A Lexical Database of Multi-Word Expressions in Portuguese

Amália Mendes and Sandra Antunes

Centro de Linguística da Universidade de Lisboa

CLUL | Centro de Linguística da Universidade de Lisboa

A Portuguese lexicon of MWE:

- includes idiomatic expressions, collocations, institutionalized phrases, favoured co-occurring forms, etc.
- contains approximately 14.000 entries
- extracted from a 50M balanced Portuguese written corpus
- selected from a sorted list of n-grams based on the MI measure and validated manually
- organized under canonical forms that include several types of variation, such as inflection, gender, lexical insertion, etc. (47.224 forms)

- Lexicon implemented on a MySQL relational database
- Includes 221.847 concordances manually verified

- Proposal for MWE annotation in running text:
  - Linking of idiomatic occurrences to the MWE-entry in the lexicon
  - Information on the canonical form and variations
  - Specific annotation guidelines for each syntactic pattern