A Corpus-Driven Study of the Variation of Co-Occurrence Patterns in Written and Spoken Registers

Amália Mendes,[1] Maria Fernanda Bacelar do Nascimento and Sandra Antunes

Abstract

This paper will focus on the study of the variation of co-occurrence patterns encountered in written and spoken registers, through the analysis of a large lexical database of corpus-extracted multiword expressions (MWEs) of European Portuguese. Those MWEs were automatically extracted from a balanced 50 million word written corpus and a 1 million word spoken corpus, furthermore statistically interpreted using lexical association measures and partially manually validated in what concerns written units.

MWEs have been and are still a challenge for linguistic analysis, lexicography and natural language processing due to their large pattern variation and to the need to put forward several linguistic levels for their analysis, namely in parameters like degree of syntactic cohesion, inflected variation and semantic compositional nature.

In this paper, we aim to revise some typologies of MWEs using a corpus-driven approach, to analyse corpus findings and their relation to MWEs categorization, and to establish possible contrastive registers based on syntactic, functional and semantic paradigms: for example, contrasts involving spoken and written texts or contrasts involving the degree of formality taken transversally in both registers.

By presenting register-specific co-occurrence patterns based on authentic data, this study will hopefully contribute to the more general categorization of MWEs in Portuguese.

[1] *e-mail*: amalia.mendes@clul.ul.pt