annotated, there was no need for us to annotate the texts with the RST model ourselves, saving us time in the research process and allowing us to focus on studying the correlation between signals and the theoretical model's relations. When annotating relations using the RST model, annotators must interpret and construct a rhetorical relation of propositions based on clues left by the author in the text. However, when we consider an analysis that has already been conducted, we need to some extent imagine what the analyst interpreted from the text, and we must search within the propositions for signals that motivated them to assign a particular relation.

Despite this challenge, the dual interpretation process does not invalidate our study, given the high agreement among the annotators of the CSTNews *corpus* (Cardoso *et al.*, 2011). This indicates that all annotators understand how the coherence relations of the theoretical model behave and are structurally organized in discourse. However, the analysis could potentially be enriched if both the identification of signals and RST relations were conducted concurrently.

The types of signals proposed in this study traverse levels of linguistic analysis and structural aspects. They were proposed based on subtypes: for instance, *Circumstantial clauses* and *Relative clauses* were identified as signals for certain relations and thus organized under the *Syntactic* type. In this sense, there are both similarities and differences from the typological proposal for English.

As observed, Taboada and Das (2013) and Liu and Zeldes (2019) organize some signals into types only, such as Graphical, Numerical, Genre, and DMs; while for other types they present a few subtypes, such as Antonyms and Lexical-chain under the Semantic type. Moreover, there are other categories that we did not propose in our study, such as Entity and Lexical. These differences can be explained by the particularities of the theoretical model itself and variations in textual genres across Portuguese and English.

Furthermore, the taxonomy presented here will be used by a group of annotators with different academic backgrounds (either in Linguistics or Computer Science) and varying

levels of experience in annotating linguistic *corpora*. So, we decided to simplify some signals; for instance, instead of proposing a distinction between “Conditional clause” and “Temporal clause”, we merged them into *Circumstantial clause*. Regarding annotators’ profiles, we organized the signals so that each belongs to a type, sharing some common functional traits, thereby facilitating manual learning and identification by the annotators.

Another point worth reflecting upon is that we worked with only one genre of text, characterized by formal linguistic registers. Therefore, it is highly likely that we will encounter other types and subtypes of signals when analyzing other *corpora*.

Although this study is pioneering in the Portuguese language,<sup>39</sup> it is important to acknowledge its initial limitation regarding the textual genre analyzed. While other academic studies have advanced toward a multigenre description, our study focused exclusively on news texts. As presented, the literature reviewed is based on pre-annotated texts using RST to identify potential DMs of this theoretical formalism. Given the absence of a multigenre *corpus* annotated with RST, the decision to work solely with news texts stems from the limited availability of pre-annotated Portuguese *corpora* with RST relations, such as the CSTNews *corpus*.

It is also important to note that in some RST relations, we were unable to identify signals other than prototypical DMs, such as in Condition and Concession. In these cases, the identified signals seem to function merely as connectors between satellite and nucleus EDUs. However, as highlighted by Pardo (2005), there are RST relations of Intentional, Semantic, and Structural types. Therefore, it is necessary to continue refining linguistic descriptions to advance and enhance our taxonomy to capture intentionality and other semantic aspects.

Finally, as pointed out by Das and Taboada (2018), RST relationships involve a cognitive component, which means there may not always be a need for a DM, as we discussed previously. This implies that, in many cases, it is possible to recognize the semantic relationship between two or more propositions and understand the intended rhetorical arrangement in the text without a DM. However, it is not always possible to identify explicit DMs in the text that indicate such relationships or arrangements, as in the example below:

- (3) [In South Korea, the leader of the opposition Grand National Party, Kang Jae-sup, called for the establishment of an investigative mission]<sup>A</sup> [to be sent to the communist country.]<sup>B</sup>

According to the annotation in the CSTNews *corpus*, the RST relation between EDUs (3A) and (3B) is Elaboration. It is evident that the two propositions are related due to the *World knowledge* used to associate “South Korea” with “communist country” by a text coreferential. However, if the scope of the Elaboration is the event presented in EDU (3B) (namely, “to be sent”), World knowledge may not be a good signal in this scenario.

<sup>39</sup> We highlight that there are previous works, such as Juliano (2017), that focus on the description of DM and DS for BP. However, we did not find researchers whose focus is on news texts and/or the proposal of a robust taxonomy of signals, as we have done in this work.

## 5 Final Remarks

Throughout this study, we aimed to highlight discourse signals that go beyond DMs in indicating coherence relations within the RST. We sought to distance ourselves from the idea that a specific signal defines a given relation, a notion closely tied to DMs. Our goal in this study was to identify a series of clues that can serve as signals for RST relations, and how these signals are organized into specific categories based on their discursive functionality.

There are still questions that motivate us to continue investigating discourse signals in future work. As demonstrated, there are some RST relations that continue to be signaled only by DMs (such as the Purpose relation), even though the focus of these relations is on the content between propositions. Therefore, it may be necessary to consider and propose signals that can capture the rhetorical structure of discourse, such as understanding which RST relations precede and follow a Purpose, for example. That is, we believe that future studies need to be conducted to analyze and refine the taxonomy proposed here.

Furthermore, aiming to automatically process, it will be necessary to investigate how individual or combined signals can indicate RST relations. Attribution relation, for instance, may involve *Conjunction*, *Source*, and *Verbal communication* in some cases, while in others, it may involve only the latter two highlighted signals. Consequently, investigating which signals can be considered elemental for a given RST relation will enable the development of automated solutions for discourse analysis in NLP.

This work is relevant to linguistic descriptive studies regarding the determination of how readers and listeners use linguistic signals to identify rhetorical relations, particularly those that are not marked by DMs. Furthermore, the proposed description could contribute to the development and/or improvement of automatic discourse analyzers, especially those classified based on (i) the use of lexical patterns, obtained manually, and (ii) the use of machine learning (supervised or semi-supervised), with the enhancement of identifying rhetorical relations between portions of a text.

Lastly, international literature highlights other signals that have not yet been studied and, consequently, do not appear in our proposed taxonomy, such as Genre. We understand that characterizing certain signals based on textual genre, for instance, will require investigating other texts. Therefore, in future work, building upon our taxonomy, we will investigate other *corpora* with different textual compositions to examine the correlation between signals and textual genres. Thus, it underscores the need to consider the dynamism and flexibility of these taxonomies.

## Funding

This work was carried out at the Center for Artificial Intelligence of the University of São Paulo (C4AI – <http://c4ai.inova.usp.br/>), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. The project was also supported by the Ministry of Science, Technology and Innovation, with resources of Law N. 8,248, of October 23, 1991, within the scope of PPI-SOFTEX, coordinated by Softex and published as Residence in

TIC 13, DOU 01245.010222/2022-44. Furthermore, it was supported by financial resources from the PRPPG-UFBA 010/2024 Call for Proposals – Support Program for Early-Career Professors/ Researchers 2024 – and the FAPESB/CNPq 004/2023 Call for Proposals – First Projects Program.

## References

ANTONIO, J. D. Mecanismos utilizados pelos destinatários do discurso para identificação de relações de coerência não sinalizadas por conectores. *DELTA: Documentação de Estudos em Linguística Teórica e Aplicada*, v. 33, n. 1, p. 79-108, 2017. DOI: <https://doi.org/10.1590/0102-445025798334674077>.

CARDOSO, P. C. F. *et al.* A Linguagem em foco: anotação de sinalizadores discursivos em textos jornalísticos. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 15., 2024, Belém/PA. *Anais [...]*. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 247-256. DOI: <https://doi.org/10.5753/stil.2024.245329>.

CARDOSO, P. C. F. *et al.* CSTNews – A Discourse-Annotated Corpus for Single and Multi-Document Summarization of News Texts in Brazilian Portuguese. In: WORKSHOP “A RST E OS ESTUDOS DO TEXTO”, 3., 2011, Cuiabá. *Anais [...]*. [S. l.]: Sociedade Brasileira de Computação, 2011. v. 1, p. 88-105. Available at: [http://www.nilc.icmc.usp.br/nilc/download/ariani/CardosoETAL\\_RST\\_2011.pdf](http://www.nilc.icmc.usp.br/nilc/download/ariani/CardosoETAL_RST_2011.pdf). Accessed on: 11 May 2025.

CARLSON, L.; MARCU, D.; OKUROWSKI, M. E. *RST Discourse Treebank*. [S. l.]: Linguistic Data Consortium, 2002. DOI: <https://doi.org/10.35111/4w31-m996>.

CRUZ, G. S. B.; SOUZA, J. W. C.; CARDOSO, P. C. F. Estratégias automáticas para análise da concordância da anotação de Sinalizadores Discursivos. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 15., 2024, Belém/PA. *Anais [...]*. Porto Alegre: Sociedade Brasileira de Computação, 2024. p. 440-444. DOI: <https://doi.org/10.5753/stil.2024.245159>.

CUNHA, I. da *et al.* A Symbolic Approach for Automatic Detection of Nuclearity and Rhetorical Relations Among Intra-Sentence Discourse Segments in Spanish. In: GELBUKH, A. (org.). *Computational Linguistics and Intelligent Text Processing*. Berlin; Heidelberg: Springer Berlin Heidelberg, 2012. p. 462-474. DOI: [http://dx.doi.org/10.1007/978-3-642-28604-9\\_38](http://dx.doi.org/10.1007/978-3-642-28604-9_38).

DANTAS, E. *et al.* *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. São Carlos: Instituto de Ciências Matemáticas e de Computação/ Universidade de São Paulo, 2024. Available at: <https://repositorio.usp.br/item/003207370>. Accessed on: 11 May 2025.

DAS, D.; TABOADA, M. RST Signalling Corpus: A Corpus of Signals of Coherence Relations. *Language Resources and Evaluation*, v. 52, n. 1, p. 149-184, 2018.

FRASER, B. An Account of Discourse Markers. *International Review of Pragmatics*, v. 1, n. 2, p. 293-320, 2009. DOI: <https://doi.org/10.1007/s10579-017-9383-x>.

FREITAS, C. Dataset e corpus. In: CASELI, H. de M.; NUNES, M. das G. V. (org.). *Processamento de linguagem natural: conceitos, técnicas e aplicações em português*. [S. l.]: BPLN – Brasileiras em PLN, 2024. p. 1-37. Available at: <https://brasileiraspln.com/livro-pln/2a-edicao/parte-dados-avaliacao/cap-dataset-corpus/cap-dataset-corpus.pdf>. Accessed on: 11 May 2025.

HERNAULT, H. *et al.* HILDA: A Discourse Parser Using Support Vector Machine Classification. *Dialogue & Discourse*, v. 1, n. 3, p. 1-33, 2010. DOI: <https://doi.org/10.5087/dad.2010.003>.

- LIU, Y.; ZELDES, A. Discourse Relations and Signaling Information: Anchoring Discourse Signals in RST-DT. *Proceedings of the Society for Computation in Linguistics*, New York, v. 2, p. 314-317, 2019. DOI: <https://doi.org/10.7275/vh3w-4240>.
- MANN, W. C.; THOMPSON, S. A. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization. *Text – Interdisciplinary Journal for the Study of Discourse*, v. 8, n. 3, 1988.
- MARCU, D. *The Theory and Practice of Discourse Parsing and Summarization*. London: MIT Press, 2000.
- MAZIERO, E. G. *Análise retórica com base em grande quantidade de dados*. 2016. Tese (Doutorado em Ciências – Ciências da computação e matemática computacional) – Universidade de São Paulo, São Carlos, 2016. DOI: <http://dx.doi.org/10.11606/t.55.2017.tde-13012017-103446>.
- MITKOV, R. Discourse Processing. In: MITKOV, R. (org.). *The Oxford Handbook of Computational Linguistics*. [S. l.]: Oxford University Press, 2022.
- PARDO, T. A. S. *Métodos para análise discursiva automática*. 2005. Tese (Doutorado em Ciências – Ciências da computação e matemática computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2005. DOI: <http://dx.doi.org/10.11606/t.55.2005.tde-29082005-172336>.
- PARDO, T. A. S.; NUNES, M. das G. V. On the Development and Evaluation of a Brazilian Portuguese Discourse Parser. *Revista de Informática Teórica e Aplicada*, v. 15, n. 2, p. 43-64, 2008. DOI: <https://doi.org/10.22456/2175-2745.7015>.
- PRASAD, R.; WEBBER, B.; JOSHI, A. The Penn Discourse Treebank: An Annotated *Corpus* of Discourse Relations. In: IDE, N.; PUSTEJOVSKY, J. (org.). *Handbook of Linguistic Annotation*. Dordrecht: Springer Netherlands, 2017. p. 1197-1217. DOI: [http://dx.doi.org/10.1007/978-94-024-0881-2\\_45](http://dx.doi.org/10.1007/978-94-024-0881-2_45).
- RODRIGUES, R.; SOUZA, J. W. C.; CARDOSO, P. C. F. Sinalizadores retórico-discursivos: revisitando a anotação RST no corpus CSTNews. In: SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL), 14., 2023, Belo Horizonte. *Anais [...]*. Porto Alegre: Sociedade Brasileira de Computação, 2023. DOI: <http://dx.doi.org/10.5753/stil.2023.234120>.
- SIDNER, C. *A Progress Report on the Discourse and Reference Components of PAL*. [S. l.]: Massachusetts Institute of Tech Cambridge Artificial Intelligence LAB, 1978.
- SOUZA, J. W. C.; DI FELIPPO, A.; PARDO, T. A. S. Investigação da identificação da redundância na sumarização multidocumento. In: WORKSHOP DE INICIAÇÃO CIENTÍFICA EM TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA, 3., 2013, Fortaleza. *Anais [...]*. [S. l.]: Sociedade Brasileira de Computação, 2013. v. 1, p. 22-24. Available at: <https://sites.icmc.usp.br/taspardo/TILic2013-SouzaEtAl.pdf>. Accessed on: 11 May 2025.
- 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. DOI: <https://doi.org/10.5087/dad.2013.211>.
- TABOADA, M.; MANN, W. C. Rhetorical Structure Theory: Looking Back and Moving Ahead. *Discourse Studies*, v. 8, n. 3, p. 423-459, 2006. DOI: <https://doi.org/10.1177/1461445606061881>.
- ZELDES, A. rstWeb – A Browser-Based Annotation Interface for Rhetorical Structure Theory and Discourse Relations. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: DEMONSTRATIONS, 2016, San Diego. *Proceedings [...]*. Stroudsburg, PA: Association for Computational Linguistics, 2016. DOI: <http://dx.doi.org/10.18653/v1/n16-3001>.

## History

Date of submission: 29/06/2024.

Date of approval: 01/07/2025.

## Responsible Editor

Andréa Machado de Almeida Mattos, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais/MG, Brasil. Lattes: <http://lattes.cnpq.br/7749222257907067>, ORCID: <https://orcid.org/0000-0003-3190-7329>, e-mail: [andreamattos@ufmg.br](mailto:andreamattos@ufmg.br).

## Authors' Contribution

All authors contributed to the construction of taxonomy and *corpus* study. Furthermore, all authors contributed to writing the text submitted to the journal.

## Declaration of Conflict of Interest

The authors declare that there is no conflict of interest.

## Data Availability Statement

The data used in this research are available in this electronic repository: <https://github.com/Projeto-RST/anotacao-sds-cstnews>.

## Use of AI

No Artificial Intelligence was used in this work.

## Reviews

As part of the commitment made by the Brazilian Journal of Applied Linguistics to Open Science, the journal publishes the reviews issued regarding its published works, when authorized by all parties involved.

## Review 1

After carefully reviewing the manuscript, I am pleased to inform you that I approve it without any reservations. The content is well-structured and meets the required standards. I am confident that it will make a valuable contribution to Applied Linguistics. Thank you for your hard work, and I look forward to seeing it move forward.

**Reviewed on:** 12/12/2024.

**Recommendation:** Accept Submission.

**Reviewer:** Eduardo Batista da Silva.

## Review 2

### Round 1

I find this paper to be a valuable contribution to the field, particularly in the area of discourse analysis and Rhetorical Structure Theory (RST) for Brazilian Portuguese. The research addresses a critical gap in identifying discourse signals beyond traditional markers and develops a taxonomy that could be instrumental for both linguistic theory and natural language processing (NLP). However, before acceptance, the manuscript requires revisions to strengthen its impact and address specific concerns.

### Strengths of the Paper

- 1 **Timely and Relevant Topic:** The exploration of discourse signals in Brazilian Portuguese is a novel and under-researched area, adding diversity to existing RST literature predominantly focused on English.
- 2 **Methodological Rigor:** The use of the CSTNews corpus, a pre-annotated resource, provides a solid empirical basis for the study. The adoption of tools like rstWeb is a modern approach for annotation.
- 3 **Contribution to NLP and Linguistics:** The taxonomy has potential applications in computational tasks like automated parsing, summarization, and coherence analysis, bridging linguistic theory and NLP practices.
- 4 **Clarity in Taxonomy Presentation:** The detailed organization of discourse signals into categories and subcategories, with examples, is a highlight and makes the findings accessible.

### Major Issues to Address

- 1 Theoretical Justification:
  - A The study heavily relies on previous works but lacks critical discussion of how its findings challenge, expand, or refine existing theories of discourse analysis.
  - B Include a more robust comparison between your taxonomy and existing taxonomies (e.g., Das and Taboada, 2018; Liu and Zeldes, 2019) to emphasize the distinctiveness of this work.

2 Generality of the Findings:

- ◆ The study is limited to formal journalistic texts. While this is acknowledged, it reduces the general applicability of the taxonomy. Address this limitation more explicitly in the discussion and propose concrete steps for expanding the research to other genres.

3 Signal Ambiguity and Annotator Agreement:

- ◆ There is limited discussion of how signal ambiguity was handled during annotation. Include inter-annotator agreement statistics or details on how discrepancies were resolved to validate the reliability of the taxonomy.

4 Methodology Clarity:

- ◆ Provide more detailed steps regarding how the signals were extracted and annotated.

### Minor Issues to Address

1 Language and Writing:

- ◆ The abstract and conclusion could be more concise while emphasizing the core contributions.

2 Practical Applications:

- ◆ Expand on the practical implications of the taxonomy for NLP tools. Discuss how it might improve existing discourse parsers or text analysis systems.

### Conclusion

The paper has substantial merit and could significantly contribute to the field of corpus linguistics and discourse analysis after addressing the issues outlined above. A revise decision allows the authors to refine their arguments, clarify methodologies, and strengthen their claims, ensuring the paper meets the journal's standards for publication.

Reviewed on: 08/01/2025.

Recommendation: Revisions Required.

### Round 2

Reviewed on: 01/07/2025.

Recommendation: Accept Submission.