LDM-PT - A Portuguese Lexicon of Discourse Markers

Amália Mendes and Pierre Lejeune
Centre of Linguistics / Faculty of Arts
University of Lisbon

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[1]. Each connective is associated to the set of its rhetorical senses, following the PDTB typology[2].

We take discourse markers as a broad category that includes cohesive devices and also pragmatic markers with interactional and modal meanings[3] but we focus for now on discourse connectives. Our immediate goal is to provide data for the annotation of discourse relations in a Portuguese Discourse Treebank, although a listing of discourse connectives will certainly prove to be useful for applications dealing with tasks such as parsing, text processing and summarization of Portuguese. 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[4] 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[5].

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 phrases, but also prepositions, which we consider in our list of connectives, an option that is common to the German lexicon DiMLex[6] and the French lexicon LEXCONN[7].

The identification of discourse connectives was first performed through a constrastive approach to English, based on the parallel Europarl corpus and on the list of connectives labelled in the

[1] M.A.K. Halliday – R. Hasan, *Cohesion in English*, London, Longman, 1976.
[2] Rashmi Prasad – Nikhil Dinesh – Alan Lee – Eleni Miltsakaki – Livio Robaldo – Aravind Joshi – Bonnie Weber, The Penn Discourse Treebank 2.0, in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, 28-30 May, 2008, 2961-2968.
[3] Maria Josep Cuenca – Maria Josep Marín Co-occurrence of discourse markers in Catalan and Spanish oral narrative. *Journal of Pragmatics 41* (2009), 903.
[4] Antonio Briz – Salvador Pons Bordería – José Portolés (dirs.) *Diccionario de partículas discursivas del español*. online since 2003, www.dpde.es.
[5] António Lopes – David Matos – Vera Cabarrão – Ricardo Ribeiro – Helena Moniz – Isabel Trancoso – Ana Isabel Mata, Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers, March 2015, http://arxiv.org/abs/1503.0914, accessed 15 January 2016.
[6] Manfred Stede, 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.
[7] Charlotte Roze – Laurence Danlos – Philippe Muller, Lexconn: a French lexicon of discourse connectives. *Revue Discours* (2012), http://discours.revues.org/8645.

PDTB. We locate discourse markers in the English corpus and inspect the Portuguese sentences to identify the corresponding connectives. We apply 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 discourse connectives denoting the same sense. The approach is close to the Translation Spotting Technique[8], 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 the limits of the category (for instance, the case of referential NPs that perform a cohesive function) and to other strategies that express coherence relations between text spans, such as paraphrases. This approach is now complemented using our preparatory work to develop a discourse treebank for Portuguese in the PDTB framework by annotating texts of the corpus CINTIL[9], a 1M word corpus annotated for part-of-speech and manually revised. Our annotation focuses on discourse connectives (conjunctions, adverbs, phrases and prepositions), taken as elements that express a two-place semantic relation filled by propositional arguments.

The lexicon is structured as pairs of discourse connectives/rhetorical senses, so as to cover polysemous connectives. The lexicon includes at the moment 210 pairs of discourse connectives/rhetorical senses and is, for now, implemented in excel format (an illustration of the discourse connectives is provided in Table 1).

Portuguese DM	Rhetorical Relation	Other rhetorical relations	English DM
com efeito, é que, na medida em que, pois, porque, visto (que)	reason	Pragmatic Cause: Justification	for
daí (que), de onde, por conseguinte, por consequência, portanto	result		hence
apesar de, embora, contudo	contrast		though

Table 1: Discourse Markers in the LDM-PT

[8] Bruno Cartoni – Sandrine Zufferey – Thomas Meyer, Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique, *Dialogue and Discourse* (2013), 68-86.

[9] Florbela Barreto *et al.*, Open Resources and Tools for the Shallow Processing of Portuguese: the TagShare project, *Proceedings of the V International Conference on Language Resources and Evaluation – LREC 2006*, Genova, 2006, 1438-1443.

Additional information on the category of the connective is provided in a required field *Category* (conjunction, preposition, adverb), and other optional information is encoded, such as the equivalent English connective, a corpus example and restrictions on the context (e.g., the presence of a negative particle, mood selection (indicative or subjunctive), modifiers). The latter might be especially important to deal with connectives that share a common rhetorical sense although they don't occur in the same contexts since "connectives are not always interchangeable and therefore cannot be treated as equivalents"[10].

We also include in the lexicon Alternative Lexicalizations (AltLex), i.e, alternative expressions that denote a cohesive relation, making it redundant to supply an implicit connective in the context[11]. For instance, the cohesive relation *contrast* is frequently denoted through the following AltLex: *acontece que* 'it happens that', *diga-se que* 'let it be said that', *dito isso* / *posto isso* 'this being said', *não deixa de ser verdade que* 'it is nevertheless true that'. We have also encountered borderline cases of intra-sentential discourse relations marked by a main causative verb[12], such as *provocar* 'to provoke', obrigar 'to force', *reduzir* 'ro reduce', which typically establish a causal coherence relation between two nominalizations[13].

The lexicon is viewed as an open list that integrates both the results of the contrastive analysis between English and Portuguese discourse connectives and of our corpus annotation following the PDTB model.

Bibliography

BARRETO, Florbela – BRANCO, António – FERREIRA, Eduardo – MENDES, Amália – BACELAR DO NASCIMENTO, Maria Fernanda – NUNES, Filipe – SILVA, João, Open Resources and Tools for the Shallow Processing of Portuguese: the TagShare project, *Proceedings of the V International Conference on Language Resources and Evaluation – LREC 2006*, Genova, 2006, 1438-1443.

BRIZ, Antonio – PONS BORDERÍA, Salvador – PORTOLÉS, José (dirs.) *Diccionario de partículas discursivas del español*, online since 2003, www.dpde.es.

[10] Cartoni *et al*., The Translation Spotting Technique, 79.

[11] Rashmi Prasad *et al.*, Realization of Discourse Relations by Other Means: Alternative Lexicalizations, in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, 2010, 1025.

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

[13] Pierre Lejeune – Amália Mendes – Nuno Martins, Some considerations on the use of main verbs to express rhetorical relations, this volume, 2016.

CARTONI, Bruno – ZUFFEREY, Sandrine – MEYER, Thomas, Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique, *Dialogue and Discourse* (2013), 68-86.

CUENCA, Maria Josep – MARÍN, Maria Josep, Co-occurrence of discourse markers in Catalan and Spanish oral narrative, *Journal of Pragmatics 41* (2009), 899–914.

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

HALLIDAY, M.A.K. – HASAN, R. *Cohesion in English*, London, Longman, 1976.

LEJEUNE, Pierre – MENDES, Amália – MARTINS, Nuno, Some considerations on the use of main verbs to express rhetorical relations, this volume, 2016.

LOPES, António – MATOS, David – CABARRÃO, Vera – RIBEIRO, Ricardo – MONIZ, Helena – TRANCOSO, Isabel – MATA, Ana Isabel, Towards Using Machine Translation Techniques to Induce Multilingual Lexica of Discourse Markers, March 2015, http://arxiv.org/abs/1503.0914, accessed 15 January 2016.

PRASAD, Rashmi – DINESH, Nikhil – LEE, Alan – MILTSAKAKI, Eleni – ROBALDO, Livio – JOSHI, Aravind – WEBER, Bonnie, The Penn Discourse Treebank 2.0., in *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, 28-30 May, 2008, 2961-2968.

PRASAD, Rashmi – JOSHI, Aravind – WEBBER, Bonnie, Realization of Discourse Relations by Other Means: Alternative Lexicalizations, in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, 2010, 1023–1031, http://www.aclweb.org/anthology/C10-2118, accessed 12 March 2016.

ROZE, Charlotte – DANLOS, Laurence – MULLER, Philippe (2012) Lexconn: a French lexicon of discourse connectives, *Revue Discours* (2012), http://discours.revues.org/8645.

STEDE, Manfred 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.