Using a discourse bank and a lexicon for the automatic identification of discourse connectives

Amália Mendes[0000−0001−6815−2674] and Iria del Río[0000−0002−4187−6485]

University of Lisbon, Center of Linguistics-CLUL, Portugal
amaliamendes@letras.ulisboa.pt
igayo@letras.ulisboa.pt

Abstract. We describe two new resources that have been prepared for European Portuguese and how they are used for discourse parsing: the Portuguese subpart of the TED-MDB corpus, a multilingual corpus of TED Talks that has been annotated in the PDTB style, and the Lexicon of Discourse Markers for Portuguese (LDM-PT). Both lexicon and corpus are used in a preliminary experiment for discourse connective identification in texts. This includes, in many cases, the difficult task of disambiguating between connective and non-connective uses. We annotated the PT-TED-MDB corpus with POS, lemma and syntactic constituency and focus on the 10 most frequent connectives in the corpus. The best approach considers word-form+POS+syntactic annotation and leads to 85% precision.

1 Introduction

While annotation levels such as POS, lemmatization, and syntactic relations have been consistently addressed for English and other languages with good results in terms of resource availability and tool development, work on the higher levels of text and discourse is still scarce, even for English. In the case of the Portuguese language, resources and tools for semantics and discourse are few, and are frequently only available for one variety of Portuguese, Brazilian or European Portuguese [3].

To be able to address discourse parsing, it is important to count on linguistically informed data that will provide the necessary input for the automatic identification of text spans explicitly connected by a discourse marker (DMs) or implicitly related through a discourse relation (also referred to as rhetorical sense), such as cause, justification, condition, elaboration, instantiation. In this paper, we describe two new resources that have been prepared for European Portuguese and how they are used for discourse parsing. One such resource is the Portuguese subpart of the TED-MDB corpus, a multilingual corpus of TED Talks that has been manually annotated in the PDTB style [18] with some adaptations required by the multilingual character of the corpus and by the specific genre (prepared speech) [34]. Another resource is the Lexicon of Discourse Markers for Portuguese (LDM-PT), that was compiled from grammars and corpora, and that provides information on a set of 222 DMs in European Portuguese.

With the goal of building an automatic system for discourse parsing in European Portuguese, we performed a first experiment focused on the automatic identification of discourse connectives. The LDM-PT lexicon provided the list of candidate connectives, and the discourse-annotated corpus was further labeled for POS and parsed with detailed syntactic categories. We then evaluated the level of ambiguity of the identification task and we investigated which linguistic information contributed more to the recognition of discourse connectives. These results shed light on the linguistic features that are especially helpful for the automatic identification task.

The paper is organized as follows: we review related resources on discourse and discourse processing, especially for Portuguese, in section 2. We present the TED-MDB corpus in 3 and the LDM-PT lexicon in 4, before addressing the automatic DM's identification in section 5.

2 Related work

Discourse parsing involves discourse connective identification (distinguishing between connective and non connective uses if required), the delimitation of the arguments of the discourse relation, and discourse relation labeling. As discourse connectives can frequently express several rhetorical relations, sense disambiguation is another required step.

There are several discourse-annotated corpora in different theoretical frameworks. The PDTB [18] style of annotation has been applied to other languages besides English, such as Turkish [33], Chinese [35], Czech [26], and applied to English and French speech data [6]. For Brazilian Portuguese, several corpora have been annotated in the RST and CST frameworks (CSTNews, CorpusTCC, Rhetalho, Summ-it) [1,14].

Lexicons of DMs are even rarer than discourse-annotated corpora. The German lexicon DiMLex [29] and the French lexicon LEXCONN [24] are two of the first initiatives. DIMLex includes 275 connectives and provides information on orthographic variants, non-connective readings, focus particle, syntactic category and, more recently, discourse relations [27]. LEXCONN describes 328 connectives and provides a syntactic category and the set of discourse relations that apply to each connective, based on SDRT. Both lexicons have inspired the development of recent lexicons for other languages, such as the Italian lexicon LiCO (173 connectives) [9]. For Spanish, the DPDE, an online dictionary of Spanish discourse markers with 210 entries, covers both written and spoken data and provides a definition, together with detailed information on each connective, such as register, prosody, formulae and comparable markers [4]. The dictionary provides a Portuguese semi-equivalent to the Spanish particles. Recently, the design of a Czech lexicon of DMs that exploits the Prague Dependency Treebank was presented in [16]. Several lexicons have been converted to the DIMLex format to integrate Connective-lex, a multilingual lexicon of discourse markers [31], [8].

Although all these lexical resources address discourse related devices, the type of unit that they contain can vary considerably. DIMLex, LEXCONN and most

of the lexicons cover discourse connectives, while the DPDE targets mainly discourse markers in speech with pragmatic and interactional meaning [7]. Even when focusing exclusively on discourse connectives, the lexicons may be restricted to the more typical categories (conjunctions, prepositions and adverbial phrases) or include a larger set of expressions that fulfill a cohesive function in a specific context. This occurs frequently in cases where the units were extracted from a discourse bank. For instance, the PDTB includes Alternative Lexicalizations [21], [22] and the Prague Discourse Treebank includes secondary connectives (and free connective phrases) [25], that fall outside the traditional categories associated to discourse connectives.

There are different approaches to discourse parsing, from rule-based methods [13] to machine learning techniques [19]. For English, [20] extracted explicit discourse connectives in the PDTB and disambiguated their senses. Other work in sense identification includes [11], as well as the CoNLL Shared Task (`http://www.cs.brandeis.edu/~clp/conll15st/`). Lopes [12] reports an experiment for fully automatic identification of multilingual lexica including Portuguese. For most languages other than English, work on discourse parsing is either scarce or non-existent. However consistent work has been developed in discourse processing for Brazilian Portuguese: the corpora annotated with discourse information have lead to manual and automatic discourse annotation in the RST and CST frameworks (RST Toolkit, DiZer, CSTParser) [1,14]. To our knowledge, no such resources exist for the European variety of Portuguese. Hence, our goal is to contribute with resources and tools for the development of state-of-the-art discourse parsers for this variety.

3 The Discourse Bank: TED-MDB

The TED-Multilingual Discourse Bank (TED-MDB) is a corpus of TED talks transcripts involving six languages (English, German, Polish, Portuguese, Russian and Turkish), annotated for discourse relations [34]. Two of the talks have been aligned and can be queried on the TextLink portal [1].

TED talks are prepared presentations delivered to a live audience. The transcripts are prepared according to the norms of written language (e.g., they include punctuation) and are translated to various languages by volunteers, and revised. An XML version of the transcripts in all languages is available at the WIT3 website [5]. The TED-MDB corpus contains 6 talks annotated in the PDTB style of annotation: discourse relations that are either explicitly marked by a discourse connective or that can be inferred from the context are labeled. These relations may hold at the inter-sentential or the intra-sentential level. In TED-MDB, both explicit and implicit relations are labeled at the inter-sentential level, while only explicit relations are annotated at the intra-sentential level.

The annotation of an explicit relation labels the discourse connective, its two arguments and its sense. TED-MDB follows the PDTB 3.0 relation hierarchy,

[1] http://ec2-18-219-79-53.us-east-2.compute.amazonaws.com:8000/ted_mdb/

which has 4 top-level senses (Expansion, Temporal, Contingency, Contrast) and their second- or in some cases third-level senses [32]. We give an example of an explicit inter-sentential (1) relation. The discourse connective is underlined, the first argument is rendered in italic, and the second argument in bold. An example of an implicit inter-sentential relation is given in (2): in this case, there is no overt connective and the annotation provides a connective that expresses the inferred relation. As in PDTB, TED-MDB considers non prototypical devices that assure coherence in the text. Such elements are labeled Alternative Lexicalizations and one such example is given in 3. The original English transcript is provided in the examples.

1. *Ela disse-me que algumas delas não correspondiam à sua marca, às suas expectativas.* Na verdade **uma das obras de tal modo não correspondia à sua marca, que ela tinha-a posto no lixo no seu estúdio.**. (She told me that a few didn't quite meet her own mark for what she wanted them to be. One of the works, in fact, so didn't meet her mark, she had set it out in the trash in her studio)[Expansion:Instantiation] (TED Talk no. 1978)
2. *esta companhia tem a visão direcionada para o que eles chamam de "o novo Novo Mundo".* (Implicit = porque) **São quatro mil milhões de pessoas da classe média que precisam de comida, de energia e de água.** (this company has their sights set on what they call "the new New World." That's four billion middle class people demanding food, energy and water.) [Contingency:Cause:Reason] (TED Talk no. 1927)
3. *muitos desses amputados do país não usavam as suas próteses.* A razão, como vim a saber mais tarde, era que **o encaixe das próteses era doloroso por não ser um encaixe perfeito**. (many of the amputees in the country would not use their prostheses. The reason, I would come to find out, was that their prosthetic sockets were painful because they did not fit well) [Contingency:Cause:Reason] (TED Talk no. 1971)

4 The Lexicon: LDM-PT

The Lexicon of Discourse Markers (LDM-PT) [15] provides a set of lexical items in Portuguese that have the function of structuring discourse and ensuring textual cohesion and coherence at intra-sentential and inter-sentential levels [10]. Each discourse marker (DM) is associated to the set of its rhetorical senses (also named discourse relations or coherence relations), following the PDTB 3.0 sense hierarchy (Webber et al., 2016).

Discourse connectives are taken in the lexicon as elements that do not vary regarding inflection, express a two-place semantic relation, have propositional arguments and are not integrated in the predicative structure. This includes conjunctions, adverbs and adverbial phrases, but also prepositions and alternative lexicalizations, as defined in the PDTB (see section 3). The DMs were taken from several sources: grammars; corpus-driven lists for the main POS, such as conjunctions and prepositions; manual contrastive approach between English

and Portuguese, based on the parallel Europarl corpus (the manual identification of connectives based on a contrastive language analysis calls attention to other lexical strategies that express coherence relations between text spans); and, mainly, the automatic extraction of the DMs that are labeled as connectives in the Portuguese part of the TED-MDB corpus.

As a result, the lexicon mainly reflects the decisions taken in the treebank in what concerns which rhetorical senses are associated with a connective. In the TED-MDB treebank, the intrinsic values of the DM are included, and values that may be triggered by adjacency between sentences and by the lexical content of the clauses are excluded. When the contexts leads to infer an additional sense, the explicit DM is labeled with its prototypical sense and an implicit relation is added to describe the sense that is inferred from the context, as in the PDTB [23]. One such example in TED-MDB is provided below: the explicit coordinate conjunction (underlined) is labeled with the sense Expansion:Conjunction (4) and an additional implicit DM (underlined and in parentheses) accounts for the inferred sense Contingency:Cause:Result (5).

4. *Estas iniciativas criam um ambiente de trabalho mais móvel* e **reduzem a nossa pegada imobiliária**. (TED talk 1927) (These initiatives create a more mobile work environment and reduce our housing footprint.)

5. *Estas iniciativas criam um ambiente de trabalho mais móvel e* (portanto) **reduzem a nossa pegada imobiliária**. (These initiatives create a more mobile work environment and consequently reduce our housing footprint.)

The lexicon includes both continuous (*porque* 'because', *então* 'then', *na verdade* 'in fact') and discontinuous units (*por um lado... por outro lado* 'on the one hand... on the other hand', *tal como... também* 'just as... so too'), and this information is part of the features of the XML structure. The typology is more detailed than the one found in the treebank: the connectives are divided in primary connectives, secondary connectives and alternative lexicalizations. The latter were described in 3. The distinction between primary and secondary connectives follows the proposal of [25]. Primary connectives are prototypical discourse connectives such as conjunctions, prepositions, adverbs and adverbial phrases. Secondary connectives are devices with a lesser degree of lexicalization, where one element (usually a deitic) is typically replaceable: *antes disso* 'before that', *da mesma maneira* 'in the same way', *nessa altura* 'at that time'.

The lexicon provides information on restrictions on the mood of the clause introduced by the DM and on its tense. For each discourse connective/sense pair, one or more English near-synonyms are listed. They are extracted, when applicable, from the DiMLex-en, compiled from data from the PDTB (Stede et al., 2017). Each entry of the lexicon provides a corpus example and information on the source of the example. Contrary to DIMLex, there is no feature in LDM-PT that identifies possible non connective uses of the DMs. The XML version of the lexicon was converted to the DIMLex format and is integrated in a multilingual resource [31] through a web app (at Connective-Lex.info) [8].

5 Automatic identification of connectives

5.1 The ambiguity of discourse connectives

Argument identification is the first step of discourse parsing and has a central role in building quality discourse representations [28]. We understand argument identification as in [11] that is, the identification of the different elements that compose a discourse relation (explicit or implicit and inter or intra- sentential): potential discourse markers and arguments.

In many cases, words that have a cohesive function in texts may also have non connective functions, that is, they are ambiguous [30]. As we mentioned in 4, the lexicon does not provide any information on those cases. For instance, the adverb *assim* 'in such a way' modifies the pronoun in (6) and does not perform a cohesive function at the discourse level. However, it is indeed a connective when connecting two sentences in (7) with a Result sense. Another very frequent case of ambiguity are coordinating conjunctions, that connect lower-level phrases such as nominal phrases (8) [2], or high-levels constituents, such as clauses and sentences (4). Only the latter cases are to be included in a discourse annotation task.

6. Isto tem que ser feito com grande precisão, mas se o conseguirmos, se conseguirmos construir esta tecnologia, se a colocarmos no espaço, poderão ver algo *assim*. (This has to be done very precisely, but if we can do this, if we can build this technology, if we can get it into space, you might see something like this.) TED Talk no. 1976
7. Eles acreditam que o ASG tem o potencial de criar impacto em riscos e receitas, *assim*, incorporar o ASG no processo de investimento é fundamental ao seu dever de agir no melhor interesse dos membros do fundo... (They believe that ESG has the potential to impact risks and returns, so incorporating it into the investment process is core to their duty to act in the best interest of fund members...) TED Talk no. 1927
8. As companhias *e* os investidores não são os únicos responsáveis pelo destino do planeta . (Companies and investors are not singularly responsible for the fate of the planet.) TED Talk no. 1927

5.2 Identification of connectives

To pursue the identification of connectives, we used a data-driven approach that exploits the information encoded in LDM-PT and in the Portuguese section of the TED-MDB corpus.

As a first step, we extracted all the explicit discourse relations in the corpus and we identified the explicit connectives with their sense (PDTB 3.0 sense

[2] Nominalizations (e.g., the destruction of the city) can be considered as equivalent to clauses and part of the discourse level, as in the PDTB (although few such cases are actually annotated), so that coordinating conjunctions connecting nominalizations would have to be identified as discourse connectives.

hierarchy). There are 275 instances of explicit connectives. These connectives correspond to 42 different word-forms with 886 cases in the corpus. Therefore, only a 31% of the possible candidates are effectively working as connectives in our data.

The ten most common connectives (by lemma) in the corpus are: *e* (and), *mas* (but), *para* (for/to), *se* (if), *quando* (when), *porque* (because), *depois* (after), *por* (for/because), *ou* (or), *então* (then). They account for 81% of the total cases (569 word-forms, 224 connectives, 345 non-connectives). Considering this fact and that we were performing a preliminary experiment, we restricted our analysis to these ten connectives.

Table 1. Distribution of word-forms, connectives and non-connectives in the corpus for the ten most common connectives.

Word-forms	**Connectives**	**NonConnectives**
569	224 - 39%	345 -61%

In our list of ten connectives, we have six conjunctions, two prepositions and two adverbs. It is interesting to note that conjunctions account for 69% of the total connectives in the corpus. In fact, a single conjunction, *e* (and), accounts for a 32% of the total occurrences of connectives in the corpus. However, only a 37% of the occurrences of the word *e* have a discourse connective function. All these aspects are relevant for testing the ambiguity of connectives.

As a second step, we automatically annotated the PT-TED-MDB corpus with lemma, POS and syntactic information. For POS and lemma, we used the Portuguese module of Freeling [17]. Freely available Portuguese parsers are scarce. We tested different options and we chose the constituency representation of the parser PALAVRAS [2] because its syntactic trees contain rich linguistic information[3]. To investigate the contribution of different linguistic features to the identification task, we first defined three levels of linguistic information: word-form of the connective; POS and lemma of the connective; word-form, POS, lemma and syntactic information involving the connective and its context. We then applied a rule-based method that makes use of these levels of linguistic information, and we measured precision (and, in some cases, recall) in the identification of connectives and non-connectives in the corpus. We describe our results in the following paragraphs.

(1) Word-form.

In this approach, we consider that each word-form that can be a connective is effectively working as a connective, and we measure precision for the identification of connectives and non-connectives. As expected, word-form is not enough to identify connectives accurately. Word-forms corresponding to the ten most

[3] Also, dependency analysis is not available in the upload interface of PALAVRAS.

common connectives are effectively connectives in a 39% of their occurrences in the corpus. That is: considering that any word that can be a connective is working as such, we obtain a precision of 39% in the identification of connectives and a 0% of precision in the identification of non-connectives (because all occurrences are considered connectives). For some connectives ambiguity is low. For example, *quando* (when) works as a connective in 94% of its occurrences. However, this connective represents only a 6% of the total use of connectives. In other cases, there is a higher level of ambiguity, as in the case of the most common connective *e* (and), mentioned above.

(2) Word-form + lemma + POS.

In this approach, we used the morphological information encoded in the LDM-PT corpus and the POS and lemma from Freeling to discriminate the connectives. Adding POS and lemma slightly improves precision: from 39% to 41% in the identification of connectives, and from 0% to 100% in the identification of non-connectives. Recall is 100% for connectives and 9% for non-connectives since, as in the previous approach, we consider most of the candidates as connectives. These results make sense considering the fact that connectives are words with low POS ambiguity. Indeed, we can see an improvement for word-forms with more than one POS (that are more or less equally frequent). This is the case of the connective *se* (if), which can be a conjunction (if) or a clitic pronoun.

(3) Word-form + lemma + POS + syntax.

In order to add syntactic information as a new layer, we used the constituency representation of the parser PALAVRAS (constraint grammar). This is the approach with the best performance. Using syntactic information, general precision increases to 85% for connectives and to a 99% in the identification of non-connectives, with a recall of 99% for connectives and a 89% for non-connectives. We experiment a slight decrease in recall for connectives and a high increase for non-connectives.

Syntactic information is especially relevant for connectives that can link different types of structures, like coordinated conjunctions. It is important to remember that the most common connective in the corpus is the copulative conjunction *e*, which accounts for 32% of all the connective cases. On the other hand, this conjunction is fairly common in the corpus, with 237 occurrences as word-form. Of these 237 occurrences, only 89 are connectives (37%) - the conjunction *e* works as a discourse connective when it links clauses (as in (4)) or sentences.

Using syntactic information from PALAVRAS' output, we can identify all the cases where *e* is linking clauses/sentences. Following this approach, we got an 89% of precision and a 100% of recall identifying the connective uses of this conjunction. Since conjunctions account for a 83.5% of the total connectives in the corpus, the use of syntactic information highly improves the results.

Connectives that are used in specific constructions could be identified with simpler approaches, like pattern matching. It is the case of the prepositions *por* (because/for) and *para* (for/to). Those connectives have a unique POS, and they work as connectives in a very specific construction: when they introduce

infinitive subordinated clauses (*para fazer isso* (to do so)). For these uses, it would be enough to identify the cases where the preposition is followed by an infinitive/adverb+infinitive. This simple approach, however, would not be enough for conjunctions like *e* (and) or *mas* (but), that can introduce multiple types of structures and which can be located far from the verb when they introduce clauses. Defining a clause with a surface pattern can be difficult and introduce a lot of errors.

6 Conclusion

We have presented work on discourse processing for Portuguese, based on LDM-PT, a new lexicon of DMs for Portuguese and on the Portuguese part of the multilingual treebank TED-MDB. Both resources account for a wide range of syntactic categories: conjunctions, prepositions, adverbs and adverbial phrases, but also alternative lexicalizations that carry a cohesive function in texts.

Both lexicon and corpus are used in a preliminary experiment for discourse connective identification in texts. This includes, in many cases, the difficult task of disambiguating between connective and non-connective uses. We annotated the PT-TED-MDB corpus with POS, lemma, using Freeling, and syntactic constituency using the PALAVRAS parser. We focus here on the 10 most frequent connectives in the corpus, and in some cases, also the most ambiguous ones between connective and non connective uses. We test the results of adding layers of annotation in our identification task. Using word-form+POS information only provides an increase in precision from 39 to 41, performing better only in cases where a word-form has more than one POS category. The approach that considers word-form+POS+syntactic annotation leads to 85% precision on the identification of connectives. Syntactic information for complex sentences, with coordinated or subordinated clauses, has a high impact in the identification of conjunctions working as connectives.

In the future, we plan to extend this approach to all the connectives in our corpus, experimenting also with a dependency representation. We want to explore the identification of connectives in nominalization structures, accounted for both in the PDTB and in the TED-MDB. Taking the discourse processing further will lead to the task of sense attribution for each discourse relation.

7 Acknowledgments

This work was partially supported by national funds through FCT - Fundação para a Ciência e a Tecnologia (under the project PEst-OE/LIN/UI0214/2013), and some of its developments were implemented in the scope of the COST Action TextLink – Structuring Discourse in Multilingual Europe3.

References

1. 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)
2. Bick, E.: The Parsing System "Palavras". Automatic Grammatical Analysis of Portuguese in a Constraint Grammar Framework. University of Arhus, Århus (2000)
3. 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)
4. Briz, S.P.B., Portolés, J.: Diccionario de partículas discursivas del español. www.dpde.es (2003)
5. Cettolo, M., Girardi, C., Federico, M.: Wit3: Web inventory of transcribed and translated talks. In: Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT). vol. 261, p. 268 (2012)
6. Crible, L.: Discourse markers and (dis)fluency across registers : a contrastive usage-based study in English and French. Ph.D. thesis, Louvain (2007)
7. Cuenca, M.J., Marín, M.J.: Co-occurrence of discourse markers in catalan and spanish oral narrative. Journal of Pragmatics **41**, 899–914 (2009)
8. Dombek, F.: Connective-lex.Info – A Web App for a Multilingual Connective Database. Bachelor thesis, Potsdam (2017)
9. Feltracco, A., Jezek, E., Magnini, B., Stede, M.: Lico: A lexicon of italian connectives. In: Proceedings of the 3rd Italian Conference on Computational Linguistics. Napoli, Italy (2016)
10. Halliday, M., Hasan, R.: Cohesion in English. Longman (1976)
11. Lin, Z., Ng, H.T., Kan, M.Y.: A PDTB-styled end-to-end discourse parser. Natural Language Engineering **20**(02), 151–184 (2014)
12. Lopes, A., Matos, D., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I., Mata, A.I.: Towards using machine translation techniques to induce multilingual lexica of discourse markers. http://arxiv.org/abs/1503.0914 (2015), accessed: 2016-01-15
13. Marcu, D.: The Theory and Practice of Discourse Parsing and Summarization. MIT Press, Cambridge, MA, USA (2000)
14. Maziero, E., Pardo, T.A.: Cstparser - a multi-document discourse parser. In: Proceedings of the PROPOR 2012 Demonstration. pp. 1–3 (2012)
15. Mendes, A., del Rio, I., Stede, M., Dombek, F.: A lexicon of discourse markers for portuguese – ldm-pt. In: Proceedings of LREC'18 (2018)
16. Mírovský, J., Synková, P., Rysová, M., Poláková, L.: Designing czedlex – a lexicon of czech discourse connectives. In: Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation (2016)
17. Padró, L., Stanilovsky, E.: Freeling 3.0: Towards wider multilinguality. In: Proceedings LREC'12 (2012)
18. PDTB Group: The Penn Discourse Treebank 2.0 annotation manual. Tech. rep., Institute for Research in Cognitive Science, University of Philadelphia (2008)
19. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp. 13–16. Association for Computational Linguistics, Stroudsburg, PA, USA
20. Pitler, E., Nenkova, A.: Using syntax to disambiguate explicit discourse connectives in text. In: Proceedings of the ACL-IJCNLP 2009 Conference Short Papers. pp. 13–16. Association for Computational Linguistics (2009)

21. Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A.K., Webber, B.L.: The penn discourse treebank 2.0. In: LREC (2008)
22. Prasad, R., Joshi, A., Webber, B.: Realization of discourse relations by other means: Alternative lexicalizations. In: Proceedings of the 23rd International Conference on Computational Linguistics: Posters. pp. 1023–1031. Association for Computational Linguistics (2010)
23. Rohde, H., Dickinson, A., Clark, C., Louis, A., Webber, B.: Recovering discourse relations: Varying influence of discourse adverbials. In: Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics. pp. 22–31 (2015)
24. Roze, C., Danlos, L., Muller, P.: LexConn: a French Lexicon of Discourse connectives. Revue Discours (2012)
25. Rysová, M., Rysová, K.: Secondary Connectives in the Prague Dependency Treebank. In: Hajičová, E., Nivre, J. (eds.) Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015). pp. 291–299. Uppsala, Sweden (2015)
26. Rysová, M., Synková, P., Mírovský, J., Hajičová, E., Nedoluzhko, A., Ocelák, R., Pergler, J., Poláková, L., Pavlíková, V., Zdeňková, J., Zikánová, Š.: Prague Discourse Treebank 2.0 (2016)
27. Scheffler, T., Stede, M.: Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In: et al., N.C. (ed.) Proceedings of LREC'16, year =
28. Soricut, R., Marcu, D.: Sentence level discourse parsing using syntactic and lexical information. In: Proceedings of NAACL'03 - Volume 1. pp. 149–156. Association for Computational Linguistics, Stroudsburg, PA, USA
29. Stede, M.: DiMLex: A lexical approach to discourse markers. In: Exploring the Lexicon - Theory and Computation. Edizioni dell'Orso, Alessandria (2002)
30. Stede, M.: Discourse Processing. Morgan & Claypool Publishers (2011)
31. Stede, M., S.T., Dombek, F.: Connective-lex.info. Potsdam University: http://connective-lex.info (2017)
32. 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)
33. Zeyrek, D., Demirşahin, I., Sevdik-Çallı, A., Çakıcı, R.: Turkish Discourse Bank: Porting a discourse annotation style to a morphologically rich language. Dialogue and Discourse **4**(2), 174–184 (2013)
34. Zeyrek, D., Mendes, A., Kurfalı, M.: Multilingual extension of pdtb-style annotation: The case of ted multilingual discourse bank. In: LREC (2018)
35. 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)