A Lexicon of Discourse Markers for Portuguese – LDM-PT

Amália Mendes[1], Iria del Rio[1], Manfred Stede[2], Felix Dombek[2]
[1] University of Lisbon, Centre of Linguistics, Portugal
[2] University of Potsdam, Germany
amaliamendes@letras.ulisboa.pt, igayo@gmail.com, stede@uni-potsdam.de, felix_dombek@hotmail.com

Abstract

We present LDM-PT, a lexicon of discourse markers for European Portuguese, composed of 252 pairs of discourse marker/rhetorical sense. The lexicon covers conjunctions, prepositions, adverbs, adverbial phrases and alternative lexicalizations with a connective function, as in the PDTB (Prasad et al., 2008; Prasad et al., 2010). For each discourse marker in the lexicon, there is information regarding its type, category, mood and tense restrictions over the sentence it introduces, rhetorical sense, following the PDTB 3.0 sense hierarchy (Webber et al., 2016), as well as a link to an English near-synonym and a corpus example. The lexicon is compiled in a single excel spread sheet that is later converted to an XML scheme compatible with the DiMLex format (Stede, 2002). We give a detailed description of the contents and format of the lexicon, and discuss possible applications of this resource for discourse studies and discourse processing tools for Portuguese.

Keywords: Discourse markers, Lexicon, Discourse treebank

1. Introduction

The Lexicon of Discourse Markers (LDM-PT) 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 (Halliday & Hasan, 1976). 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).

We consider that discourse connectives do not vary regarding inflection, they 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 we discuss in section 4.

Our immediate goal is to provide data for the annotation of discourse relations in a Portuguese discourse treebank, although a listing of DMs will certainly prove to be useful for applications dealing with tasks such as parsing, text processing and summarization of Portuguese.

We revisit in Section 2 other lexicons of DMs, their features and structure schemata; we discuss in Section 3 the acquisition of the DMs that we integrate in our lexicon and in Section 4 the information provided for each DM. In Section 5, we present the way this information is structured and the result in XML format, while we discuss in Section 6 the use of such a lexicon in discourse studies and its applications in the automatic processing of discourse. Finally, we provide some concluding remarks in sSection 7.

2. Related work

A lexicon of DMs may be restricted to discourse connectives, i.e., devices that assure cohesion at intra and inter-sentential levels (typically, conjunctions and adverbial phrases) or it can have a larger scope by also including pragmatic markers with interactional and modal meanings (Cuenca and Marín, 2009). Even under a more restrictive perspective, there are differences in the set of categories included in lexicons. The question is additionally related to the acquisition method: while a lexicon that is compiled manually and is informed mainly by grammars and dictionaries will be more restrictive in terms of the categories and items listed, a lexicon (semi-) automatically derived from a discourse treebank will typically include a larger set of devices that the annotators have found to fulfil a cohesive function. Example of such cases are the Alternative Lexicalizations included in the Penn Discourse Treebank (PDTB) (Prasad et al., 2008; Prasad et al., 2010) and the secondary connectives (and free connective phrases) in the Prague Discourse Treebank (Rysová and Rysová, 2015), that fall outside the traditional categories associated to discourse connectives.

There are few lexicons of DMs currently available, although recent initiatives are reported for several languages. The German lexicon DiMLex (Stede, 2002) includes 275 connectives and provides information on orthographic variants, non-connective readings, focus particle and syntactic category. The association of discourse relations to each connective in DiMLex is described in Scheffler and Stede (2016). The Italian lexicon LiCO contains 173 connectives and follows closely the DiMLex structure (Feltracco et al, 2016). For French, there is LEXCONN, a large lexicon with 328 connectives, with information on their syntactic category and their discourse relation, based on SDRT (Roze et al., 2012). The DPDE is an online dictionary of Spanish DMs with 210 entries in html format. The DMs are not labelled with a rhetorical sense, but a definition is provided, together with detailed information on each connective, such as register, prosody, formulae and comparable DMs (Briz et al, 2003). Recently, the design of a Czech lexicon of DMs that exploits the Prague Dependency Treebank was presented in Mírovský et al. (2016).

Lexical resources available for Portuguese deal essentially with content words and even those focusing on multi word expressions favour content expressions. However, the DPDE online does provide a Portuguese equivalent to the set of Spanish discourse particles, and an experiment in the fully automatic identification of multilingual lexica including Portuguese has been reported (Lopes et al., 2015). In this context, the LDM-PT lexicon provides a new resource for discourse studies in Portuguese.

3. The acquisition of DMs

The identification of DMs comes from several sources. First of all, we used a list of single and phrasal elements belonging to grammatical classes, such as conjunctions and prepositions, compiled during the preparatory work for the POS annotation of the Reference Corpus of Contemporary Portuguese (Généreux et al., 2012).

We also automatically identify the DMs that are labelled as connectives in the Portuguese part of the TED-MDB corpus (Zeyrek et al., 2018). The TED-Multilingual Discourse Bank, or TED-MDB, is a parallel corpus of English TED talks transcripts and their translations in 5 languages (German, Russian, Polish, Portuguese and Turkish). The transcripts are manually annotated at the discourse level following the goals and principles of PDTB (Prasad et al., 2014). For each language, trained or experienced annotators go through each transcribed talk and proceed sentence by sentence, by identifying the type of relation (e.g. explicit, implicit, AltLex), the sense (using PDTB 3.0 sense hierarchy) and the arguments. The annotations are then discussed in multilingual group meetings where all TED-MDB members are physically present, to check annotation consistency. We refer to Zeyrek et al. (2018) for a detailed account of the annotation process.

To populate the lexicon, we retrieve the list of explicit and implicit connectives and the alternative lexicalizations that were marked in the corpus. This data inform the type of DMs that we include. Indeed, deriving the lexicon entries from the corpus annotation work leads us to include categories that are less typical of DMs, as we discuss in section 4.

Furthermore, we conducted a manual contrastive approach between English and Portuguese, based on the parallel Europarl corpus and on the list of English connectives of the PDTB (Mendes and Lejeune, 2016). We located DMs in the English corpus and inspected the Portuguese sentences to identify the corresponding DM. We applied a manual approach with several goals in mind: to procure fully accurate data, to identify potential new senses of the Portuguese connectives, to spot semantic and pragmatic differences between DMs denoting the same sense. The approach is close to the Translation Spotting Technique (Cartoni et al., 2013), although our motivation is not to capture the different meanings of a given connective in the source language but to acquire a diversified set of connectives in Portuguese. The manual identification of connectives based on a contrastive language analysis brings our attention to other lexical strategies that express coherence relations between text spans.

4. Contents of the Lexicon

The lexicon is structured as pairs of DMs/rhetorical senses, so as to cover polysemous markers. The lexicon includes at the moment 252 pairs of DMs/rhetorical senses. A unique numerical identifier is attributed to each DM/rhetorical sense pair. Additionally, there is a Comment feature available to add any observation or open discussion regarding the DM.

Rhetorical sense

We adopt the PDTB 3.0 sense hierarchy with 4 top-level senses (Comparison, Contingency, Expansion, Temporal) and second- and in some cases third-level senses (Webber et al., 2016). For instance, the DM *de modo que* 'so' is labelled as Contingency:Cause:Result, while the DM *da mesma forma que* 'in the same way as' is labelled Comparison:Similarity. DMs have frequently more than one possible rhetorical sense. Working on a lexicon of DMs involves a tension between multiplying the rhetorical senses of a DM or keeping a limited set of what may be considered as the prototypical or core values of the connective. Again, the acquisition method informs the results: the annotators of a discourse treebank will frequently choose different rhetorical senses for a single connective according to the context and this will be reflected in a treebank-driven lexicon. In our case, many of the DMs that are included in the lexicon are acquired from our work on TED-MDB. Here, the method followed the proposal of the PDTB: when the contexts lead to infer an additional sense, the explicit DM is labelled with its prototypical sense and an implicit connective is proposed and annotated with the inferred sense (Rohde et al., 2015). One example of such annotation in the Portuguese section of the TED-MDB Treebank is provided in (1): the explicit coordinate conjunction (underlined) is labelled with the sense Expansion:Conjunction (cf. 1a) and an additional implicit DM (underlined and in parentheses) accounts for the inferred sense Contingency:Cause:Result (cf. 1b).

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

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

As a result, the lexicon reflects the decisions taken in the treebank: we describe the intrinsic values of the DM independently of values that may be triggered by adjacency between sentences and by the lexical content of the clauses. For future automatic applications, we aim to combine the information in the lexicon with the data in the treebank related to explicit DMs that have been complemented by an implicit connective to account for inferred senses.

Internal structure of the DM

Two complementary features, inspired by the information in the DiMLex lexicon, describe the internal structure of the DMs. On the one hand, each DM is defined as continuous or discontinuous. Examples of discontinuous DMs are *por um lado… por outro lado* 'on the one hand… on the other hand', *tal como… também* 'just as… so too'. Discontinuous DMs are described as having two orthographic segments.

On the other hand, DMs are described as composed of a single token or as a multiword unit (phrasal). In the case of discontinuous DMs, each orthographic segment is also described in terms of single or phrasal. For instance, the conjunction *logo* 'thus' is a DM with a single token, the conjunction *logo que* 'as soon as' is a phrasal continuous marker, and *tal como… também* 'just as…so too' is a discontinuous DM, where orthographic part 1 is phrasal (*tal como*) and orthographic part 2 is single (*também*).

Type

We adopt a three-category typology: primary connectives, secondary connectives and Alternative Lexicalizations. The distinction between primary and secondary connectives follows the proposal of Rysová and Rysová (2015). Primary connectives are prototypical discourse connectives such as conjunctions, prepositions, adverbs and adverbial phrases. Secondary connectives are other devices that assure cohesion but show a lesser degree of lexicalization than the prototypical discourse connectives. Instances of secondary connectives in the lexicon frequently involve one element that may be replaced, such as deitics:

- *antes disso* 'before that', *da mesma maneira* 'in the same way', *nessa altura* 'at that time', *nesse caso* 'in that case', *nesta perspetiva* 'in this perspective / accordingly', *nessa perspetiva* 'in that perspective / accordingly', *neste sentido* 'in this sense / accordingly', *razão pela qual* 'reason for which'*, motivo por que* 'motive for which' .

We also include in the lexicon Alternative Lexicalizations (AltLex), i.e, alternative expressions that denote a cohesive relation, following the PDTB typology (Prasad et al., 2010). What we mark as Alternative Lexicalizations are cases more or less equivalent to a third type in Rysová and Rysová's proposal, labelled 'free connective phrases', that differ from secondary connectives because they carry specific lexical content that restricts their use to a limited set of contexts. Examples of alternative lexicalizations:

- *não deixa de ser verdade que* 'it is nevertheless true that', *isto não significa que* 'but that doesn't mean that', *um dia depois* 'one day later'

We have also encountered borderline cases of intra-sentential discourse relations marked by a main causative verb (Danlos, 2006), such as *provocar* 'to provoke', obrigar 'to force', *reduzir* 'to reduce', which typically establish a causal coherence relation between two nominalizations (Lejeune et al., 2016).

While these alternative lexicalizations were with no doubt required to capture coherence relations in the annotations of texts, it was debated whether or not to include them in the lexicon, since they fall outside the obvious POS categories. We decided to include these expressions because they might prove useful for applications in automatic discourse relation identification and labelling. Their categorization in a specific category (AltLex) makes it possible to isolate and exclude them if required.

Category of the DM

Additional information on the category of the connective is provided in a required field *Category.* For the primary and secondary types, there are 4 categories:

- subordinate conjunction (csu),
- coordinate conjunction (cco),
- preposition (prep),
- and adverb and adverbial phrases (adv).

For the AltLex type, we use the categories above if applicable. In other cases, we give here information about the category of the semantic nucleus of the expression. For instance, *isto não* <u>*significa*</u> *que* 'but that doesn't mean that' (arg2-as-denier) is labelled as Category = verb. This allows us to quickly retrieve all verb based alternative expressions that assure coherence relations in texts.

Restrictions on the context

The lexicon provides information on restrictions on the mood of the clause introduced by the DM: we consider indicative as the default value and label as subjunctive otherwise. There is also information on the tense of the clause introduced by the DM: the default is a finite tense and we provide information if otherwise, such as infinitive, inflected infinitive and participle. The last two are illustrated in (2) and (3), respectively (we underline the DM and show the inflected infinitive form in italic).

(2) <u>Apesar de</u> não *terem* sido colegas, a amizade delas durava desde o tempo da Faculdade. (CRPC) 'Although they had not been colleagues at school, their friendship lasted since college time.'

(3) <u>Uma vez</u> *ultrapassada* a "fase de admissibilidade", o SEF emite uma autorização de residência válida por 60 dias e renovável por 30 até ser tomada uma decisão final. (CRPC) 'Once the 'phase of admissibility' is overcome, the SEF issues a green card valid for 60 days and renewable for 30 more days until a final decision is taken.'

Frequent modifiers of the DM, if any, are also indicated in the lexicon, although not consistently. One such case is the frequent presence of the adverb *muito* 'very' before the conjunction *embora* 'although': *muito embora*. These features might be especially important to deal with connectives that share a common rhetorical sense although they do not occur in the same contexts since "connectives are not always interchangeable and therefore cannot be treated as equivalents" (Cartoni et al., 2013).

English near-synonym

We provide one or more English near-synonyms for each DM/sense pair. We choose, when applicable, one of the entries of the DiMLex-en, compiled from data from the PDTB, and provide the unique identifier of the DM in the English lexicon (Stede et al., 2017).

Corpus Example

Finally, we provide for each entry of the lexicon a corpus example and information on the source of the example. Examples originate mostly from: (i) the Reference Corpus of Contemporary Portuguese, available through CQPweb[1] (Généreux et al., 2012); (ii) from the Portuguese subpart of the TED-MDB discourse treebank (in the case of alternative lexicalizations, because we tend to provide examples from a native corpus of Portuguese, namely CRPC, in what concerns primary and secondary connectives); (iii) and from Europarl texts, when they are identified through a contrastive approach with English.

5. Format of the lexicon

The integration of the lexicon into different types of applications requires structured information in a machine-readable format such as XML. But while machine-friendly and extremely rich and hierarchical, XML is certainly less human-friendly than a simple spread sheet that allows the immediate comparison and filtering of the entries. We have adopted a mixed approach for the lexicon of Portuguese: data is entered in a single spread sheet and later converted to an XML scheme compatible

[1] http://alfclul.clul.ul.pt/CQPweb/crpcfg16/

with the DiMLex format. The first row of the excel data sheet makes explicit how the field is later on converted to a structured xml file through a perl script. We follow the main components of DiMLex and consider four top-level main components: *Orthographical*, *Syntactic*, *Semantic*, *Synonym* and *Examples*. There are some differences in the contents of each component due to specificities of each project. In LDM-PT, each row corresponds to an association of DM/category/meaning. So, the same word form will occur in two different rows if it has two different categories or two different meanings. This is handled differently in DiMLex (a single entry aggregates different categories and meanings).
The syntactic component <syn> includes information on type, category, context restrictions (mood and tense) and modifiers of the DM. The semantic component <sem> states the 3-level sense. Finally, there are three additional components: <synonym>, <example> and <comments>.

We illustrate an XML entry of the lexicon in Figure 1.

```
<dmarkers>
  <dmarker word="a fim de que" id="dm1">
        <orth1 type="cont">
            <part1 type="phrasal">a fim de que</part1>
            <part2 type=""></part2>
        </orth1>
        <syn>
            <type>primary connective</type>
            <cat>csu</cat>
            <context>
                  <mood>subjunctive</mood>
                  <tense></tense>
            </context>
            <modifier1></modifier1>
            <modifier2></modifier2>
         </syn>
        <sem>
            <relationl1>contingency</relationl1>
            <relationl2>purpose</relationl2>
            <relationl3>arg2-as-goal</relationl3>
        </sem>
        <synonym lexicon="dimlex-en" entry-id="22">so
        that</synonym>
        <examples>
            <example1 source="CRPC">Por fim , a
Comissão sugere um sistema de etiquetagem das viaturas
a fim de que o cliente possa fazer uma escolha com
melhor conhecimento de causa. </example1>
            <example2 source=""></example2>
            <example3 source=""></example3>
          </examples>
          <comment></comment>
  </dmarker>
```

Figure 1: Full XML entry of the continuous and phrasal DM *a fim de que* 'so that'

The top-level <dmarker> component includes attributes regarding the word form of the DM and its numerical id. The Orthographical <orth> component (more than one <orth> component can be included to deal with variants such as initial capital letter and contractions) has an attribute type to describe the continuous our discontinuous nature of the DM. Continuous DMs are described in the subcomponent part1 as belonging to the type single or phrasal. We illustrate in Figure 2 the <orth> component of a discontinuous DM: the type of the <orth> component is "discont" and each part (part1 and part2) are labelled as phrasal or single.

```
<dmarkers>
  <dmarker word="tal como…também" id="dm235">
        <orth1 type="discont">
              <part1 type="phrasal">tal como</part1>
              <part2 type="single">também</part2>
        </orth1>
```

Figure 2: <orth> component of the XML entry of the discontinuous DM *tal como … também* 'just as…so too'

The lexicon was later converted to the DIMLex format to be integrated in the multilingual resource Connective-Lex.info (Stede et al., 2017)[2] through a web app (Dombek, 2017). Due to the different entry structure of LDM-PT and DIMLex, the split-up entries for ambiguous connectives in LDM-PT had to be merged by grouping them, first by word, then by word class. Of the described fields unique to this lexicon, only type was taken over into the DiMLex representation, as a new type attribute for the entry tag, so that it can be displayed by the app. Neither the sense tagset nor the POS tagset had to be converted using mappings. Some used POS tags, e.g. verb, are not specifically represented in the app, but no mapping is necessary for this, as the app automatically represents all unknown tags as 'other'.

6. Applications

Resources with encoded discourse information like LDM-PT have different applications. First of all, they provide data for the annotation of discourse relations in discourse treebanks.
This information can be used directly for manual annotation, in the development of semi-automatic tools (Aleixo and Pardo, 2008), or in fully automatic systems that perform discourse parsing (Pardo and Nunes, 2008; Ziheng et al, 2014; Maziero et al, 2015).
Secondly, they can be integrated in NLP applications dealing with tasks like automatic summarization, information extraction, text generation, machine translation and sentiment analysis (Taboada and Mann, 2006), as well as in the new field of argumentation mining (Peldszus and Stede, 2013).
Finally, linking monolingual lexicons through a pivot English lexicon leads to a multilingual resource and provides data for multilingual applications.

7. Conclusion and future work

We have presented LDM-PT, a new lexicon of DMs for Portuguese. The set of DMs included in the lexicon is based on several sources, ranging from frequency lists extracted from a corpus of contemporary Portuguese, to a multilingual discourse treebank (TED-MDB) and contrastive analysis with English DMs. This accounts for the wide range of syntactic categories that are included in the resource: conjunctions, prepositions, adverbs and adverbial phrases, but also alternative lexicalizations that carry a cohesive function in texts.

[2] http://connective-lex.info

The rich set of features is inspired by both the DiMLex and the LEXCONN lexicons, and covers orthographical information, syntactic category, rhetorical relations, restrictions on the context, examples and an English near-synonym. The latter feature has enabled the linking of LDM-PT in connective-lex.info, a multilingual platform of lexicons of DMs.

The lexicon includes for now 252 pairs of discourse connectives/rhetorical senses. The coverage and sense inventory of the lexicon will be validated in the near future by comparing the set of rhetorical labels for each DM in the lexicon with the TED-MDB corpus, and also with a random selection of contexts from different genres taken from the CRPC corpus.

We plan to enlarge this resource by including pragmatic markers with interactional and modal meaning found in our spoken corpora of Portuguese. Also, our objective is to use the lexicon to automatically pre-annotate DMs in a discourse treebank of Portuguese and to develop automatic tools for discourse parsing.

8. Acknowledgements

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 Europe[3].

9. Bibliographical References

Cartoni, B., Zufferey, S. and Meyer, T. (2013). Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique, *Dialogue and Discourse* (2013), 68-86.

Cuenca, M. and Marín, M. J. (2009) : Co-occurrence of discourse markers in Catalan and Spanish oral narrative, *Journal of Pragmatics 41* (2009), 899–914.

Danlos, L., (2006) "Discourse Verbs" and Discourse Periphrastic Links. In Sidner, C., Harpur, J., Benz, A., Kühnlein, P. (eds), *Proceedings of the Second Workshop on Constraints in Discourse*, Maynooth, Ireland, 2006. 59–65.

Dombek, F. (2017) *Connective-lex.Info – A Web App for a Multilingual Connective Database*. Bachelor Thesis, Department of Linguistics, University of Potsdam.

Généreux, M., I. Hendrickx, A. Mendes (2012) A Large Portuguese Corpus On-Line : Cleaning and Preprocessing. In: Caseli, H. et al. (eds.) *Computational Processing of the Portuguese Language. Proceedings of the 10th International Conference PROPOR1012*. Berlin, Heidelberg: Springer-Verlag, pp. 113-120.

Halliday, M.A.K. and Hasan, R. (1976). *Cohesion in English*. London, Longman.

Lejeune, P., Mendes, A. and Martins, N. (2016) Some considerations on the use of main verbs to express rhetorical relations. In Degand, L., C. Dér, P. Furkó and B. Webber (eds.) *Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference*, Budapest, 11-14 April 2016, 89-92.

Lopes, A., Matos, D., Cabarrão, V., Ribeiro, R., Moniz, H., Trancoso, I., Mata, A. I. (2016). Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers, March 2015, http://arxiv.org/abs/1503.0914, accessed 15 January 2016.

Mírovský, J., Synková, P., Rysová, M., Poláková, L. (2016). Designing CzeDLex – A Lexicon of Czech Discourse Connectives, *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation*, Seoul, South Korea, 449-457.

Peldszus, A. and Stede, M. (2013): From argument diagrams to argumentation mining in texts: A survey. *International Journal of Cognitive Informatics and Natural Intelligence* (IJCINI), 7(1):1–31, 2013.

Prasad, R., Dinesh, N., Lee, A., Miltsakaki, E., Robaldo, L., Joshi, A., and Webber, B. (2008): The Penn Discourse TreeBank 2.0. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'08),* 2961-2968.

Prasad, R., Joshi, A., & Webber, B. (2010). Realization of discourse relations by other means: alternative lexicalizations. *Proceedings of the 23rd International Conference on Computational Linguistics: Posters,* pp. 1023-1031. Association for Computational Linguistics.

Prasad, R., Webber, B., & Joshi, A. (2014). Reflections on the Penn Discourse Treebank, comparable corpora, and complementary annotation. *Computational Linguistics*.

Rohde, H. Dickinson, A., Clark, C., Louis, A. Webber, B. (2015). Recovering discourse relations: Varying influence of discourse adverbials *Proceedings of the EMNLP 2015 Workshop on Linking Models of Lexical, Sentential and Discourse-level Semantics*, pp. 22–31, Lisboa, Portugal.

Rysová, M. and Rysová, K. (2015). Secondary connectives in the Prague Dependency Treebank, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, Uppsala, Sweden, August 24–26 2015, 291–299.

Scheffler, T. and Stede, M. (2016). Adding Semantic Relations to a Large-Coverage Connective Lexicon of German. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC'2016),* 1008-1013.

Taboada, M. and Mann, W. C. (2006). Applications of rhetorical structure theory. Discourse Studies 8 (4): 567–588.

Webber, B., Prasad, R., Lee, A., & Joshi, A. (2016). A Discourse-Annotated Corpus of Conjoined VPs. LAW X, p. 22.

Zeyrek, D., A. Mendes, M. Kurfali (2018) Multilingual Extension of PDTB-Style Annotation: The Case of TED Multilingual Discourse Bank. In *Proceedings of the 7th Language Resources and Evaluation Conference* (*LREC'2018*).

10. Language Resource References

Aleixo, P. & Pardo, T. (2008). CSTTool: Uma Ferramenta Semi-Automática para Anotação de Córpus pela Teoria Discursiva Multidocumento CST. NILC-TR-08-03. NILC. Maio de 2008.

Briz, A., Pons Bordería, S. and Portolés, J. (dirs.) (2003). *Diccionario de partículas discursivas del español*, online since 2003, www.dpde.es.

Feltracco, A., Jezek, E., Magnini, B. and Stede, M. (2016) LICO : A Lexicon of Italian Connectives, *Proceedings of the Third Italian Conference on Computational*

[3] www.textlink.metu.edu.tr

Linguistics (CLiC-it 2016), 2016, Napoli, December 5-7, 2016.

Maziero, E., Hirst, G. and Pardo, T. (2015). Adaptation of Discourse Parsing Models for the Portuguese Language. 140-145.

Mendes, A. and Lejeune, P. (2016). LDM-PT. A Portuguese Lexicon of Discourse Markers. In Degand, L., C. Dér, P. Furkó and B. Webber (eds.) *Conference Handbook of TextLink – Structuring Discourse in Multilingual Europe Second Action Conference*, Budapest, 11-14 April 2016, 89-92.

Pardo, T. and Nunes, M. (2008). On the development and evaluation of a Brazilian Portuguese discourse parser. Journal of Theoretical and Applied Computing, vol. 15, pp. 43–64, 2008.

Roze, C., Danlos, L. and Muller, P. (2012): Lexconn: a French lexicon of discourse connectives, *Revue Discours* (2012), http://discours.revues.org/8645.

Stede, M. (2002) DiMLex: A Lexical Approach to Discourse Markers, in A. Lenci – V. Di Tomaso (ed.), *Exploring the Lexicon - Theory and Computation*, Alessandria (Italy), Edizioni dell'Orso, 2002.

Stede, M., Scheffler, T. and Dombek, Felix (2017). Connective-lex.info, Potsdam University. http://connective-lex.info.

Ziheng, L., Hwee, T. N. and Min-Yen, K. (2014). A PDTB-Styled End-to-End Discourse Parser. Natural Language Engineering, 20, pp 151-184. Cambridge University Press.