Rev. Estud. Ling., Belo Horizonte, v. 31, n. 3, p. 1643-1675, 2023.

Systematic Review of Studies on Rhetorical Structure Theory (RST)

Revisão sistemática de estudos sobre Rhetorical Structure Theory (RST)

Jackson Wilke da Cruz Souza
Universidade Federal da Bahia (UFBA), Camaçari, Bahia / Brasil
jackcruzsouza@gmail.com
https://orcid.org/0000-0003-1881-6780

Paula Christina Figueira Cardoso
Universidade Federal do Pará (UFPA), Belém, Pará / Brasil
paulastm@gmail.com
https://orcid.org/0000-0003-3621-8960

Roana Rodrigues
Universidade Federal de Sergipe (UFS), São Cristóvão, Sergipe / Brasil
roana@academico.ufs.br
https://orcid.org/0000-0002-7748-8716

Abstract: This paper presents a systematic review of studies published between 2010 and 2022 couched within the theoretical-methodological framework of Rhetorical Structure Theory (RST). Using "Publish or Perish" software, we extracted from Web 760 works related to RST and, considering the number of citations, we analyzed the first 100 results that were organized and described based on their abstracts. For didactic purposes, we classified these studies into the following criteria: (i) works that couldn't be analyzed due to accessibility issues; (ii) works focusing on theorization and the description of various linguistic phenomena; (iii) studies using *corpus* creation and exploration; and (iv) investigations on computational applications in Natural Language Processing (NLP). In addition, among the data collected, we conducted a brief analysis of RST works developed by Brazilian researchers. As a result, we present an overview of RST studies in the last decade, allowing for the creation of research programs that consider the projects already developed and the advances of the area in Brazil and worldwide.

Keywords: RST; discourse; coherence relations; natural language processing.

eISSN: 2237-2083 | DOI: 10.17851/2237-2083.31.3.1643-1675

Resumo: Este artigo apresenta uma proposta de revisão sistemática de trabalhos publicados entre 2010 e 2022 que têm como base teórico-metodológica a Teoria da Estrutura Retórica, mais comumente nomeada RST (*Rhetorical Structure Theory*). Utilizando a ferramenta "Publish or Perish", foram extraídos da *Web* 760 artigos relacionados com RST.Considerando o número de citações, os 100 primeiros resultados foram analisados, organizados e descritos com base em seus resumos. Para fins didáticos, optou-se por classificar tais estudos em: (i) trabalhos não analisados, devido a alguma impossibilidade de acesso; (ii) trabalhos em RST com foco em teorização e na descrição de variados fenômenos linguísticos; (iii) trabalhos em RST com uso de compilação e exploração de *corpus*; e (iv) trabalhos em RST e aplicações computacionais, na área de Processamento de Língua Natural (PLN). Além disso, dentre os dados levantados, realizou-se uma breve análise de trabalhos em RST desenvolvidos por pesquisadores brasileiros. Como resultado, é possível ter acesso a um panorama dos estudos da área na última década, possibilitando a criação de programas de investigação, levando em consideração os projetos já desenvolvidos no Brasil e no mundo.

Palavras-chave: RST; discurso; relações de coerência; processamento de línguas naturais.

Recebido em 10 de abril de 2023.

Aceito em 24 de novembro de 2023.

1 Introduction

The relationships established between the elements within a text for the construction of meaning are quite complex, even for human interpretation. Therefore, the annotation and processing of discourse data is seen as a significant challenge for linguistic description and Natural Language Processing (NLP), also known as Computational Linguistics, an area dedicated, roughly speaking, to the creation of computing resources that can understand, interpret and manipulate human language.

Among the various proposals for describing the rhetorical relations - or coherence relations - established in a text, the contributions of Rhetorical Structure Theory (RST) stand out, a theory initially proposed by William Mann and Sandra Thompson, in the late 1980s. According to Hirata-Vale and Oliveira (2014), RST forms a part of the so-called North American West Coast Functionalism, which understands language as a flexible system, molded in and by use. Unlike other functionalist

approaches, RST does not work at the complex clauses level, but at the discourse level, investigating the explicit and implicit propositional contents between parts of the text to construct and interpret coherent and cohesive discourses. The authors also point out that the theory is used both in descriptive linguistic works and in research of NLP.

Thus, considering the relevance of the theory for linguistic and computational studies, we present an overview of works based on RST in the 21st century, specifically scientific investigations carried out under this theoretical basis published between 2010 and 2022[1]. This paper is the result of discussions carried out within an interinstitutional research project dedicated to analyzing rhetorical relations in Brazilian Portuguese (BP) under the theoretical assumptions of RST. In order to establish an agenda for the group's investigative actions, it seemed prudent to carry out, initially, a systematic survey of RST work in the world.

Therefore, this paper outlines the steps we took to conduct a systematic review of RST, structured as follows: in Section 2, we present the theoretical-methodological foundations of RST. Section 3 states the methodology we employed to survey RST research conducted in the past decade. In Section 4, we examine the primary subjects and research findings in the field. Finally, we offer concluding remarks and outline future research directions within the scope of this research project.

2 Rhetorical Structure Theory

RST was developed in the 1980s at the University of Southern California, in the United States, by a group of researchers interested in Natural Language Generation. According to Taboada and Mann (2006), initially RST aimed at developing a model that to guide computational text generation, however, it was adopted by researchers from diverse areas and for different purposes, such as teaching, description, and NLP, helping in the better understanding of the text and in the proposal of a conceptual structure of the coherence relations.

For RST, the minimal element of the analysis are the units, which are close to the concept of clauses used in traditional BP grammars.

[1] This temporal cut was mainly made considering the contributions of Taboada and Mann (2006), which is also of bibliographical nature, and presents the main RST research from its inception until that moment.

Units are constituted by nucleus (N), the most important part; and satellite (S), which, despite playing a secondary role, in some cases, can contribute to a better understanding of the N. Each rhetorical relation is defined in terms of four fields: constraints on the N; constraints on the S; constraints on the combination of N and S; and effect (achieved on the text receiver). Relations composed of one nucleus and one satellite are named mononuclear relations. On the other hand, in multinuclear relations, two or more nuclei participate and have the same importance. The relationships are traditionally structured in a tree-like form.

The RST taxonomy is flexible, resulting in different numbers of coherence relations based on the particular project and the language being studied. However, Taboada and Mann (2006) warn about cases of an increase in the number of relations, as having too many possibilities for classification results in greater difficulty in manually analyzing texts. There are different taxonomic proposals for RST relationships, such as the one made by Mann and Thompson (1987), who proposed 24 coherence relations. These two proposals were based on analyses of English texts. For BP, we can mention the contributions of Pardo (2005), who presented 32 coherence relations[2]. In (1), there is an example of *Explanation* relation, taken from Pardo (2005, p. 169):

(1) [and the readability index is calculated,] [*that is,* an indicator of difficulty in understanding the text.][3]

In Example (1), the *Explanation* relation is characterized by an N (*and the readability index is calculated*) that presents an event or situation; and an S (*that is, an indicator of difficulty in understanding the text*) with no filling restrictions. From the N+S relation, it is established that S explains how and/or why the event or situation presented in N occurs or came to occur. This relationship causes the effect in the reader of

[2] Coherence relations in Brazilian Portuguese, according to Pardo (2005): antithesis, attribution, background, circumstance, comparison, concession, conclusion, condition, contrast, elaboration, enablement, evaluation, evidence, explanation, interpretation, joint, justify, list, means, motivation, non-volitional cause, non-volitional result, otherwise, parenthetical, purpose, restatement, same-unit, sequence, solutionhood, summary, volitional cause and volitional result.

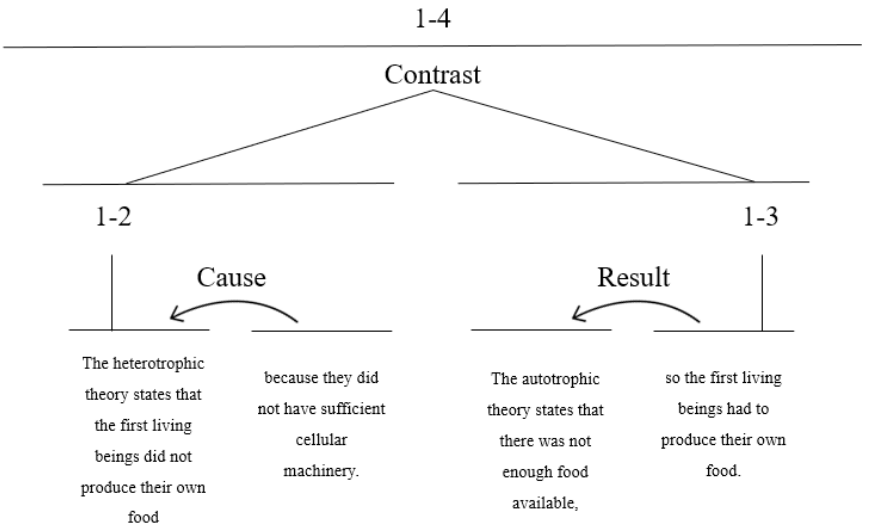
[3] The original text in Portuguese by Pardo (2005, p. 169) is: "[e é calculado o índice de legibilidade,] [**isto é**, um indicador de dificuldade de entendimento do texto.]".

recognizing that S is the reason for N or that S explains how N occurs. Still in (1), the discourse marker (*that is*) is the textual signal of *Explanation*.

While the provided example showcases an over connective, which enhances the comprehension and categorization of the rhetorical relations (coherence relation), Antonio (2017, p. 105) stipulates that these relations are rooted in semantics rather than form. This semantic foundation allows for their establishment and interpretation autonomously, irrespective of the presence of the explicit connective markers. Hence, there is a need for a study of other signals, beyond explicit discourse markers, to adequately describe and annotate rhetorical relations. These signals may include punctuation marks, phonological elements such as intonation, morphosyntactic features like verb tense, semantic elements such as the interplay between states-of-affairs, cognitive factors such as the activation of referents from a global cognitive model, among others (Antonio, 2017; Das; Taboada, 2018).

In Figure 1, based on Antonio (2017, p. 105) and translated from the original language - Portuguese, the *Contrast* relationship is observed even without the existence of an explicit discourse marker in the text is illustrated:

Figure 1 – Example of RST relations



Source: Antonio (2017, p. 88).

As Antonio (2017) explains in Figure 1[4], the Contrast relation is primarily due to the morphological markers of the morphemes *hetero* and *auto* in the lexemes heterotrophic and autotrophic. In essence, RST is a descriptive theory that employs selective, structured forms to provide explicit representations of a text's coherence and organization. Its structure facilitates the development of rigorously annotated corpora. This theory contributes significantly to various NLP applications, including automatic summarization, anaphora resolution, automatic translation, polarity classification of sentences in opinion blogs, and more (Cardoso, 2014, p. 37-38). According to Taboada and Mann (2006), we can categorize the diverse applications of RST into four major domains, which include:

- RST and NLP: parsing, summarization, argument evaluation, automatic translation, essay evaluation, among others.
- RST and cross-linguistic studies: study of different languages, making comparisons and cross-linguistic generalizations.
- RST and dialogue and multimedia: studies that totally (or partially) use RST to describe the relationships established in more "dynamic" phenomena, such as dialogical interactions and multimedia environments (textual formatting, hypertexts, text and video, text and figures, text, and gestures, etc.).
- RST and discourse analysis, argumentation, and writing: RST is used to describe and understand the structure of texts, as well as its relationship with other phenomena such as anaphora and cohesion. Thus, in this category, there are studies based on RST for the elaboration of discourse analysis, studies of argumentation and the analysis and teaching of writing.

To examine the evolution of RST in the 21st century, the following sections will present the methodology and data analysis of studies conducted on RST over the last decade. As already explained, this type of investigation carried out in a systematic way helps to establish

[4] The original text in Portuguese by Antonio (2017, p. 86) is: "[A teoria heterotrófica diz que os primeiros seres vivos não produziam seu próprio alimento] [porque não tinham uma maquinaria celular suficiente.] [A teoria autotrófica diz que não havia alimento suficiente], [daí os primeiros seres vivos tiveram que produzir seu próprio alimento.]".

an overview of research projects for other languages and, mainly, for BP, which contributes to the delimitation of the state-of-the-art and for possible directions for investigations in the area.

3 Methodology

This investigation is characterized by conducting a thorough bibliographical review. According to Gil (2002), this type of research is based on previously elaborated materials, primarily books and scientific papers. For this purpose, the author categorizes the bibliographic sources into three types: books (as reference reading), periodical publications (including academic journals and magazines), and various printed materials. Cervo, Bervian, and Silva (2007) emphasize that bibliographical research can be considered a fundamental component of all scientific studies, but it can also stand alone as an independent research method.

In addition to bibliographical research, a bibliometric approach has been employed as a methodological strategy in this study. This approach involves extracting metrics to assess the pertinence and relevance of the works analyzed. Moreira, Guimarães, and Tsunoda (2020) highlight several possibilities in bibliometric studies, including: (i) identifying current advancements in specific knowledge areas; (ii) providing a comprehensive basis for evaluating scientific publications; and (iii) assessing academic production. Additionally, the authors emphasize the ability of bibliometrics to uncover specific perspectives within a scientific field or knowledge domain by examining the individuals and institutions involved in the research or the applications derived from the studies.

Zupic and Čater (2015) highlight the procedures adopted in bibliometric studies. The authors point out the need to define a research question; select appropriate bibliometric methods; choose the bibliometric methods used to answer such a question; select the database; use bibliometric software; and decide which visualization method to use to represent the findings generated from the chosen tool.

Moreira, Guimarães and Tsunoda (2020) analyzed several bibliometric software, and among them, we chose to use Publish or Perish (Harzing, 2007) in this work. According to the authors, although Publish or Perish software has more limitations regarding the visualization of retrieved bibliometric data, it is possible to analyze a series of databases in the same search. This criterion justifies our choice for this software

because in previous tests, it analyzed only databases separately (such as Scopus and Web of Science), failing to consider studies in relevant repositories for NLP, such as the Association for Computational Linguistics (ACL) Anthology[5].

We used two essential search criteria in the tool: search term ("Rhetorical Structure Theory")[6] and publication year of the work (from 2010 to 2022), resulting in 760 occurrences. In the quantitative analysis, we examined the 100 studies that presented five or more citations, representing, at first, a greater circulation among RST specialists and researchers. The results obtained from the software were organized in a table in .xls format. For this research, we observed the following data: (i) Number of citations, (ii) Authors, (iii) Title of the work, (iv) Year of publication, and (v) Source. We manually included Language, Study Area, and Application (where applicable) in the analysis.

We divided the set of 100 academic works among three researchers and classified them into four categories, initially based on the classification by Taboada and Mann (2006) and according to the identification of specificities between the works. As a result, we identified that 23 papers only mentioned RST but did not have it as their main focus of study and/or needed free access, leading them to be disregarded in this research. In the next section, we present the classification of studies, their themes and impacts.

Despite the significance and contributions of many studies, including those in the Portuguese language, there are several hypotheses that justify why they were not systematically cited throughout our study: (i) the ACL-Anthology is not considered an scientific indexed database, which may lead to works not being retrieved by databases; (ii) even though there is a clear scientific contribution in many investigations, they did not reach the citation threshold we set for our analysis; (iii) as

[5] We recognize the relevance of repositories in the NLP field, such as the ACL Anthology (https://aclanthology.org), and the website dedicated to RST studies worldwide (https://www.sfu.ca/rst/index.html), which contains manuals, resources, tools, and bibliographies in the field. However, it is important to emphasize that this research only described the works that emerged as a result of the bibliometric research conducted, following the described methodological procedures. Web pages accessed in December 2022.

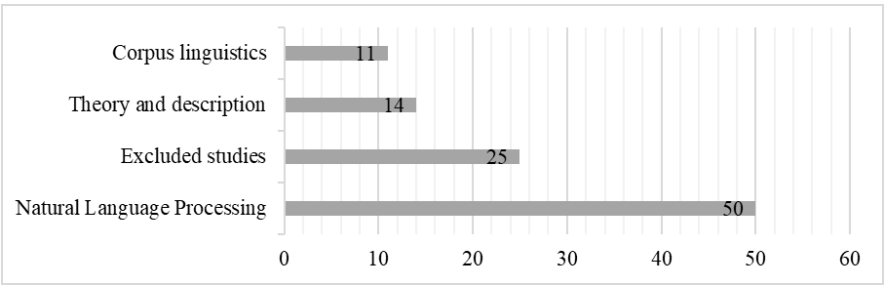
[6] The only search term used was "Rhetorical Structure Theory," as there was no significant difference in the results when combined with related terms such as "discourse" or the acronym "RST."

we used the Google database, we measured the impact factor of works using the H-5 index, which resulted in recent works published in the last five years were less prevalent in our analysis.

4 Bibliometric Analyses

We proceeded to the effective analysis of their themes and contents, based mainly on the data described in their abstracts. The findings refer to data collected in October 2022, with studies published until June of the same year. Figure 2 shows the distribution of studies based on the areas identified in this investigation.

Figure 2 – Distribution of works by area



Source: Prepared by the authors.

Figure 2 organizes the works into four areas. In Theory and description, we group studies that characterize and identify RST relations, in addition to research that recovers what the literature proposes as a theory for the model. In *Corpus* Linguistics, we organized works that explore, compile and/or annotate linguistic corpora according to the RST model. In Natural Language Processing, we selected studies that approach RST from computational applications. Finally, in Excluded studies, we point out the works in which it was impossible to have access or only mentioned RST, without being the focus of the research.

It is important to emphasize that, although we have organized the studies in these categories, many of them move between areas - or can contribute with discussions to other categories. The main discussions and contributions of these studies will be presented in the next sections.

4.1 RST and Theoretical and Descriptive Works

The works presented in this section emphasize RST guidelines for computational applications and/or present descriptions of different natural language phenomena with this theory as a basis. Fourteen works were analyzed, all written in English, although their content involves the description of other natural languages, individually or based on contrastive studies (German, Arabic, Basque, Spanish, and English).

The three theoretical studies listed here conduct a bibliographic survey of the area of automatic summarization (Alami *et al*., 2015) and automatic identification of fake news (Conroy, Rubin, Chen, 2015; Oshikawa, Qian, Wang, 2020). Although they have different topics, they present RST as a foundational theory for the mentioned descriptive-computational endeavors, which are contemporary threads and concerns for NLP.

On the other hand, studies that propose descriptions and analysis based on RST have different objects of study, classified into three topics, as follows:

a) Identification and analysis of coherence relations: Das and Taboada (2013) claim that, until the time of the publication of their work, research on RST focused on analyzing only discourse markers as signs of coherence relations, considering that any other interpretation would be understood as an implicit relation - not explicit. Their study, however, goes against the grain of these works in giving visibility to other signs (morphology, lexical, syntax, semantic, graphical, etc) for the interpretation of relation. In the same direction, we can mention the contributions of Jasinskaja and Karagjosova (2020). Although the authors propose a predominantly theoretical work, they understand that coherence relations go beyond analyzing discourse markers and anaphoric phenomena, contemplating discussing the different classes of relations that aim to establish discourse coherence. The study of coherence relations can also contribute to understanding other language phenomena, as exemplified by the investigations of Matthiessen (2015) and Matthiessen and Teruya (2015), who, based on RST, analyze the semantic organization of texts in English from different linguistic registers.

b) Analysis of coherence relations in different genres: In addition to discussing the rhetorical relations themselves, some investigations highlight the particularities of these relations considering specific documents and textual genres, as in the work of Taboada and Habel (2013), who discuss coherence relations in multimodal documents (which present textual and visual elements); the research of Peldszus and Stede (2013, 2016), which consider coherence relations and the construction of arguments in a *corpus* of short micro texts; the analyzes by Abrahamson and Rubin (2012), which compare lay (consumer) and professional (physician) discourse structures in answers to health questions; and Green's (2010) work that presents a study of argument presentation in a biomedical *corpus* within the framework of RST.

c) Comparative analysis of coherence relations: The comparative studies discussed here refer to establishment of coherence relations in texts of different languages. Da Cunha and Iruskieta (2010) propose a contrastive study of rhetorical structures in a parallel *corpus* of medical texts in Spanish and Basque. The results indicate that, in translation processes, the rhetorical structure needs to be considered as much as the syntactic structure. Discrepancies between the choices of coherence relations were also visible in the investigation by Taft *et al*. (2011). The authors analyze texts written in English by native and foreign speakers (Chinese and Spanish speakers) and conclude that rhetorical achievements and preferences are different according to the mother tongue of each research participant.

The works seem to elucidate the interests related to RST in recent years, as follows: (i) the development of resources for NLP; (ii) the identification and detailed analysis of coherence relations in one or between languages; and/or (iii) the study of the signal markers (which go beyond of the already well-studied discourse markers) that trigger rhetorical interpretations.

4.2 RST and *Corpus* Linguistics

In recent decades, *Corpus* Linguistics has witnessed significant advances. Historically, *corpus* annotation had been predominantly confined to the domains of morphology, syntax and semantics. However, in the last 20 years, there has been a notable expansion into discourse-

level annotation; not without enormous efforts. Prominent exemplars of discourse annotation frameworks include RST-DT - Discourse Treebank Rhetorical Structure Theory (Carlson, Marcu, 2001), SDRT - Segmented Discourse Representation Theory (Asher, Lascarides, 2003) and PDTB - Penn Discourse Treebank (Prasad *et al*. 2008).

In the current research project, we have categorized 11 works related to *Corpus* Linguistics. It is worth emphasizing that a substantial portion of these studies appears to straddle the interface between *Corpus* Linguistics and the domains of descriptive linguistics or NLP. This overlap arises from the fact that a majority of these studies leverage corpus data for conducting linguistic analyses and/or implementing computational applications. Consequently, our categorization decision was primarily influenced by the prominence assigned to processes related to corpus construction, segmentation, and annotation. In the following topics, we present general considerations about these investigations:

a) *Corpus* construction: Respecting its particularities and goals, the work of Van Der Vliet *et al*. (2011), Da Cunha, Torres-Moreno and Sierra (2011) and Iruskieta (2013) present the process of creating, segmentation and discourse annotation of *corpus* in Dutch, Spanish and Basque, respectively. Cardoso *et al*. (2011) created and annotated CSTNews[7], a Portuguese-language *corpus* composed of texts from BP newspapers and abstract and extractive summaries, mono and multi-documents, produced automatically and manually. It is a resource with different layers of annotation (morphosyntactic, discourse and topical, for example) and, more specifically concerning the discourse level, presents annotation based on two theories: RST model and Cross-document Structure Theory (CST)[8]. In addition, we can mention the contributions of Zeldes (2017), the GUM *corpus* composed of news, interviews, instructional and informative texts in English. That work is different because it was developed in a classroom context, demonstrating the possibility of creating rich linguistic resources

[7] CSTNews, available at: http://nilc.icmc.usp.br/CSTNews/login/?next=/CSTNews/. Accessed in December 2022.

[8] RADEV, Dragomir. A common theory of information fusion from multiple text sources step one: cross-document structure. *In*: *1st SIGdial workshop on Discourse and dialogue*. 2000. p. 74-83.

in pedagogical practices. Das and Taboada (2018) present the RST Signaling *Corpus*[9], an annotated corpus for coherence relations signals. The *corpus* includes annotations of discourse markers considered the most typical signals in discourse, and a wide range of other signals, such as reference, lexicon, semantics, syntactic, graphic and gender characteristics as potential indicators of coherence relations. Finally, we report the research by Zhong *et al*. (2020), in which the manual and semi-automatic process of compiling and analyzing a *corpus* of simplified English texts is described, to identify the strategies used and predict the exclusion of phrases for textual simplification.

b) *Corpus* and contrastive/comparative studies: The work of Iruskieta, Da Cunha, and Taboada (2015) represents studies in RST that contrast such relations through the construction, annotation, and analysis of multilingual *corpora*. The authors compare coherence relations in texts written in English, Spanish, and Basque. Notably, they display substantial similarities. The principal aim of these studies is to introduce a novel qualitative methodology for contrasting coherence structures across different languages and to elucidate the reasons behind disparities in coherence structures within translated texts. The remaining studies examined RST annotation in relation to different options for annotating discourse. Stede *et al*. (2016) present an annotation of 112 short texts, and *corpus* analysis in two approaches: RST and SDRT, which made it possible to establish correlations between the annotations taken and between the structure of the discourse and the argumentation. Additionally, the research by Stab *et al*. (2014) addresses the structure of arguments, offering insights into the process of discourse annotation with the intent of modeling argument components and structure within persuasive essays at the sentence level. On the other hand, Sanders *et al*. (2021) propose a unified framework for annotating rhetorical structures derived from various theoretical perspectives, including PDTB, RST and SDRT.

We recognize that we present a small sample of works that establish a direct relationship between RST and *Corpus* Linguistics. However, we believe it is important to keep them as a separate category, precisely to emphasize their relevance in linguistic-descriptive-

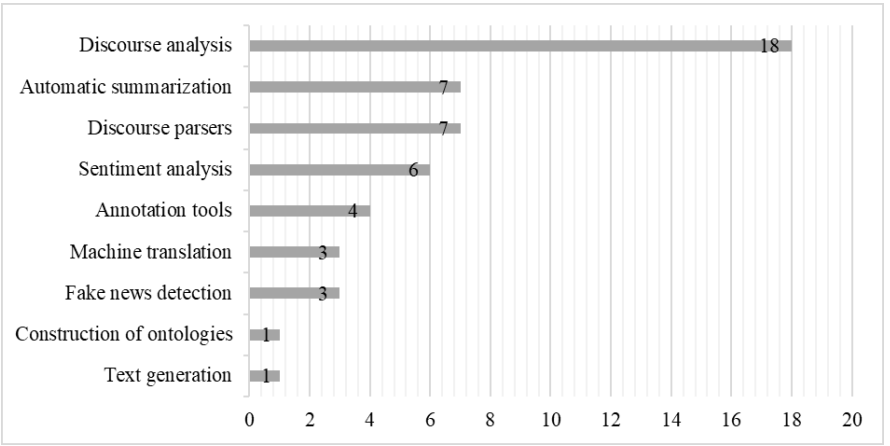
[9] Available at: https://catalog.ldc.upenn.edu/LDC2015T10. Accessed in December 2022.

computational studies and to serve as a basis for the development of another research.

4.3 RST and NLP

We categorized 50 works that explore the intersection of RST to NLP. To accomplish this, we consider studies that have RST as a central topic and that present some linguistic-computational application. As Figure 3 illustrates, we grouped the works into nine categories based on the type of application.

Figure 3 – Distribution of NLP works by categories



Source: Prepared by the authors.

a) Construction of ontologies: ontologies are used to organize and represent information, highlighting the relationship that eventually exists in that representation. In this category, we found only one study by Mitrović *et al*. (2017). The authors elucidate the utility of rhetorical figures for the extraction of argumentative mining purposes in different text genres, such as implicit hate speech, fake news, and complex arguments. The authors highlight that the adoption of a formal ontological representation of rhetorical figures has the potential to exert a beneficial impact on research endeavors associated with argumentative mining.

b) Text generation: in this category, the focus of works around natural language generation originating from the output of a computational system. A distinct work, authored by Konstas and Lapata (2013), tackles the challenge of text generation from a database by employing a trainable generation system that encompasses content selection and ordering. Content planes are intuitively represented through a set of grammatical rules that operate at the document level and are autonomously acquired from training data. The authors have developed two approaches: first, inspired by RST, involves representing the document as a tree of discourse relationships between database records; second, requiring minimal linguistic sophistication, employs tree structures to depict overarching patterns of database record sequences within a document. Konstas and Lapata assert that their experimental evaluations yielded satisfactory results for both methodologies when compared to the current state-of-the-art approaches.

c) Automatic summarization: this NLP area aims to automatically produce a smaller, coherent and cohesive version of a source text from discourse analysis. In this category, we have classified works that incorporated RST annotation, either through manual or automatic process, with a specific emphasis on discourse markers. RST offers distinct advantages for summarization by identifying the nucleus as the most salient information when compared to the satellite. In certain communicative contexts, the satellite information can be omitted without detriment to text comprehension. The majority of works in this area are dedicated to extractive summarization, where the summary is constructed by joining unaltered sentences from the source text. Consequently, they may encounter challenges related to the coherence between selected segments for the summary, as discussed by Hirao *et al*. (2013) and Li, Thadani and Stent (2016) discuss. On the other hand, abstract summarization, as presented by Le and Le (2013), allows for adaptations and rewritings within the summary without altering the primary content. In terms of discourse units, it was observed that some works focus on sentence-level analysis (*e.g*., Louis; Joshi; Nenkova, 2010; Azmi; Al-Thanyyan, 2012; Kikuchi *et al*., 2014) and others emphasize segment-level analysis (*e.g*., Uzêda; Pardo; Nunes, 2010; Li; Thadani; Stent, 2016).

d) Discourse parsers: studies in this category are centered on the development and/or improvement of discourse parsers based on the RST model, catering to various languages. Notable instances include an English parser named HILDA, initially proposed by Hernault *et al*. (2010), which was subsequently refined by Feng and Hirst (2012) through the incorporation of linguistic filters and sentence context. Muller *et al*. (2012) pioneered the creation of the first RST parser for the French language. Additionally, Joty, Carenini and Ng (2015) introduced the CODRA parser for English, while Surdeanu, Hicks and Valenzuela-Escárcega (2015) offered proposed two parsers for English: one employing resources dependent on dependency syntax and another incorporating information from constituent and dependency syntax, along with coreference data from RST. Anita and Subalalitha (2019) presented the Thirukkural Discourse Parser for Tamil, and Lin *et al*. (2019) developed a neural framework for sentence-level discourse analysis based on the RST model for the English language.

e) Discourse analysis: this category comprises works that focus on the automated analysis of the discourse of a text's discourse, interpreting it as an understanding it as a highly elaborate underlying structure that interconnects all its content, thus imbuing it with coherence. There were a total of 18 works pertaining to the domain of automatic discourse analysis. Notably, these studies conducted automated analyses of discourse structure within diverse textual genres, such as argumentative, interviews and posts of social media posts. These analyses employed varying approaches, including *linguistic* methods utilizing combination of words or discourse markers (*e.g*., Biran; Rambow, 2011a, 2011b; Feng; Lin; Hirst, 2014; Jansen; Surdeanu; Clark (2014); Li; Li; Hovy, 2014; Hayashi; Hirao; Nagata, 2016; Katz; Albacete, 2016; Li; Sun; Joty, 2018; Kobayashi *et al*., 2020); *hybrid* techniques that combine Machine Learning (ML) methods with the presence of discourse markers in texts (*e.g*., Allen; Carenini; NG, 2014; Wang; Li; Wang, 2017; Morey; Muller; Asher, 2017, 2018); and or *computational* approaches involving unsupervised ML methods (*e.g*., Li; Li; Chang, 2016; Braud; Plank; Sogaard, 2016; Ji; Smith, 2017; Chakrabarty *et al*., 2020). An exception to these methodologies was observed in the work of Ge and Herring (2018), which adopted a *multimodal* approach by analyzing rhetorical and discourse structures using sequences of emojis in Chinese texts. The

authors employed computer-mediated discourse analysis to investigate possible pragmatic meanings that could be captured by strings of emojis and their rhetorical relations from Chinese social media. The results demonstrated that these sequences pragmatically functioned as verbal utterances and established relationships with textual units.

f) Machine translation: we noted three works related to RST and machine translation. First, the research conducted by Tu, Zhou and Zong (2013), which applies RST in an automatic translation system from Chinese to English. This research follows a structured three-stage process that involves construction of an RST tree, extracting rules, and performing translation. Secondly, the multilingual research led by Guzmán *et al*. (2014) spans English, French, German and Spanish. This work investigates the utilization of rhetorical structure to enhance machine translation evaluation. The evaluation is based on assessing the similarity of kernels of subtrees which allows for a comparison of the rhetorical structure of each. Finally, the research by Joty *et al*. (2014), which utilizes discourse structure and neural networks to compare the discourse tree of a machine translation with that of the human reference, enabling a detailed analysis of the quality of machine-generated translations.

g) Sentiment analysis: works in this category are dedicated to enhancing discourse analysis through the classification of polarity, taking into account the semantic embedded within the rhetorical structure connecting sentences and paragraphs. Our analysis has revealed a spectrum of outcomes, ranging from parsers, exemplified by Heerschop *et al*. (2011), who discern the significance of textual content through RST relations, to broader frameworks employed by researchers such as Zhou *et al*. (2011), Chenlo, Hogenboom and Losada (2014), Bhatia *et al*. (2015), Hogenboom *et al*. (2015) and Kraus, and Feuerriegel (2019). We emphasize the works in this category encompass diverse text genres, such as journalistic texts, blog texts, and product reviews.

h) Annotation tools: the creation of *corpora* dedicated to RST and discourse parsers had a bigger growth when compared with the development of annotation tools. Notably, the most widely recognized annotation tools, RSTTool and the ISI RST Annotation Tool, are no longer receiving updates. In our bibliometric investigation, we have identified two annotation resources that align with contemporary technological

standards and requirements: RSTWeb (Zeldes, 2016) and TreeAnnotator (Helfrich *et al*., 2018). Both of these tools are browser-based, enabling project managers to gather data, without the need for file exchange with annotators. Moreover, they facilitate the tracking of progress and the automatic recording of annotation processes.

i) Fake news detection: in this category, we have classified works focused on the identification of fake news, a topic of interest and relevance to the NLP in recent years. Researchers have been instrumental in highlighting the field's concern with establishing connections between RST, particularly concerning the analysis of textual structure and its coherence, and the detection of fake news (*e.g*., Rubin, Vashchilko, 2012; Rubin, Conroy, Chen, 2015; Rubin, Lukoianova, 2015). These studies have proposed various approaches to differentiate genuine stories from deceptive ones, a task that, as indicated by conducted experimentals, presents a challenge even for human classification. Furthermore, we have come across research conducted by Jansen, Surdeanu and Clark (2014), which introduces a model for reclassifying responses to real questions found on the web. This model employs two discourse representations: one centered around discourse markers and the other grounded in RST relations.

The works explored in the NLP category, as illustrated, encompass a spectrum of applications and approaches, occasionally leaning more towards computational foundations and at other times emphasizing linguistic aspects. It is evident that some of the more recent research endeavors extend beyond the confines of automatic processing of rhetorical text relations. They address broader themes and requirements, including but not limited to well-established NLP applications like sentiment analysis and fake news detection.

4.4 RST Studies in Brazil

Given the methodological decisions we adopted in this study, it is regrettable that we were unable to provide a more comprehensive description and analysis of some undoubtedly pertinent research conducted in Brazil. Among the 100 works analyzed, 3 are about BP, 2 focusing on RST (Uzêda *et al*., 2010; Cardoso *et al*., 2011), and 1 just mentioning the theory (Maziero *et al*., 2010). It's noteworthy

that all of these papers were written in English and developed at the Interinstitutional Center for Computational Linguistics (NILC), whose head office is at the University of São Paulo (USP/São Carlos)[10].

In evaluating the entirety of the 760 works resulting from the search conducted within Publish or Perish software, it was observed that 52 studies are about BP. Although, this research did not undertake the listing of which among these investigations merely reference RST and which ones use it as a theoretical-methodological foundation. Nevertheless, it is noteworthy that a prevailing number of these works were produced through the collaborative efforts of professor-researchers Thiago Alexandre Salgueiro Pardo (USP/São Carlos) and Juliano Desiderato Antonio (State University of Maringá - UEM).

Collectively, there are 16 works authored by Thiago Alexandre Salgueiro Pardo, all of which are associated with projects focused on NLP. These research endeavors have contributed, to varying degrees, to the development of DiZer (DIscourse analyZER for BRazilian Portuguese)[11], a discourse parser based on RST for BP. On the other hand, Juliano Desiderato Antonio is credited with 21 works, predominantly centered on descriptive language studies. These works encompass a broad spectrum of analyses, including the examination of discourse signals, elocution and rhetorical relations in texts of different registers and genres.

It's essential to acknowledge that several influential factors play a crucial role in the likelihood of scientific work receiving citations within the academic-scientific community. These factors include (i) the language and channels publication, (ii) the publication format (whether in conferences or journals) and (iii) the impact factor of scientific communication channels. While this research did not specifically address these variables, it is well-established that they directly influence the visibility and citations rates, or citation works in BP. As a prospective avenue for future research, it is imperative to conduct a systematic review of these BP studies. Such a review will enable the identification of works that genuinely employ RST as a theoretical and methodological foundation, as well as elucidate their primary themes and contributions. which ones actually use RST as a theoretical-methodological basis, their main themes and contributions.

[10] NILC, available at: https://sites.google.com/view/nilc-usp/. Accessed in December 2022.
[11] DiZer, available at: http://nilc.icmc.usp.br/dizer2/. Accessed in December 2022.

5 Final Remarks

The main goal of this paper was to provide an overview of studies based on RST in the last decade, with the organization and description of relevant works, in number of citations, in the area. Therefore, for didactic purposes, we proposed a classification of the works listed in three major areas: (i) theoretical and descriptive studies; (ii) *Corpus* linguistics; and (iii) NLP.

It is important to emphasize that the applied methodology, meant that most of the works that use the RST model for Portuguese language were not resumed (only 3 investigations appeared among the most cited works). In total, out of the 760 works listed by the Publish or Perish tool, the number of researches on Portuguese becomes 52, which represents 6.8% of the sample. The fact that the vast majority of these works were produced in Portuguese stands out, being a possible explanation for the fact of their low citation, despite the impact and contribution of research. Corroborating this reflection, there is also the issue that English was the language predominantly analyzed and/or processed in the most cited studies.

Taking the temporal dimension into account, a conspicuous trend has emerged in the adoption of RST as a foundational framework in the realm of NLP. We have discerned a paradigmatic shift commencing from 2015, which has propelled the state-of-the-art forward, particularly concerning textual methodologies and genres. This transition has witnessed the increasing popularity and robustness of deep learning models, including neural networks. This observation holds significant implications for prospective research endeavors, as it offers a promising avenue for surmounting the limitations associated with the sole reliance on discourse markers for the identification of RST relations.

From this systematic review, in addition to preparing an overview of research on RST in recent times, it was possible to list different points for future investigations, with a focus on BP, for the establishment of an agenda of work, namely:

a) We verified, as carried out in the works of Das and Taboada (2013; 2018), the possibility of revisiting research and notes in RST of BP, in order to consider other textual elements, in addition to the traditional discourse markers, for the determination of rhetorical relations. In addition, we foresee the diversification of textual genres for analysis - as in the works of Green (2010) and Peldszus and Stede (2013; 2016).

This expansion includes the incorporation of user-generated content, such as product reviews, tweets, comments, and more;

b) We highlight the importance of research endeavors focused on the segmentation, annotation, and comparative analysis of parallel and/or compatible corpora encompassing Portuguese language variants and other natural languages. This approach mirrors the methodology employed by Iruskieta, Da Cunha, and Taboada (2015) in their investigation involving Basque, Spanish, and English. We particularly emphasize our interest in descriptive and comparative studies that juxtapose the Portuguese and Spanish languages. This interest is informed by the intricate political-linguistic dynamics in South America and is aligned with the academic backgrounds and research interests of our project members;

c) We underscore our dedication to the segmentation and identification of units of meaning within the framework of Rhetorical Structure Theory (RST). This commitment is directed towards providing direct contributions to Natural Language Processing (NLP) applications, thus extending the research initiated by Cardoso (2014). The areas of focus include automatic summarization, a continuation of Cardoso's work, sentiment analysis as conducted by Zhou *et al*. (2011), Chenlo, Hogenboom, and Losada (2013), and the detection of fake news as highlighted by Rubin and Lukoianova (2015). These research directions align with the contemporary and highly pertinent themes within the field.

While acknowledging the potential for exploring various topics related to RST, a well-defined work schedule, which may be adopted by other research teams, appears to address the primary concerns of the field at this time. In our project, our immediate focus is on the examination of the intricacies and advancements pertaining to points (a) and (b). This entails conducting linguistic-oriented scientific research, involving the annotation and analysis of discourse markers (both explicit and implicit), and their comparative analysis with other natural languages.

Acknowledgment

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.

Authors' Contributions

All the authors contributed to the literature review in this study and participated in drafting the text.

References

ALAMI, N.; MEKNASSI, M.; NOUREDDINE, R.A.I.S. Automatic texts summarization: Current state of the art. *Journal of Asian Scientific Research,* [*s.l.*], v. 5, n. 1, p. 1-15, 2015. DOI: https://doi.org/10.18488/journal.2/2015.5.1/2.1.1.15

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 (Online), [*s.l.*], v. 33, p. 79-108, 2017. DOI: https://doi.org/10.1590/0102-445025798334674077

ASHER, N.; LASCARIDES, A. *Logics of conversation*. 1ªed. Cambridge/UK: Cambridge University Press, 2003.

CARDOSO, P. C. F. *Exploração de métodos de sumarização automática multidocumento com base em conhecimento semântico-discursivo*. 2014. 182p. Tese (Doutorado). Universidade de São Paulo. Instituto de Ciências Matemáticas e de Computação/Universidade de São Paulo. 2014.

CARLSON, L.; MARCU, D. Discourse tagging reference manual. *ISI Technical Report ISI-TR-545*, [*s.l.*.], v. 54, n. 2001, p. 56, 2001.

CERVO, A.L.; BERVIAN, P.A.; SILVA, R. *Metodologia científica*. 6. ed. São Paulo: Pearson Prentice Hall, 2007.

CONROY, N.K.; RUBIN, V.L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. *In*: *Proceedings of the association for information science and technology*, USA: John Wiley & Sons, Ltd, 2015. v. 52, n. 1, p. 1-4. DOI: https://doi.org/10.1002/pra2.2015.145052010082

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

GIL, A.C. *Como elaborar projetos de pesquisa*. São Paulo: Atlas, 2002.

HARZING, A.W. Publish or perish software. *Computer Software*, 2007.

HIRATA-VALE, F. B. M.; OLIVEIRA, T. P. Modelos e Métodos de Análise Funcionalista. *In*: GONÇALVES, A. V.; GÓIS, M. L. S. (orgs.). *Ciências da Linguagem*: O Fazer Científico -Volume 2. Campinas: Mercado de Letras, 2014.

MANN, W.C.; THOMPSON, S.A. *Rhetorical structure theory*: Description and construction of text structures. Netherlands: Springer, 1987.

MAZIERO, E.G.; PARDO, T.A.S.; DA CUNHA, I.; TORRES-MORENO, J.; SANJUAN, E. DiZer 2.0-an adaptable on-line discourse parser. *In*: *Proceedings of the III RST Meeting* (8th Brazilian Symposium in Information and Human Language Technology). Mato Grosso/Brazil, p. 50-57, 2011.

MOREIRA, P.S.C.; GUIMARÃES, A. J. R.; TSUNODA, D. F. Qual ferramenta bibliométrica escolher? Um estudo comparativo entre *softwares*. *P2P e Inovação*, [*s.l.*], v. 6, p. 140-158, 2020. DOI: https://doi.org/10.21721/p2p.2020v6n2.p140-158

OSHIKAWA, R.; QIAN, J.; WANG, W.Y. A survey on natural language processing for fake news detection. *In*: *Proceedings of the 12th Conference on Language Resources and Evaluation*. Marseille/France, 2020. p.6086-6093. DOI: https://doi.org/10.48550/arXiv.1811.00770

PARDO, T.A.S. *Métodos para análise discursiva automática*. 2005. 211f. Tese (Doutorado em Ciências de computação e Matemática computacional). - Instituto de Ciências Matemáticas e de Computação/ Universidade de São Paulo. 2005.

PRASAD, R.; DIBSH, N.; LEE, A.; MILTSAKAKI, E.; ROBALDO, L.; JOSHI, A.; WEBBER, B. The Penn Discourse TreeBank 2.0. *In*: *Proceedings of Language Resources and Evaluation*. Marrakech/ Morocco, 2008. p. 2961-2968.

TABOADA, M.; MANN, W.C. Applications of rhetorical structure theory. *Discourse studies,* [*s.l.*], v. 8, n. 4, p. 567-588, 2006.DOI: https://doi.org/10.1177/1461445606064833

ZUPIC, I.; ČATER, T. Bibliometric methods in management and organization. *Organizational research methods,* [*s.l.*], v. 18, n. 3, p. 429-472, 2015. DOI: https://doi.org/10.1177/1094428114562629

Appendix A – Studies on RST analyzed in the systematic review

AREA	REFERENCES
RST and Heoretical and Descriptive Studies	ABRAHAMSON, J.A.; RUBIN, V.L. Discourse structure differences in lay and professional health communication. *Journal of Documentation*, [*s.l.*], v. 68, n. 6, p. 826-851, 2012. DOI: https://doi.org/10.1108/00220411211277064
	ALAMI, N.; MEKNASSI, M.; NOUREDDINE, R.A.I.S. Automatic texts summarization: Current state of the art. *Journal of Asian Scientific Research*, *[s.l.*], v. 5, n. 1, p. 1-15, 2015. DOI: https://doi.org/10.18488/journal.2/2015.5.1/2.1.1.15
	CONROY, N.K.; RUBIN, V.L.; CHEN, Y. Automatic deception detection: Methods for finding fake news. *In*: P*roceedings of the association for information science and technology*, USA: John Wiley & Sons, Ltd, 2015. v. 52, n. 1, p. 1-4. DOI: https://doi.org/10.1002/pra2.2015.145052010082
	DA CUNHA, I.; IRUSKIETA, M. Comparing rhetorical structures in different languages: The influence of translation strategies. *Discourse Studies*, [*s.l.*], v. 12, n. 5, p. 563-598, 2010. DOI: https://doi.org/10.1177/1461445610371054
	DAS, D.; TABOADA, M.. Explicit and implicit coherence relations: A corpus study. *In*: *Proceedings of the 2013 annual conference of the Canadian Linguistic Association*. Victoria: University of Victoria, 2013
	GREEN, N.L. Representation of argumentation in text with rhetorical structure theory. *Argumentation*, [*s.l.*], v. 24, n. 2, p. 181-196, 2010. DOI: https://doi.org/10.1007/s10503-009-9169-4
	JASINSKAJA, K.; KARAGJOSOVA, E. Rhetorical relations. *In:* MATTHEWSON, L.; MEIER, C.; RULLMANN, H.; ZIMMERMANN, T.E. (Eds.), *The companion to semantics*. Oxford: Wiley. 2020.
	MATTHIESSEN, C.M. Register in the round: Registerial cartography. *Functional Linguistics*, [*s.l.*], v. 2, n. 1, p. 1-48, 2015. DOI: https://doi.org/10.1186/s40554-015-0015-8
	MATTHIESSEN, C.M.; TERUYA, K. Grammatical realizations of rhetorical relations in different registers. *Word*, [*s.l.*], v. 61, n. 3, p. 232-281, 2015. DOI: https://doi.org/10.1080/00437956.2015.1071963

	OSHIKAWA, R.; QIAN, J.; WANG, W.Y. A survey on natural language processing for fake news detection. *In*: *Proceedings of the 12th Conference on Language Resources and Evaluation*. Marseille/France, 2020. p.6086-6093. DOI: https://doi.org/10.48550/arXiv.1811.00770
	PELDSZUS, A.; STEDE, M. From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence*, [*s.l.*], v. 7, n. 1, p. 1-31, 2013. DOI: https://doi.org/10.4018/jcini.2013010101
	PELDSZUS, A.; STEDE, M. Rhetorical structure and argumentation structure in monologue text. *In*: P*roceedings of the Third Workshop on Argument Mining (ArgMining2016)*, Berlin/Germany, 2016. p. 103-112. DOI: https://doi.org/10.18653/v1/W16-2812
	TABOADA, M.; HABEL, C.. Rhetorical relations in multimodal documents. *Discourse studies*, [*s.l.*], v. 15, n. 1, p. 65-89, 2013. DOI: https://doi.org/10.1177/1461445612466468
	TAFT, M.; KACANAS, D., HUEN, W.; CHAN, R. An empirical demonstration of contrastive rhetoric: Preference for rhetorical structure depends on one›s first language. *Intercultural Pragmatics*, [*s.l.*], v.8, n.4, p.503-516, 2011. DOI: https://doi.org/10.1515/iprg.2011.023
Corpus Linguistics	CARDOSO, P.C.F.; MAZIERO, E.G.; JORGE, M.L.R.C.; SENO, E.M.R.; DI FELIPPO, A. RINO, L.H.N.; NUNES, M. G.V.; PARDO, T.A.S. CSTnews: A discourse-annotated corpus for single and multi-document summarization of news texts in Brazilian Portuguese. *In*: *Proceedings of the 3rd RST Brazilian Meeting*. Cuiabá/Brazil, 2011. p. 88-105.
	DA CUNHA, I.; TORRES-MORENO, J.; SIERRA, G. On the development of the RST Spanish Treebank. *In*: *Proceedings of the 5th Linguistic Annotation Workshop*. 2011. p. 1-10.
	DAS, D.; TABOADA, M. RST Signalling Corpus: A corpus of signals of coherence relations. *Language Resources and Evaluation*, [*s.l.*], v. 52, n. 1, p. 149-184, 2018. DOI: https://doi.org/10.1007/s10579-017-9383-x
	IRUSKIETA, M.; ARANZABE, M.J.; ILARRAZA, A. D.; GONZALEZ-DIOS, I.; LERSUNDI, M.; LACALLE, O. L. The RST Basque TreeBank: an online search interface to check rhetorical relations. *In*: *Proceedings of 4th workshop RST and discourse studies*. Fortaleza/Brazil, 2013. p. 40-49.

	IRUSKIETA, M.; DA CUNHA, I.; TABOADA, M. A qualitative comparison method for rhetorical structures: identifying different discourse structures in multilingual corpora. *Language resources and evaluation,* [*s.l.*], v. 49, n. 2, p. 263-309, 2015. DOI: https://doi.org/10.1007/s10579-014-9271-6
	SANDERS, T.J.M.; DEMBER, V.; KOEK, J.; SCHOLMAN, M.C.J.; ASR, F;.T.; ZUFFEREY, S.; EVERS-VERMEUL, J. Unifying dimensions in coherence relations: How various annotation frameworks are related. *Corpus Linguistics and Linguistic Theory*, [*s.l.*], v. 17, n. 1, p. 1-71, 2021. DOI: https://doi.org/10.1515/cllt-2016-0078
	STAB, C.; KIRSCHNER, C.; ECKLE-OHLER, J.; GUREVYCH, I. Argumentation mining in persuasive essays and scientific articles from the discourse structure perspective. *In*: *Proceedings of ArgNLP*. Forli-Cesena/Italy, 2014. p. 21-25.
	STEDE, M. AFANTENOS, S. PELDSZUS, A.; ASHER, N.; PERRET, J.Parallel discourse annotations on a corpus of short texts. *In*: *10th International Conference on Language Resources and Evaluation*. European Language Resources Association: Portorož/Slovenia, 2016. p. 1051-1058.
	VAN DER VLIET, N.; BERLÁNOVICH, I. BOUMA, .G.; EGG, M.; REDEKER, G. Building a discourse-annotated Dutch text corpus. *Bochumer Linguistische Arbeitsberichte*, [*s.l.*] v. 3, p. 157-171, 2011.
	ZELDES, A. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, [*s.l.*], v. 51, n. 3, p. 581-612, 2017. DOI: https://doi.org/ 10.1007/s10579-016-9343-x
	ZHONG, Y.; JIANG. XU, W.; LI, J.J. Discourse level factors for sentence deletion in text simplification. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*. New York/USA, 2020. v. 34, n.5, p.9709-9716. DOI: https://doi.org/10.48550/arXiv.1911.10384
NLP – Construction of Ontology	MITROVIĆ, J.; O'REILLY, C.; MLADENOVIĆ, M.; HANDSCHUH, S. Ontological representations of rhetorical figures for argument mining. *Argument & Computation*, [*s.l.*] v. 8, n. 3, p. 267-287, 2017. DOI: https://doi.org/10.3233/AAC-170027
NLP – Text Generation	KONSTAS, I.; LAPATA, M. Inducing document plans for concept-to-text generation. *In*: *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Seattle/USA, 2013. p. 1503-1514.

NLP – Automatic Summarization	AZMI, A.M.; AL-THANYYAN, S. A text summarizer for Arabic. *Computer Speech & Language*, [*s.l.*], v. 26, n. 4, p. 260-273, 2012. DOI: https://doi.org/10.1016/j.csl.2012.01.002
	HIRAO, T.; YOSHIDA, Y.; NISHINO, M.; YASUDA, N.; NAGATA, M. Single-document summarization as a tree knapsack problem. *In*: *Proceedings of the 2013 Conference on empirical methods in Natural Language Processing*. Association for Computational Linguistics: Seattle/USA, 2013. p.1515-1520.
	KIKUCHI, Y.; HIRAO, T.; OKUMURA, M.; NAGATA, M. Single document summarization based on nested tree structure. *In*: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Baltimore/USA, 2014. p. 315-320. DOI: https://doi.org/10.3115/v1/P14-2052
	LE, H.T.; LE, T.M. An approach to abstractive text summarization. *In*: *Proceedings of International Conference on Soft Computing and Pattern Recognition*. 2013. p. 371-376. DOI: https://doi.org/10.1109/SOCPAR.2013.7054161
	LI, J.J.; THADANI, K.; STENT, A. The role of discourse units in near-extractive summarization. *In*: *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics: Los Angeles/USA, 2016. p.137-147. DOI: https://doi.org/10.18653/v1/W16-3617
	LOUIS, A.; JOSHI, A.K.; NENKOVA, A. Discourse Indicators for Content Selection in Summarization. *In*: *Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics: Tokyo/Japan, 2010. p. 147–156.
	UZÊDA, V.R.; PARDO, T.A.S.; NUNES, M.G.V. A comprehensive comparative evaluation of RST-based summarization methods. ACM *Transactions on Speech and Language Processing (TSLP)*, [*s.l*], v. 6, n. 4, p. 1-20, 2010. DOI: https://doi.org/10.1145/1767756.1767757
NLP – Discourse Parsers	ANITA, R.; SUBALALITHA, C. N. Building discourse parser for Thirukkural. *In*: *Proceedings of the 16th International Conference on Natural Language Processing*. NLP Association of India: Hyderabad/India 2019. p. 18-25.

	FENG, V.W.; HIRST, G. Text-level discourse parsing with rich linguistic features. *In*: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics*. Jeju Island/Korea: ACL Anthology. 2012. p. 60-68
	HERNAULT, H.; PRENDINGER, H.; DU VERLE, D.A.; ISHIZUKA, M. HILDA: A discourse parser using support vector machine classification. *Dialogue & Discourse*, [*s.l.*], v. 1, n. 3, p. 1-33, 2010. DOI: https://doi.org/10.5087/dad.2010.003
	JOTY, S.; CARENINI, G.; NG, R.T. Codra: A novel discriminative framework for rhetorical analysis. *Computational Linguistics*, [s.l.], v. 41, n. 3, p. 385-435, 2015. DOI: https://doi.org/10.1162/COLI_a_00226
	LIN, X., JOTY, S., JWALAPURAM, P.; BARI, M.S. A unified linear-time framework for sentence-level discourse parsing. *In*: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics: Florence/Italy, 2019, p.4190–4200. DOI: https://doi.org/10.18653/v1/P19-1410
	MULLER, P.; AFANTENOS, S.; DENIS, P.; ASHER, N. Constrained decoding for text-level discourse parsing. *In*: *Proceedings of 24th International Conference on Computational Linguistics (COLING 2012)*. Chung-Li/Taiwan, 2012. p. 1883–1900.
	SURDEANU, M.; HICKS, T.; VALENZUELA-ESCARCEGA, M.A. Two practical rhetorical structure theory parsers. *In*: *Proceedings of the 2015 conference of the North American chapter of the association for computational linguistics*: Demonstrations. Association for Computational Linguistics: Denver/USA, 2015. p. 1-5. DOI: https://doi.org/10.3115/v1/N15-3001
NLP – Discourse Analysis	ALLEN, K.; CARENINI, G.; NG, R. Detecting disagreement in conversations using pseudo-monologic rhetorical structure. *In*: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Doha/Qatar, 2014. p. 1169-1180. DOI: https://doi.org/10.3115/v1/D14-1124
	BIRAN, O.; RAMBOW, O. Identifying justifications in written dialogs by classifying text as argumentative. *International Journal of Semantic Computing*, [*s.l.*], 2011b, v. 5, n. 04, p. 363-381. DOI: https://doi.org/10.1109/ICSC.2011.41

	BIRAN, O.; RAMBOW, O. Identifying justifications in written dialogs. *In*: *Proceedings of 2011 IEEE Fifth International Conference on Semantic Computing*. California/Estados Unidos, 2011a. p. 162-168. DOI: https://doi.org/10.1109/ICSC.2011.41
	BRAUD, Chloé; COAVOUX, Maximin; SØGAARD, Anders. Cross-lingual RST discourse parsing. *In*: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*. Valencia/Spain, 2017. p.292-304. DOI: https://doi.org/10.48550/arXiv.1701.02946
	CHAKRABARTY, T.; HIDEY, C.; MURESAN, S.; MCKEOWN, K.; HWANG, A. AMPERSAND: Argument Mining for PERSuAsive oNline Discussions. *In*: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, Hong Kong/China, 2019. p.2933-2943. DOI: https://doi.org/10.18653/v1/D19-1291
	FENG, V.W.; LIN, Z.; HIRST, G. The impact of deep hierarchical discourse structures in the evaluation of text coherence. *In*: *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*. Dublin/Irlanda, 2014. p. 940-949.
	GE, J.; HERRING, S. C. Communicative functions of emoji sequences on Sina Weibo. *First Monday*, [*s. l.*], v. 23, n. 11, 2018. DOI: https://doi.org/10.5210/fm.v23i11.9413
	HAYASHI, K.; HIRAO, T.; NAGATA, M.. Empirical comparison of dependency conversions for RST discourse trees. *In*: *Proceedings of the 17th annual meeting of the special interest group on discourse and dialogue*. Los Angeles/EUA, 2016. p. 128-136. DOI: https://doi.org/10.18653/v1/W16-3616
	JANSEN, P.; SURDEANU, M.; CLARK, P. Discourse complements lexical semantics for non-factoid answer reranking. *In*: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore/USA, 2014, v.1, p. 977-986. DOI: https://doi.org/10.3115/v1/P14-1092
	JI, Y.; SMITH, N.A. Neural Discourse Structure for Text Categorization. *In*: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver/Canada, 2017. p.996–1005. DOI: https://doi.org/10.18653/v1/P17-1092

	KATZ, S.; ALBACETE, P.L. A tutoring system that simulates the highly interactive nature of human tutoring. *Journal of Educational Psychology*, [*s.l.*], vol. 105, nº. 4, p.1126–1141, 2013.
	KOBAYASHI, N.; HIRAO, T.; KAMIGAITO, H.; OKUMURA, M.; NAGATA, M. Top-down RST parsing utilizing granularity levels in documents. *In*: *Proceedings of the AAAI Conference on Artificial Intelligence*. [*s.l.*], 2020. p. 8099-8106. DOI: https://doi.org/10.1609/aaai.v34i05.6321
	LI, J.; LI, R.; HOVY, E. Recursive deep models for discourse parsing. *In*: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*. Doha/Catar: Association for Computational Linguistics, 2014. p. 2061-2069. DOI: https://doi.org/10.3115/v1/D14-1220
	LI, J.; SUN, A.; JOTY, S.R. SegBot: A Generic Neural Text Segmentation Model with Pointer Network. *In*: P*roceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*. Stockholm, 2018. p. 4166-4172. DOI: https://doi.org/10.24963/ijcai.2018/579
	LI, Q.; LI, T.; CHANG, B. Discourse Parsing with Attention-based Hierarchical Neural Networks. *In*: *Proceedings of Conference on Empirical Methods in Natural Language Processing*. Austin/USA, 2016. p. 362-371. DOI: https://doi.org/10.18653/v1/D16-1035
	MOREY, M.; MULLER, P.; ASHER, N. A dependency perspective on rst discourse parsing and evaluation. *Computational Linguistics*, [*s.l.*], v. 44, n. 2, p. 197-235, 2018. DOI: https://doi.org/10.1162/COLI_a_00314
	MOREY, M.; MULLER, P.; ASHER, N. How much progress have we made on RST discourse parsing? A replication study of recent results on the RST-DT. *In*: *Proceedings of Conference on Empirical Methods on Natural Language Processing (EMNLP 2017)*. Copenhagen/Denmark, 2017. p. 330-1335. DOI: https://doi.org/10.18653/v1/D17-1136
	WANG, Y.; LI, S.; WANG, H. A two-stage parsing method for text-level discourse analysis. *In*: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Vancouver/Canada, 2017. p. 184-188. DOI: https://doi.org/10.18653/v1/P17-2029

NLP – Machine Translation	GUZMÁN, F.; JOTY, S.; MÀRQUEZ, L.; NAKOV, P. Using discourse structure improves machine translation evaluation. *In*: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Baltimore/USA, 2014. p. 687-698. DOI: https://doi.org/10.3115/v1/P14-1065
	JOTY, S.; GUZMÁN, F.; MÀRQUEZ, L.; NAKOV, P. DiscoTK: Using discourse structure for machine translation evaluation. *In*: *Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore/USA*, 2014. p. 402-408. DOI: https://doi.org/10.3115/v1/W14-3352
	TU, M.; ZHOU, Y.; ZONG, C. A novel translation framework based on rhetorical structure theory. *In*: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. Sofia/ Bulgaria, 2013. p. 370-374.
NLP – Sentiment Analysis	BHATIA, P.; JI, Y.; EISENSTEIN, J. Better document-level sentiment analysis from RST discourse parsing. *In*: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Lisbon/Portugal, 2015. p. 2212-2218. DOI: https://doi.org/10.48550/arXiv.1509.01599
	CHENLO, J.M.; HOGENBOOM, A.; LOSADA, D.E. Rhetorical structure theory for polarity estimation: An experimental study. *Data & Knowledge Engineering*, [*s.l.*], v.94, p.135-147, 2014. DOI: https://doi.org/10.1016/j.datak.2014.07.009
	HEERSCHOP, B.; GOOSSEN, F.; HOGENBOOM, A.; FRASINCAR, F.; KAYMAK, U.; DE JONG, F. Polarity analysis of texts using discourse structure. *In*: *Proceedings of the 20th ACM international conference on Information and knowledge management*. New York/USA,2011. p. 1061-1070. DOI: https://doi.org/10.1145/2063576.2063730
	HOGENBOOM, A.; FRASINCAR, F.; DE JONG, F., & KAYMAK, U. Using rhetorical structure in sentiment analysis. *Communications of the ACM*, [*s.l.*], v. 58, n. 7, p. 69-77, 2015. DOI: https://doi.org/10.1145/2699418
	KRAUS, M.; FEUERRIEGEL, S. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, [*s.l.*], v. 118, p. 65-79, 2019. DOI: https://doi.org/10.48550/arXiv.1704.05228

	ZHOUM, L; LI, B.; GAO, W.; WEI, Z.; WONG, K. Unsupervised discovery of discourse relations for eliminating intra-sentence polarity ambiguities. *In*: *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics: Edinburgh/UK, 2011. p.162-171.
NLP – Tools Annotation	HELFRICH, P.; RIEB, E.; ABRAMI, G.; LÜCKING, A.; MEHLER, A. TreeAnnotator: versatile visual annotation of hierarchical text relations. *In*: *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*. European Language Resources Association: Miyazaki/Japan, 2018. p.1958-1963.
	MARCU, D. *The theory and practice of discourse parsing and summarization*. MIT press: Cambridge, 2000.
	O'DONNELL, M. RSTTOOL 2.4-A markup tool for rhetorical structure theory. *In*: *INLG'2000 Proceedings of the First International Conference on Natural Language Generation*. Association for Computational Linguistics: Mitzpe Ramon/Israel, 2000. p. 253-256. DOI: https://doi.org/10.3115/1118253.1118290
	ZELDES, A. rstWeb-a browser-based annotation interface for Rhetorical Structure Theory and discourse relations. *In*: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: Demonstrations. Association for Computational Linguistics: San Diego/USA, 2016. p.1-5. DOI: https://doi.org/10.18653/v1/N16-3001
NLP – Fake News Detection	RUBIN, V.L.; VASHCHILKO, T. Identification of truth and deception in text: Application of vector space model to rhetorical structure theory. *In*: *Proceedings of the Workshop on Computational Approaches to Deception Detection*. Association for Computational Linguistics: Avignon/France, 2012. p. 97-106
	RUBIN, V.L.; CONROY, N.J.; CHEN, Y. Towards news verification: Deception detection methods for news discourse. *In*: P*roceedings of the Hawaii International Conference on System Sciences*. Kauai/ USA, 2015. p. 5-8. DOI: https://doi.org/10.13140/2.1.4822.8166
	RUBIN, V.L.; LUKOIANOVA, T. Truth and deception at the rhetorical structure level. *Journal of the Association for Information Science and Technology*, [*s.l.*], v. 66, n. 5, p. 905-917, 2015. DOI: https://doi.org/10.1002/asi.23216

Source: Prepared by the authors.