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

We take discourse markers as a broad category that includes cohesive devices and also pragmatic markers with interactional and modal meanings 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 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.

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 and the French lexicon LEXCONN.

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

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, 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, 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).

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

Table 1: Discourse Markers in the LDM-PT

---


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 context11. 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 verb12, such as provocar ‘to provoke’, obrigar ‘to force’, reduzir ‘to reduce’, which typically establish a causal coherence relation between two nominalizations13.

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**


---

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


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


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


