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.
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.

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. 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. 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.]

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 \textit{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 \textit{Contrast} relationship is observed even without the existence of an explicit discourse marker in the text is illustrated:

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{example_rst_relations}
\caption{Example of RST relations}
\end{figure}

\begin{itemize}
\item The heterotrophic theory states that the first living beings did not produce their own food.
\item because they did not have sufficient cellular machinery.
\item The autotrophic theory states that there was not enough food available, so the first living beings had to produce their own food.
\end{itemize}

As Antonio (2017) explains in Figure 1\(^4\), the Contrast relation is primarily due to the morphological markers of the morphemes \textit{hetero} and \textit{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 21\(^{\text{st}}\) 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.

We used two essential search criteria in the tool: search term (“Rhetorical Structure Theory”) 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
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 CSTNews7, 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.

---


in pedagogical practices. Das and Taboada (2018) present the RST Signaling Corpus, 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-

---

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)\textsuperscript{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)\textsuperscript{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.


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


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


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


## Appendix A – Studies on RST analyzed in the systematic review

<table>
<thead>
<tr>
<th>AREA</th>
<th>REFERENCES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author(s)</td>
<td>Title and Details</td>
</tr>
<tr>
<td>-----------</td>
<td>------------------</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
</tr>
</tbody>
</table>

**NLP – Construction of Ontology**


**NLP – Text Generation**


|--------------------------|-----------------------------------------------------------------------------------|


Source: Prepared by the authors.