LDM-PT - A Portuguese Lexicon of Discourse Markers

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

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. Each connective is associated to the set of its rhetorical senses, following the PDTB typology. The lexicon contains for now 210 pairs of discourse markers/rhetorical senses.

Lexical resources available for Portuguese deal essentially with content words and even those focusing on multi word expressions favour content expressions.

The goal is to provide data for:

- The annotation of discourse relations in a Portuguese Discourse Treebank;
- Applications for parsing, text processing and summarization of Portuguese.

2. Discourse markers

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. We consider that discourse connectives:

- Do not vary regarding inflection;
- Express a two-place semantic relation;
- Have propositional arguments and;
- Are not integrated in the predicative structure.

3. Methodology

The identification of discourse connectives follows a dual approach:

(i) a manual contrastive approach to English

Using the list of connectives labelled in the PDTB, we locate those connectives in the English subpart of the parallel corpus Europarl and inspect the equivalent Portuguese sentences to identify the corresponding connectives.

The method is close to the **Translation Spotting Technique** (Cartoni & Zufferey, 2013). Our motivation however is to acquire a diversified set of connectives in Portuguese and not to capture the different meanings of a given connective in the source language.

We apply a **manual approach** to:

- procure fully accurate data,
- identify potential new senses of the Portuguese connectives,
- spot semantic and pragmatic differences between discourse connectives denoting the same sense,
- Identify other strategies that express coherence relations between text spans, such as alternative lexicalizations.

(ii) Corpus annotation

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.

4. Content of the lexicon

The lexicon is structured as pairs of discourse connectives/rhetorical senses, so as to cover polysemous connectives.

The connectives are conjunctions, prepositions, adverbs and phrases, and also Alternative Lexicalizations (AltLex), i.e., alternative expressions that denote a cohesive relation, making it redundant to supply an implicit connective in the context.

Examples of AltLex in the lexicon and their rhetorical sense:

acontece que	'it happens that'	contrast
diga-se que	'let it be said that'	contrast
dito isso / posto isso	'this being said'	contrast
não deixa de ser verdade que	'it is nevertheless true that'	contrast
provocar	'to provoke'	cause:result
obrigar	'to force'	cause:result
reduzir	'to reduce'	cause:result

Fields of the lexicon:

- Rhetorical sense (first-level required + two-levels of granularity following PDTB
- Additional rhetorical sense
- Category (required) (conjunction, preposition, adverb, alternative lexicalization)
- Restrictions on the context. E.g., the presence of a negative particle, mood selection. (Especially important to deal with connectives that share a common rhetorical sense although they are not interchangeable)
- Equivalent English connective in PDTB
- Modifiers of the connective
- Source (Europarl or corpus annotation)
- Corpus example
- Comments

Connective	ainda que	então	então
Sense – first level	comparison	contingency	temporal
Sense – second level	contrast	cause	asynchronous
Sense – third level		result	precedence
Category	conjunction	adverb	adverb
Restrictions	subjunctive	-	-
English connective	so that	then	then
Modifiers			
Source	Europarl		

XML format

```
<dmarker type="connective" cat="conj" id="dm1" relationl1="contingency" relationl2=”purpose"
relationl3=”arg2-as-goal" source="EC”>
   <form>a fim de que</form> <context mood="subj"/>
   <ENsynonym>so<Ensynonym><PTsynonym dmarker="dm149"/>
</dmarker>

<dmarker type="connective" cat="altlex" id="dm182" relationl1="contingency" relationl2="cause"
relationl3="reason” source=“AN””>
   <form>razão pela qual</form
   <example>Finalmente, o estudo concluiu que a prática da actualização anual e a separação
   entre danos totais e danos parciais, deixaram agora de se verificar, razão pela qual os
   consumidores têm sobejas razões para se congratular.</example>
</dmarker>
```

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.

Selected references:

Cartoni, B., S. Zufferey, T. Meyer (2013) Annotating the Meaning of Discourse Connectives by Looking at their Translation: The Translation Spotting Technique, *Dialogue and Discourse* (2013), 68-86.
Cuenca, M. J., M. Marín (2009) Co-occurrence of discourse markers in Catalan and Spanish oral narrative, *Journal of Pragmatics 41* (2009), 899–914.
Prasad, R., A. Joshi, B. Webber (2010) Realization of Discourse Relations by Other Means: Alternative Lexicalizations, in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, Beijing, 2010, 1023–1031.

U LISBOA | UNIVERSIDADE DE LISBOA FLUL LETRAS LISBOA
FCT Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA
CLUL | Centro de Linguística da Universidade de Lisboa