A Comparative Study of Discourse Signal Proposals in Rhetorical Structure Theory

Gabriela Fortalesa[1], Jackson Wilke da Cruz Souza[2]

[1] Departamento de Letras e Artes - Universidade Estadual de Feira de Santana
Feira de Santana, Bahia / Brasil

[2]Programa de Pós-Graduação em Língua e Cultura - Universidade Federal da Bahia
Salvador, Bahia / Brasil

{gabrielafortalesa,jackcruzsouza}@gmail.com

Abstract. *This paper presents a comparative literature review on the Discourse Signals (DSs) of Rhetorical Structure Theory (RST) across six languages. The analysis demonstrates that coherence signaling is not universal, depending heavily on linguistic typology and textual modality. It becomes evident that typologically distant languages and new modalities require profound structural, syntactic, and hypertextual adaptations of the theory's original taxonomy.*

1. Introduction

Rhetorical Structure Theory (RST) (Mann and Thompson, 1988) has established itself as one of the most robust theoretical models for mapping textual organization and coherence. Furthermore, the theory has made significant contributions to a wide range of applications in the field of Natural Language Processing (NLP), especially in discourse processing (Souza, Cardoso and Rodrigues, 2024).

Historically, the identification of rhetorical relations and discourse signaling focused almost exclusively on explicit Discourse Markers (DMs), such as conjunctions and prepositions. This paradigm, however, was transformed by the foundational work of Das and Taboada (2018), who expanded the notion of rhetorical relation signaling by showing that coherence extends beyond traditional connectives, as illustrated in (1).

(1) John is tall. Mary is short.

In this example, the two sentences are implicitly related according to the traditional view, since there is no DM connecting them. However, it can be observed that the two propositions may be linked through the *Contrast* relation despite the absence of an explicit connection between them. In this case, more recent studies in the literature indicate that the signaling may occur through the antonymy between the lexical items “tall” and “short”. As a result, Das and Taboada (2018) formalized a comprehensive taxonomy of non-connective (DSs) encompassing morphological, syntactic, semantic, graphical, and genre-related features.

As pointed out by Liu and Zeldes (2023), RST trees and relations vary widely in the ways they are signaled. According to the authors, signaling depends heavily on text type and genre, as well as on cross-linguistic differences. However, it can be observed that the

taxonomic proposals for DSs presented in previous studies often overlap and/or differ from one another, sometimes only with regard to the label assigned to a given type of signal. In this context, a cross-sectional investigation of the existing studies could enrich current taxonomic proposals, especially with respect to Portuguese.

With this in mind, the present study aims to examine how discourse signals, across different languages, may be adapted to distinct genre and text-type contexts within the scope of RST. To this end, this paper is organized into four sections in addition to this Introduction. Section 2 presents the methodology adopted in this bibliographic study. Section 3 discusses the results and findings. Finally, Section 4 presents the concluding remarks and outlines directions for future research based on this study.

2. Methodology

This study was primarily based on a bibliographic and bibliometric review methodology (Soares, Picolli, and Casagrande, 2018), aiming to conduct a comparative literature review. The bibliographic survey was carried out using the Google Scholar and ACL Anthology databases. To this end, combinations of keywords (including their variations and translations) such as "Rhetorical Structure Theory," "RST," "discourse signals," "signalling," and "discourse markers" were used.

The preliminary screening was conducted through the reading of titles and abstracts, as suggested by Moreira, Guimarães, and Tsunoda (2020). Based on this process, studies presenting clear evidence of the formulation, adaptation, or adoption of discourse signal proposals were selected. Additionally, the cross-reference tracking technique (snowballing) (Ecar *et al.*, 2020) was employed, whereby references to other texts served as indicators for identifying further relevant studies. As a result, a corpus of 47 studies was compiled.

For the construction of the final comparative framework, only articles that adopted RST as their exclusive theoretical framework were retained. In this sense, studies that, although based on RST, operated through theoretical hybridization and/or multi-layer annotation were likewise excluded from the comparative scope, leaving a final total of 11 papers.

3. Results and discussion

Following the bibliographic review, the studies included relied on a wide variety of corpus genres. For English, Das and Taboada (2018) compiled a journalistic corpus, whereas Gessler, Liu, and Zeldes (2019) adopted a multi-genre corpus comprising social media posts, tutorials, news articles, and interviews. Regarding Brazilian Portuguese, Pardo (2005) analyzed a corpus of academic (scientific) texts, Antonio (2017) used spoken language transcripts, the study by Rodrigues, Souza and Cardoso (2023) relied on journalistic texts, and Gama and Souza (2024) employed data collected from social media. For Spanish and Basque, Da Cunha et al. (2011, 2012) built a multi-genre corpus including scientific, academic, and didactic texts, as well as news reports. In Chinese, Cao and Iruskieta (2018) compiled a multi-genre corpus consisting of academic texts, news articles, and advertisements, whereas Peng, Liu, and Zeldes (2022) expanded this approach by

incorporating academic texts, interviews, news articles, and tutorials. Seemann and Scheffler (2025) adopted an experimental corpus composed of constructed items and sentences for German, while Toldova *et al.* (2017) employed a multi-genre Russian corpus containing magazine, journalistic, and blog texts.

STUDIES	EXAMPLE
Das and Taboada (2018) - English	**"Although** L&T hadn't raised money for 38 year"
Gessler, Liu and Zeldes (2019) - English	"Three issues begin trading **today** [...] one began **last week**"
Pardo (2005) - Brazilian Portuguese	"Software Engineering is an evolving discipline [...] **with the primary objective** of producing high-quality, low-cost software."
Antonio (2017) - Brazilian Portuguese	(You defeat the Indians, keep the land for yourself)
Rodrigues, Souza and Cardoso. (2023) - Brazilian Portuguese	Defense Minister Nelson Jobim **announced** (...) **that** economist Solange Vieira…
Gama and Souza (2024) - Brazilian Portuguese	Let's see if even on a weak day, it gains some strength... **RT** @Live_Trade: #ecor3 closing 15' above 12.53 is already a good sign
Da Cunha *et al.* (2011; 2012) - Spanish	"[Los niños aprenden matemáticas] [**al** resolver problemas.]" (Children learn mathematics **by** solving problems)
Cao and Iruskieta (2018) - Chinese	**Jǐnguǎn** hái méiyǒu qǔdé zuìzhōng jiéguǒ, **dànshì** wǒmen rènwéi gāi móxíng"(**Although** the final results have not yet been obtained, we believe that the model…)
Seemann and Scheffler (2025) - German	"**Nachdem** er **ja** immer gesagt hatte, ich konnte ihn jederzeit anrufen, habe ich das gestern auch gemacht." ('**Because** he always said I could call him anytime, I did just that yesterday ')
Toldova *et al.* (2017) - Russian	There is no example of a complete sentence in the article, only marker-relation pairs, such as "**однако/а/но**" ('however/but'), "который" ('which'); "поэтому" ('so/therefore')

Table 1. Some illustrative examples drawn from the studies.

From Table 1, all the examples were translated to English from original studies and or idioms. For Brazilian Portuguese (BP), the pioneering study by Pardo (2005) focused on the extraction of explicit DMs from academic texts in order to create the first automatic parser for the language. This was later expanded by Rodrigues, Souza and Cardoso (2023), who systematized the use of combined signals in journalistic texts. As a result, the most recent study focused on the mapping of implicit relations, local syntax, and adaptation to the digital environment by identifying cues such as markers combined with verbal tenses. Still regarding BP, the study by Antonio (2017) illustrates that, when moving away from the standard of formal written texts, the signaling proposal needed to encompass interactional phenomena in spoken language (such as intonation and pauses) and to initiate a hypertextual paradigm for

User-Generated Content (UGC). Concerning the study by Gama and Souza (2024) on social media, for instance, it is noteworthy that elements inherent to the digital platform (such as links, emojis, and retweets) function as independent and domain-specific DSs.

With regard to Russian, the same principle of mapping non-connective items was adopted. This decision guided the formulation of the lexical inventory proposed by Toldova *et al.* (2017) for the structuring of discourse trees. In that study, non-connective items were mapped based on their co-occurrence, resulting in an inventory of content words (lemmas). As for the studies on English, Gessler, Liu, and Zeldes (2019) expanded the original model in order to anchor signals to the exact tokens in the text, including, for example, synonymy and visual layout features. Studies on German indicate that the disambiguation of rhetorical relations frequently occurs through modal aspects in colloquial sentences, making use of strong particles such as ja and doch, for instance.

With regard to languages that are typologically distant from one another, the studies presented more substantial innovations based on the proposal by Das and Taboada (2018), as observed in the work on Chinese. The authors highlighted the requirement for paired connectives and the high frequency of implicit relations, evidenced by simultaneous double markers (such as *jinguan* and *danshi*) and prenominal relative clauses.

Although most causal and concessive relations are marked by explicit lexical connectives across all the languages analyzed, the survey reveals that discourse signaling is far from uniform. First, there is typological variation in the amount of marking required for the same relation: Spanish and Portuguese rely on a single concessive connective, whereas Chinese requires an obligatory correlative pair. Second, not every signal is lexical: in spoken BP, clause order alone is sufficient to mark Condition; and in Chinese, the syntactic position of a relative clause itself triggers the same-unit relation. Third, DSs, such as modal particles in German and platform-specific abbreviations ("RT") in BP, show that the category of DS extends beyond the closed class of traditional connectives. Finally, marker ambiguity (Spanish "*al*", Portuguese "*que*", German "*doch*") recurs across typologically distant languages, suggesting that the functional polysemy of DSs is a general phenomenon rather than a language-specific one.

4. Final remarks

This study presented a bibliographic review and a comparative mapping of how discourse signals have been categorized and adapted across different languages and text genres from a strictly RST-based perspective. The aim was to demonstrate that coherence transcends the use of traditional markers, requiring expanded signaling frameworks capable of accounting for structures and phenomena related to syntax and to features associated with different modalities, such as spoken language and digital media.

It should be emphasized that the present study falls within the scope of research dedicated to discourse processing, particularly in BP. Recent findings demonstrate the relevance of RST as a theoretical framework, covering a wide range of languages and text genres, while also enabling notable applications in NLP.

As future work, we suggest expanding the scope of analysis to include hybrid or multi-layer theoretical models of discourse processing beyond RST. This approach will enable the exploration of larger corpora and additional language families, as well as deeper contrastive analyses of discourse signals. By explicitly comparing different taxonomies, future research can assess the impact of their incorporation or adaptation in studies on BP. Finally, as this study did not aim to propose a taxonomy for DSs in BP, future work following our exploratory and comparative analyses will focus on developing a data-driven taxonomy for BP.

Acknowledgements
The project was 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. In addition, this work received financial support from the FAPESB/CNPq Call for Proposals 004/2023 - First Projects Program.

References

Antonio, J. D. (2017). 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*, 33(1), p. 79-108. DOI: 10.1590/0102-445025798334674077

Cao, S., da Cunha, I., & Iruskieta, M. (2018). The RST Spanish-Chinese treebank. In: *Proceedings of the Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, p. 156-166.

Da Cunha, I., Torres-Moreno, J. M., & Sierra, G. (2011). On the development of the RST Spanish Treebank. In: *Proceedings of the 5th Linguistic Annotation Workshop*, p. 1-10.

Da Cunha, I., et al. (2012). DiSeg 1.0: The first system for Spanish discourse segmentation. *Expert Systems with Applications*, 39(2), p. 1671-1678. DOI: 10.1016/j.eswa.2011.06.058

Dantas, E., Santa Bárbara, L., Pereira, M., Gama, N., Almeida, T., Souza, J. W. C., Cardoso, P. C. F., & Rodrigues, R. (2024). *Manual de anotação de sinalizadores discursivos em textos jornalísticos*. Relatório Técnico nº 447, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo.

Das, D., & Taboada, M. (2018). RST Signalling *Corpus*: A corpus of signals of coherence relations. *Language Resources and Evaluation*, 52(1), p. 149-184. DOI: 10.1007/s10579-017-9383-x

Ecar, M. *et al.* (2020) Software process improvement diagnostic: A snowballing systematic literature review. *In: 2020 XLVI Latin American Computing Conference (CLEI)*. IEEE. p. 156-164. DOI: 10.1109/CLEI52000.2020.00025

Gama, N. S., & Souza, J. W. C. (2024). *Estudo preliminar sobre sinalizadores discursivos para Conteúdo Gerado por Usuários*. Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. DOI: 10.5753/stil.2024.245060.

Gessler, L., Liu, Y. J., & Zeldes, A. (2019). A discourse signal annotation system for RST trees. In: *Proceedings of the Workshop on Discourse Relation Parsing and Treebanking 2019*, p. 56-61. DOI: 10.18653/v1/W19-2708

Liu, Y. J., & Zeldes, A. (2023). Why Can't Discourse Parsing Generalize? A Thorough Investigation of the Impact of Data Diversity. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, p. 3112-3130. DOI: 10.48550/arXiv.2302.06488

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3), p. 243-281.

Moreira , P.S.C.; Guimarães, A. J. R. & Tsunoda, D. F. (2020). Qual ferramenta bibliométrica escolher? Um estudo comparativo entre softwares. P2P e Inovação, [s.l.], v. 6, p. 140-158. DOI: 10.21721/p2p.2020v6n2.p140-158

Pardo, T. A. S. (2005). *Métodos para análise discursiva automática*. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo. DOI: 10.11606/T.55.2005.tde-29082005-172336

Peng, S., Liu, Y. J., & Zeldes, A. (2022). GCDT: A Chinese RST Treebank for Multigenre and Multilingual Discourse Parsing. In: *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, p. 382-391. DOI: 10.18653/v1/2022.aacl-short.47

Rodrigues, R., Souza, J. W. C., & Cardoso, P. C. F. (2023). Sinalizadores retórico-discursivos: revisitando a anotação RST no córpus CSTNews. In: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, SBC, p. 249-257. DOI: 10.5753/stil.2023.234120

Seemann, H. J., & Scheffler, T. (2025). German modal particles as discourse Signals. *Dialogue & Discourse*, v. 16, n. 1, p. 1-30. DOI: 10.5210/dad.2025.101

Soares, S. V., Picolli, I. R. A. & Casagrande, J. L (2018). Pesquisa bibliográfica, pesquisa bibliométrica, artigo de revisão e ensaio teórico em administração e contabilidade. *Administração: ensino e pesquisa*, v. 19, n. 2, p. 308-339. DOI: 10.13058/raep.2018.v19n2.970

Souza, J. W. C., Cardoso, P. C. F., & Rodrigues, R. (2024). Systematic Review of Studies on Rhetorical Structure Theory. *Revista de Estudos da Linguagem*, 31(3), p. 1643–1675. DOI: 10.17851/2237-2083.31.3.1643-1675

Toldova, S., *et al.* (2017). Rhetorical relation markers in Russian RST Treebank. In: *Proceedings of the 6th Workshop on Recent Advances in RST and Related Formalisms*, p. 29-33. DOI: 10.18653/v1/W17-3604