CRPC-DB A discourse bank for Portuguese

Amália Mendes*[0000−0001−6815−2674] and Pierre Lejeune[0000−0002−2495−9043]

Center of Linguistics, School of Arts and Humanities, University of Lisbon, Lisbon, Portugal
{amaliamendes,lejeune}@letras.ulisboa.pt

Abstract. We present a new resource for discourse studies in Portuguese, the CRPC Discourse Bank (CRPC-DB). CRPC-DB follows the Penn Discourse Treebank style of annotation. The annotation is performed on the PAROLE corpus, a free subset of the Reference Corpus of Contemporary Portuguese (CRPC) that includes news, fiction and didactic/scientific texts. The discourse bank covers explicit and implicit relations at intra and inter-sentential levels, and includes for now a total of 14,436 discourse relations. We present the main guidelines of our annotation and discuss specific cases. An experiment in inter-annotator agreement was performed and holds results of 0.88 F1-score for discourse relation identification, 0.71 Cohen's K for the classification of discourse relation types, and 0,75 for top-level sense classification. The CRPC-DB will be distributed free of charge through the PORTULAN CLARIN infrastructure.

Keywords: Discourse bank · discourse relations · text coherence · PDTB-style of annotation

1 Introduction

We introduce the CRPC-DB, a Discourse Bank for Portuguese annotated according to the Penn Discourse Treebank (PDTB) scheme [22]. The corpus is labeled for discourse relations (also referred to as rhetorical relations or coherence relations), such as cause and condition, that hold between two spans of text and contribute to ensure the overall cohesion and coherence of the text. The scheme follows the principles of the PDTB annotation proposal and includes the updates of the PDTB 3.0 version [29]. The annotation is applied over the PAROLE corpus, a written subset of the Reference Corpus of Contemporary Portuguese (CRPC) [13] available on the ELRA catalogue [1]. The CRPC-DB, as a new resource for Portuguese in the PDTB framework, can be easily compared with similar projects for other languages, as well as compared with resources in other frameworks. It is also a source of linguistic insight into discourse relations and discourse connectives, and has immediate applications for discourse parsing, as well as texts related tasks, such as summarization, argumentation mining

* Author to whom correspondence should be addressed.

[1] http://catalogue.elra.info/en-us/repository/browse/ELRA-W0024$_0$1

and identification of complexity levels. In section 2 we revise work on discourse banks in several languages and different frameworks, and we specifically address resources that have been developed for the Portuguese language. We introduce the contents of the corpus in section 3.1, the annotation scheme in section 3.2 and the annotation process in 3.3. The results of an inter-annotator agreement experiment are presented in section 4 and we conclude in section 5.

2 Related work

As semantics, pragmatics and discourse are increasingly the focus of linguistics and NLP, several discourse banks marking coherence relations have been created for different languages and in different discourse frameworks such as Rhetorical Structure Theory or RST [12], Segmented Discourse Representation Theory or SDRT [5], the Penn Discourse Treebank or PDTB [22], and the Cognitive approach to Coherence Relations or CCR [25]. The model of the PDTB has been applied to English [22] and used with many other languages, such as Arabic [1], Chinese [33], Hindi [20], Italian [28], Tamil [24], and Turkish [32]. Some of these discourse banks cover all or part of the components of the PDTB scheme, and some adaptations have been made to accommodate specific linguistic properties of certain languages [23], but the core of discourse types is quite stable and makes it possible to use PDTB as a source of contrastive studies. The PDTB style of annotation has been applied to Portuguese to a small sample of TED Talks in the TED multilingual Discourse Bank - TED-MDB [31]. This multilingual and parallel discourse bank includes 6 talks that were annotated with explicit intra and inter-sentential relations, and with explicit inter-sentential relations. It follows the PDTB 2.0 scheme in terms of discourse relations types (Explicit, Implicit, AltLex, EntRel and NoRel) but adopts the PDTB 3.0 sense hierarchy [29]. To deal with the specific nature of the TED Talks transcripts, a new top-level sense named Hypophora was added to the hierarchy, in order to annotate contexts where the speaker asks a question and answers it himself to appeal to the public. Other discourse annotation efforts have produced resources for Brazilian Portuguese in the RST framework: corpora annotated with discourse information (CSTNews [3], CorpusTCC [19], Rhetalho [21], Summ-it [11]) and discourse parsing tools (RST Toolkit, DiZer, CSTParser) [2,17]. However, the number of resources for discourse studies is still scarce for Portuguese, especially European Portuguese, and are very much needed for the development of parsing tools.

3 The CRPC-DB

3.1 Raw corpus and pre-processing

The corpus is composed of written texts from different genres: newspapers, fiction and didactic / scientific texts taken from the PAROLE corpus, a subset of the CRPC [13]. The texts were tokenized using the LX-tokenizer which separates punctuation marks from words, detects sentence boundaries and deals with contracted forms and clitics in Portuguese [10]. The annotation consists of marking the connectives and the arguments of the connective, and consequently text tokenization is required prior to the annotation, in order to isolate connectives that are contracted with the following article or pronoun (e.g., "ao contrário de_ o"). Newspaper articles are usually short and were kept in full, but long texts from the other two genres were reduced to a maximum size of around 10.000 words. The discourse banks contains 65 texts and a total of 85.510 tokens. More information on the corpus is provided in Table 1. The corpus is not balanced in terms of text types. Newspapers is the dominant text type, while fiction is only a small part of our data. This follows from the fact that the PAROLE corpus, and the total CRPC, themselves are not balanced. In the Parole corpus, fiction texts are fewer but much longer. Our decision to select only a sample of the fiction texts (to prevent the inclusion of texts of very different lengths) is also the reason why fiction is underrepresented. This could be mitigated in future versions to provide data for contrastive studies of discourse relations in different text types.

Table 1. Number of files, words and relations per text type in the CRPC-DB

genre	no. of files	no. of tokens	no. of relations
newspaper	308	177,457	11,232
didactic/scientific	4	38,566	2,452
fiction	3	8,255	752
Total	315	224,278	14,436

3.2 Annotation scheme

The CRPC-DB is annotated according to the PDTB scheme: we consider that discourse relations are relations that ensure coherence and hold between two arguments that have properties of abstract objects [6], such as eventualities. As a result, we annotate verbal predicates but also nominalizations that are part of a discourse relation. The PDTB-style of annotation follows a lexicalist approach, as each discourse relation is marked by a connective. This connective is either explicit in the context, or the sense is inferred and a connective is supplied by the annotator. Contrary to RST, the two arguments of a relation are not distinguished in a structure Nucleus-Satellite. The decision as to which is argument 1 and which is argument 2 follows from the lexicalist approach of PDTB: the second argument is the one introduced by the connective. The annotation of the CRPC-DB applies at intra and inter-sentential levels and uses the relation types of the PDTB 3.0 (Explicit, Implicit, Alternative Lexicalization (AltLex), Alternative LexicalizationC (AltLexC), Entity Relation and No Relation. The only exception is the new relation type Hypophora, which is not considered in our scheme (see treatment of Question-Answer pairs in this section). We follow the sense hierarchy of the PDTB 3.0 version [29], extended with additional senses that will be discussed in this section.

The relation is considered Explicit when there is an overt connective that denotes the meaning of the relation, as in example 1 hereunder. For readability, we present the examples as non-tokenized text. In all examples, we underline the connective and render arg1 in italics and arg2 in bold. Connectives include (single or multi word) conjunctions, prepositions and adverbs, and also parallel connectives, i.e., pairs of connectives which are discontinuous and function as a single connective unit (e.g. *não só... mas também* 'not only... but also'). The discourse relation may also be lexically expressed by elements that do not fall into the category of connectives. These are alternative lexicalizations (AltLex) such as "the reason for this is that", "an example is" (example 2). Another type of relation is expressed by lexico-syntactic constructions (AltLexC) that signal specific coherence relations, such as the inversion of the auxiliary expressing condition, or constructions or "so (Adj/Adv) that" expressing result (example 3). When no connective or alternative lexicalization is found, the relation is considered Implicit and the annotator has to supply a connective that could occur in that context (example 4). Entity Relations (EntRel) are used when an Entity is introduced in the first argument and the second argument provides additional information on that entity, frequently as a parenthetical segment in the flow of discourse. NoRel is applied when there is no visible relation between two sentences (typically cases of topic shift).Both EntRel and NoRel apply specifically at inter-sentential level, between sentences.

For each relation of the type Explicit, Implicit, AltLex and AltLexC, a sense is provided, out of the sense hierarchy of the PDTB 3.0. The set of senses is divided in 4 top-level senses: Temporal, Contingency, Comparison and Expansion, further subdivided in a two or three-level set of senses. For instance, one subsense of Contingency is Contingency:Cause:Reason, and one subsense of Expansion is Expansion:Conjunction. In cases of ambiguity, the annotator may label the relation with two senses. In the CRPC-DB, both explicit and implicit relations are annotated, at both intra-sentential (examples (1) and (2)) and inter-sentential levels (example (4)). Contrary to the PDTB, we do not mark attribution (information related to the source and degrees of factuality of the abstract objects) at this stage of our work.

1. *A situação poderá mesmo agravar-se*, pois **passados os primeiros dias de Janeiro não se vislumbram sensíveis alterações** [Explicit; Contingency:Cause+Belief:Reason+Belief (The situation may even get worse, since after the first days of January no significant changes are expected)

2. *No caso daqueles situados entre a Terra e o Sol - Mercúrio e Vénus - essas "laçadas", como em tempos se lhes chamava, envolvem o Sol*, razão por que **se avistam ora à esquerda ora à direita do Sol (...)** [AltLex; Contingency:Cause:Result] (In the case of those located between the Earth and the Sun - Mercury and Venus - those 'loops' - as they used to be called - circle

the Sun, reason why they are visible either on the left or on the right of the Sun)

3. *faz logo de início considerações* <u>*tão tão óbvias*</u>*,* **<u>que</u> parecem lugares comuns (...)** [AltLexC; Contingency:Cause:Result] ([He/she] makes from the very beginning considerations that are so obvious that they seem commonplaces

4. *Este ano, a Primavera chegou mais cedo.* [Implicit = <u>de facto</u> 'indeed'] **Estamos, em Março, a viver alegremente o clima de Maio.** [Implicit; Expansion:Specification: arg2-as-detail] (This year, spring came earlier. We are, in March, happily enjoying the climate of May.)

During the annotation, we apply several principles that define the extension of the arguments of a relation and the annotation of conjoined structures, noun phrases and relative clauses.

Extension of the arguments. The extension of the arguments follow the minimality principle: an argument contains the minimal and sufficient amount of information required for the interpretation of the relation. If there is another span of text related to the arguments, they may be annotated as supplementary information (Sup1 and Sup2, for Arg1 and Arg2 respectively). Except for EntRel and NoRel relations, the minimality principle allows the annotator to select parts of the sentences as arguments of the relation. Or instead, to select multiple sentences as an argument if such information is required, for instance, when arg2 expresses a summary of a previous set of sentences.

Conjoined elements. In cases of conjoined verbal phrases (VPs), only constituents not shared by both arg1 and arg2 are considered in the relation. For instance, in the example "os agricultores *olham para o céu* e *desesperam* (farmers look at the sky and despair), none of the arguments include the subject "the farmers" because it is shared by both arguments. VP coordination only applies to cases where both arg1 and arg2 include a verb. When the second argument is verbless, the coordinated spans are not annotated. For instance, in the sentence "Depois de amanhã, Viana Batista discutirá o problema com Alberto João Jardim e, no dia seguinte, com Mota Amaral" (after tomorrow, Viana Batista will discuss the problem with Alberto João Jardim and, the day after, with Mota Amaral) we don't consider that the span "and the day after" is an argument because it lacks the verb. An exception to the previous rule, and to our option to avoid interpreting contexts as involving elided linguistic material, are cases where each of the conjoined arguments has its own subject but arg2 lacks a verb. These cases are understood as a clause with an elided verb and are annotated. For instance, the sentence "Os anticiclones estão associados a condições de bom tempo e os sistemas depressionários ou frontais, à chuva" (Anticyclones are associated with good weather conditions and low-pressure or frontal systems with rain) is interpreted as equivalent to "and low-pressure or frontal systems [are associated] with rain".

Noun phrases. A noun phrase (NP) is annotated as an argument of a connective when: (i) there is an existential interpretation (e.g. Dada a grande diversidade de fontes sonoras, a resolução dos problemas (...) 'Given the great diversity of sound sources, the resolution of problems' is interpreted as "Given the [existence] of a great diversity"; (ii) when the head noun is a nominalization (e.g. "utilização" 'utilization').

Question-answer pairs. We encountered in our corpus several contexts containing questions. This is the case of newspaper articles, when transcribing an interview, but also of newspaper articles that inform about an event and include part of the declarations made by an intervening party. Furthermore, in fiction texts, it is frequent to find dialogues that include questions and responses. Other discourse banks had to deal with question-answer (QA) pairs in different types of data. Most include a specific set of senses to label those contexts. For instance, the STAC corpus, a corpus of situated multiparty dialogues [4,7] annotated in the style of the SDRT [5], uses labels such as Question-Answer Pair (QAP). The section of the Wall Street Journal that has been annotated in the RST framework [12] labels QA pairs with specific senses combined with the concepts of nucleus and satellite (e.g., Question-Answer-N). In the TED-MDB (a corpus of transcriptions of TED Talks), cases where the speaker asks a question and answers it are labelled with a new top-level sense Hypophora, the name of a pragmatic figure of speech with an appealing function [15,16]. Contrary to these perspectives, the PDTB 2.0 [22] doesn't treat QA pairs with any special sense but the annotation of QA pairs has been revised in the PDTB 3.0 [30]: a new relation type Hypophora is added.

The contexts of QAP found in the CRPC-DB can involve truly interactive contexts (interviews), with two speakers, and contexts with a single speaker, as in phatic contexts of hypophora, frequent for instance in textbooks: the author presents a question that does not, of course, constitute a true request for information, but rather constructs what could be the question of a "second virtual speaker" [14]. The question frequently establishes a break in the flow of discourse with a topic-comment function. Contrary to RDT, the PDTB doesn't include topic-commment relations. Also, in QA pairs there is a single proposition instead of two abstract objects required in a relation in the PDTB: the answer to a global question provides the truth value of the proposition and the answer to a partial question provides the value of the variable identified in the question. In the CRPC-DB, we annotate QA pairs as other sentence sequences in the discourse bank, similarly to the PDTB 2.0 approach. For example, there is an implicit relation in the QA pair: "Quais as razões deste facto? Vamos procurá-las através de um estudo pormenorizado de cada continente." (What are the reasons for this fact? We will try to find them through a detailed study of each continent.) The answer provides additional information and allows the development of a topic, so the meaning of Specification is assigned (Expansion:Level-of-detail:arg2-as-detail) (see [7]). But to be able to identify the QA contexts all QA pairs that are truly interactive are labelled with a new top-level sense QAP, and cases of hypophora are labelled with the subsense QAP-Hypophora. In cases of doubt as to whether there are one or two speakers, the annotation is conservative and the QAP tag is chosen. When there is no relationship between the question and the next segment (that is, when the next segment does not directly refer to the question), it is noted as NoRel. The diversity of contexts of enunciation and the different functions performed at the textual level by QA pairs is a natural area to explore the concept of Attribution and its future application within the CRPC-DB.

3.3 Annotation process

The corpus is manually annotated at the discourse level by one trained annotator, who follows the principles of the PDTB [23], and is then revised by an experienced annotator. After the discussion of remaining differences between the two annotators and a third experienced member of the team, the final annotation is adjudicated by the experienced annotator.

Contrary to the PDTB, where connectives were annotated one at a time throughout the corpus, here the annotator reads all the text and annotates all the relations that are found, without pre-annotation of lexical cues. This methods guarantees that the annotator is not conditioned to identify certain relations and ignore others. An assessment of 3 different workflow strategies is reported in [26]: they conclude that an approach that proceeds one text at the time (either by annotating the relations sequentially as they appear in the text or by annotating first explicit and then implicit relations in one text) performs better than the PDTB approach. We apply the full text approach and annotate all the relation types sequentially. However, in especially difficult texts, what proved useful was to annotate first intra-sententially and then inter-sententially, to deal with one level at a time. An annotation manual has been elaborated for the Portuguese discourse bank, and is followed by the annotators. The manual is frequently revised after the discussion of differences between annotators.

Results. The total number of discourse relations in the CRPC-DB is 14,436. There are 365 segments marked as NoRel, and 53 marked as EntRel. The remaining relations are Explicit, Implicit and Alternative Lexicalizations (AltLex and AltLexC), to which a sense is attributed. We provide information on the distribution of the relations per type and per top-level sense in Table 2. One interesting result, that can be compared with other discourse banks, is the fact that implicit relations are more frequent than explicit ones. For instance, in the parallel corpus TED-MDB Portuguese stands out due to the higher number of implicits, compared to other languages. Also, when comparing the aligned data of TED-MDB, the authors found that Portuguese showed a stronger tendency to implicitation (translating an explicit relation in the source language as an implicit one in the target language). A comparison with the PDTB shows a different pattern in English, where explicit relations are more frequent (18,459) than implicit ones (16,224). Nevertheless, in the Portuguese CRPC-DB only the top-level sense Expansion occurs more frequently as an implicit relation, suggesting that implicitness may be strongly linked to rhetorical senses.

Table 2. Frequency of discourse relations per text type and top-level sense in the CRPC-DB (Expl.=Explicit; AL=AltLex; ALC=AltLexC; Impl.=Implicit)

	News				Did./sc.				Fiction				Total
Sense	Expl.	AL	ALC	Impl.	Expl.	AL	ALC	Impl.	Expl.	AL	ALC	Impl.	
Temporal	702	47	0	211	77	7	1	29	86	0	0	67	1,227
Contingency	1,006	113	26	473	199	39	2	161	32	1	3	29	2,084
Comparison	892	18	6	403	171	4	0	75	30	0	0	28	1,627
Expansion	2,057	63	0	4,941	432	22	1	1115	178	0	0	286	9,095
Total	4,657	241	32	6,028	879	72	4	1,380	326	1	3	410	14,033

4 Inter-annotator agreement experiment

In order to check the consistency of the annotations in the CRPC-DB, we performed an inter-annotator agreement (IAA) experiment. In our experiments, we selected three texts from the CRPC-DB, which were coded by a second experienced annotator. We then use the data coded by the two raters to evaluate three aspects: identification of discourse units, classification of relations and classification of senses. For the identification of units, we calculate agreement on discourse relation spotting, i.e. whether or not the annotators identified a relation between the same discourse units. As in [31], we do not adopt a strict approach in terms of arguments spans. We only require a match between the selected connectives (for the Explicits and AltLexes), and a match of the end point of the first text span and the beginning of the second span point. Following [18], we computed results of 0.8 for precision, 0.86 for recall and 0.88 for F1 score. To perform the calculations, we consider as "correct" the annotations of the first annotator.

For the classification of relations, we measured agreement among the common annotations on the discourse relation type (whether or not the discourse relation identified in two sets of annotations is of the same type, e.g. Explicit, AltLex, etc.). We also measured agreement on the sense of the discourse relation, i. e., whether or not the discourse relation identified in two sets of annotations is of the same top-level sense of PDTB's relation hierarchy. We report observed agreement and Cohen's kappa in Table 3. Annotating discourse relations is a complex task as the annotator has to infer semantic and pragmatic values from the connective and the context, and has to be aware of relations that hold at intra and also at inter sentential levels. Taking into account these challenges,

Table 3. Agreement on classification of discourse relation and top-level sense

	Observed agreement	Cohen's k
Classification of discourse relations	0.83	0.71
Classification of top-level senses	0.84	0.75

we consider that the F1 score of 0.88 indicates a high similarity in terms of spotting discourse relations. For IAA values, similar to [31], we consider a kappa of 0.70 as a good standard [27] and table 3 shows that this level is reached for the classification of both discourse relations and senses, suggesting a consistent and reliable annotation in the CRPC-DB.

5 Final remarks

A survey of available language resources and tools for Portuguese pointed out that, while tagged and parsed corpora were available, few resources existed at the discourse level [8]. The results of this survey are still valid today, and the CRPC-DB addresses this shortage of data for discourse studies and applications in Portuguese, especially in what concerns European Portuguese by offering a corpus annotated with a set of 14,436 discourse relations. The CRPC-DB provides annotated data in a widely used format, the PDTB scheme, that enables contrastive linguistic studies of different languages in what concerns the nature of the connectives, the frequency of explicit and implicit relation types, and also the challenges that language properties impose on the annotation scheme.

We reported an experiment in inter-annotator agreement that provided good results, considering the challenging task of discourse annotation: we obtained 0.88 F1-score for discourse relation identification, 0.71 Cohen's K for the classification of discourse relation types, and 0,75 for top-level sense classification. In the future, we plan to address attribution, as a crucial part of discourse studies. Another important aspect will be to parse our corpus to cross-reference the coherence relations with the syntactic relations that hold between the arguments (e.g., the different syntactic patterns to express Cause: subordination, conjunction, juxtaposition). Our goal in preparing this new resource is two-fold: to make available real contexts annotated with discourse relations that provide data for the linguistic analysis of cohesion and coherence relations in Portuguese; and to provide training data for the development of automatic tagging systems of discourse relations. We believe it might prove equally useful for linguistics and NLP. The CRPC-DB will be distributed free of charge through the PORTULAN CLARIN infrastructure [2] [9].

6 Acknowledgements

This work was partially supported by PORTULAN CLARIN—Research Infrastructure for the Science and Technology of Language, funded by Lisboa2020, Alentejo2020 and FCT—Fundação para a Ciência e Tecnologia under the grant PINFRA/22117/2016, and by FCT under the project UIDP/00214/2020. Some of its developments were implemented in the scope of the COST Action TextLink – Structuring Discourse in Multilingual Europe. We wish to thank the anonymous reviewers for their helpful comments.

[2] https://portulanclarin.net

References

1. Al-Saif, A., Markert, K.: The Leeds Arabic Discourse Treebank: Annotating discourse connectives for Arabic. In: Proceedings of LREC'2010. pp. 2046–2053
2. Aleixo, P., Pardo, T.A.: Csttool: um parser multidocumento automático para o português do Brasil. In: Proceedings of the IV Workshop on MSc Dissertation and PhD Thesis in Artificial Intelligence – WTDIA. pp. 140–145 (2008)
3. Aleixo, P., Pardo, T.A.: CSTNews: Um córpus de textos jornalísticos anotados segundo a teoria discursiva multidocumento CST (cross-document structure theory). Tech. Rep. NILC-TR-08-05, Núcleo Interinstitucional de Lingüística Computacional NILC, Universidade de São Paulo (2008)
4. Asher, N., Hunter, J., Morey, M., Farah, B., Afantenos, S.: Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In: The Tenth International Conference on Language Resources and Evaluation (LREC 2016) (2016)
5. Asher, N., Lascarides, A.: The semantics and pragmatics of presupposition. Journal of Semantics **15**(2), 239—-299 (1988)
6. Asher, N.: Reference to Abstract Objects in Discourse. Kluwer, Dordrecht (1993)
7. Asher, N., Muller, P., Bras, M., Ho-Dac, L.M., Benamara, F., Afantenos, S., Vieu, L.: ANNODIS and related projects: Case studies on the annotation of discourse structure. In: Ide, N., Pustejovsky, J. (eds.) Handbook of Linguistic Annotation, pp. 1241–1264. Springer (2017)
8. Branco, A., Mendes, A., Pereira, S., Henriques, P., Pellegrini, T., Meinedo, H., Trancoso, I., Quaresma, P., Lima, V., Bacelar, F.: The Portuguese Language in the Digital Age / A Língua Portuguesa na Era Digital. Springer, Heidelberg (2012)
9. Branco, A., Mendes, A., Quaresma, P., Gomes, L., Silva, J., Teixeira, A.: Infrastructure for the science and technology of language PORTULAN CLARIN. In: LREC 2020 Worskhop IWLTP 2020 – 1st International Workshop on Language Technology Platforms. pp. 1–7. ELRA (2020)
10. Branco, A., Silva, J.: Contractions: breaking the tokenization-tagging circularity. In: Lectures Notes in Artificial Intelligence. pp. 167–170. Springer (2003)
11. Carbonel, T., Fuchs, J.T., Rino, L.: Anotação parcial de estruturas retóricas (RST) do Corpus Summ-it. Tech. Rep. NILC-TR-04-07, Núcleo Interinstitucional de Lingüística Computacional NILC, Universidade de São Paulo (2007)
12. Carlson, L., Marcu, D.: Discourse tagging reference manual. Tech. Rep. ISI-TR-545 (2001)
13. Généreux, M., Hendrickx, I., Mendes, A.: Introducing the reference corpus of contemporary portuguese on-line. In: Calzolari, N., Choukri, K., Declerck, T., Dogan, M.U., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S. (eds.) LREC'2012 – Eighth International Conference on Language Resources and Evaluation. pp. 2237–2244. European Language Resources Association (ELRA), Istanbul, Turkey (May 2012)
14. Grésillon, A., Lebrave, J.L.: Qui interroge qui et pourquoi? In: La langue au ras du texte, pp. 57–132. Presses Universitaires de Lille (1984)
15. Lanham, R.: A Handlist of Rhetorical Terms. University of California Press, Berkeley (1991)
16. Mayoral, J.A.: Figuras Retóricas. Editorial Sintesis, Madrid (1994)
17. Maziero, E., Pardo, T.A.: CSTParser - a multi-document discourse parser. In: Proceedings of the PROPOR 2012 Demonstration. pp. 1–3 (2012)
18. Mírovský, J., Mladová, L., Zikánová, Š.: Connective-based measuring of the inter-annotator agreement in the annotation of discourse in PDT. In: COLING 2010: Posters. pp. 775–781. Coling 2010 Organizing Committee, Beijing, China (Aug 2010), `https://www.aclweb.org/anthology/C10-2089`

19. Nunes, M.V., Pardo, T.A.: A construção de um corpus de textos científicos em português do Brasil e sua marcação retórica. Tech. Rep. NILC-TR-03-08, Núcleo Interinstitucional de Lingüística Computacional NILC, Universidade de São Paulo (2003)
20. Oza, U., Prasad, R., Kolachina, S., Sharma, D.M., Joshi, A.: The Hindi Discourse Relation Bank. In: Proc. of the 3rd Linguistic Annotation Workshop. pp. 158–161. Association for Computational Linguistics (2009)
21. Pardo, T., Seno, E.: Rhetalho: Um corpus de referência anotado retoricamente. In: Anais do V Encontro de Corpora (2005)
22. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L.: The penn discourse treebank 2.0. In: Proceedings of LREC'2008. pp. 2961–2968 (2008)
23. Prasad, R., Webber, B., Joshi, A.: Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. Computational Linguistics **40**(4), 921–950 (2014)
24. Rachakonda, R.T., Sharma, D.M.: Creating an annotated Tamil corpus as a discourse resource. In: Proceedings of the 5th Linguistic Annotation Workshop. pp. 119–123. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011), `https://www.aclweb.org/anthology/W11-0414`
25. Sanders, T., Spooren, W., Noordman, L.: Toward a taxonomy of coherence relations. Discourse Processes **15**, 1–35 (04 1992)
26. Sharma, H., Dakwale, P., Sharma, D.M., Prasad, R., Joshi, A.K.: Assessment of different workflow strategies for annotating discourse relations: A case study with HDRB. In: CICLing (1). Lecture Notes in Computer Science, vol. 7816, pp. 523–532. Springer (2013)
27. Spooren, W., Degand, L.: Coding coherence relations: Reliability and validity. Corpus Linguistics and Linguistic Theory **6**(2), 241–266 (2010)
28. Tonelli, S., Riccardi, G., Prasad, R., Joshi, A.: Annotation of discourse relations for conversational spoken dialogs. In: Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
29. Webber, B., Prasad, R., Lee, A., Joshi, A.: A discourse-annotated corpus of conjoined VPs. In: Proc. of the 10th Linguistics Annotation Workshop. pp. 22–31 (2016)
30. Webber, B., Prasad, R., Lee, A., Joshi, A.: The Penn Discourse Treebank 3.0 annotation manual. Tech. rep., Institute for Research in Cognitive Science (2019)
31. Zeyrek, D., Mendes, A., Grishina, Y., Kurfalı, M., Gibbon, S., Ogrodniczuk, M.: TED multilingual discourse bank (TED-MDB): a parallel corpus annotated in the PDTB style. Language Resources and Evaluation **54**, 587–613 (04 2020)
32. Zeyrek, D., Webber, B.L.: A discourse resource for turkish: Annotating discourse connectives in the metu corpus. In: IJCNLP. pp. 65–72 (2008)
33. Zhou, Y., Xue, N.: PDTB-style discourse annotation of Chinese text. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1. pp. 69–77. Association for Computational Linguistics (2012)