A Lexical Database of Multi-Word Expressions in Portuguese

Amália Mendes and Sandra Antunes

Centre for Linguistics at the University of Lisbon

WG1

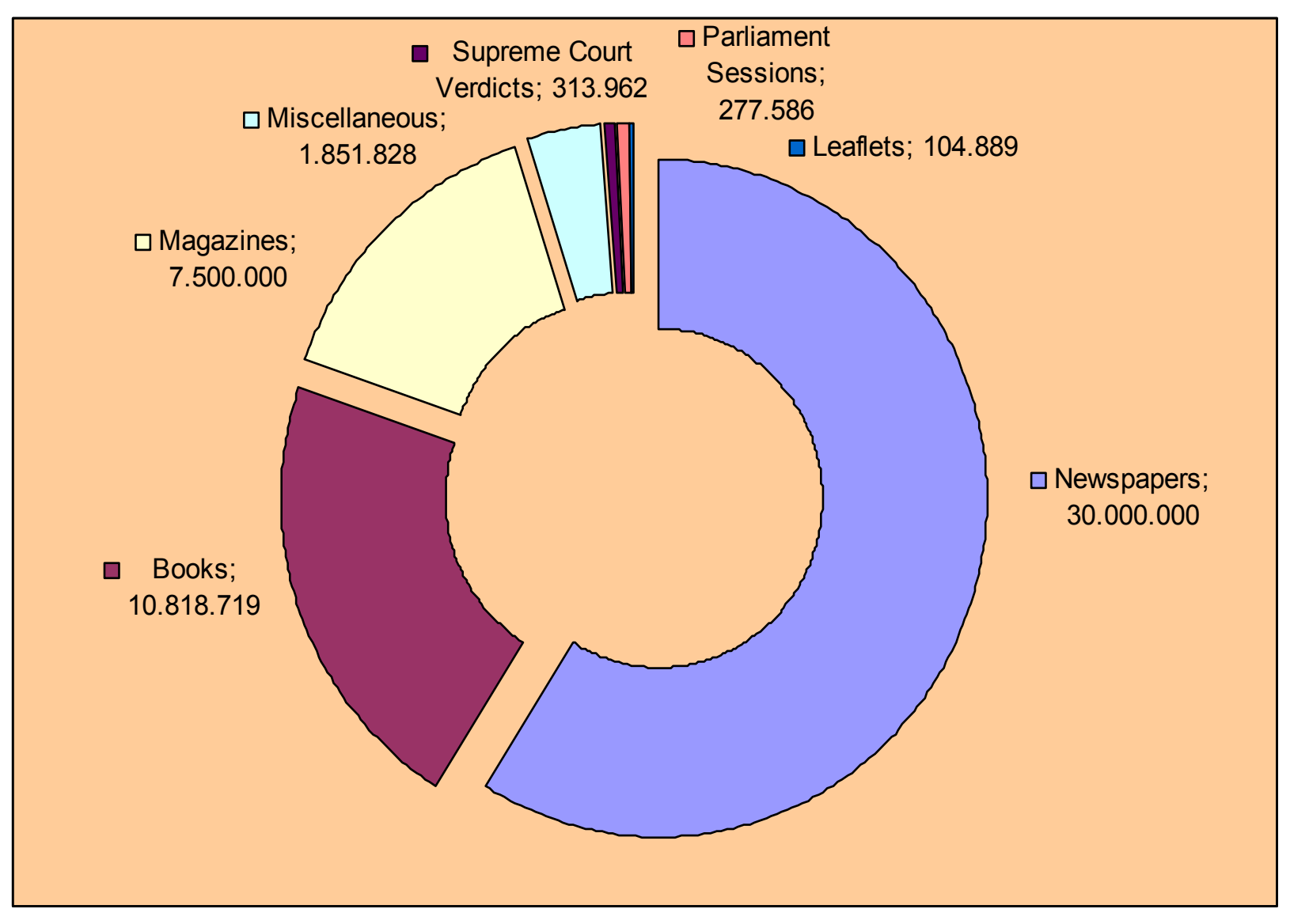
1 – Introduction

We present an overview of our work on multiword expressions (MWE) in Portuguese. We discuss the methodology followed to extract a lexicon of MWE from a 50 million words corpus, as well as the typology of MWE encountered in our data. We sketch a proposal for the annotation of idiomatic expressions in running text, aiming to create a resource that allows us not only to analyze their behavior in context, but also to evaluate automatic MWE's identification systems.

2 – Lexicon and Corpus

http://www.clul.ul.pt/sectores/linguistica_de_corpus/manual combinatorias online.php

We have developed a Portuguese MWE lexicon, implemented on a MySQL relational database that was extracted from a 50M balanced Portuguese written corpus with the following design:



This Portuguese Lexicon of MWE:

- ➢ contains approximately 14.000 entries;
- ➢ was selected from a sorted list of n-grams based on the mutual information (MI) measure and validated manually;
- ➢ is organized under canonical forms that include several types of variation, such as inflection, gender, lexical insertion, etc.
- ➢ includes idiomatic expressions, collocations, institutionalized phrases, favoured co-occurring forms, with different degrees of lexicalization and different levels of idiomaticity:
  - ✓MWEs can be totally idiomatic, when the global meaning can not be recovered by the sum of the individual meanings of its elements – *deitar água na fervura* / *to pour oil on trouble waters* = to calm down a situation;
  - ✓MWEs can be partially idiomatic, when one or more elements are used in their literal meaning, while others have an idiomatic meaning – *saúde de ferro* / *iron health* = strong health (an allusion to the strength of the metal);
  - ✓MWEs can be compositional, but receive an additional meaning – *deitar as mãos à cabeça* / *to grab one's head + to be unbelievable/incomprehensible.*

3 – Evaluation of lexical association measure and frequency

Based on our previous studies in MWE automatic extraction and evaluation, we selected candidates with MI values between 8 and 10 and frequency ≥ 6. However, during manual validation, we found significant expressions that showed very different MI values and frequencies, raising the question of the performance of statistical measures in identifying MWE:

➢ MWE with low MI and high frequency (they contain very high-frequent words in the corpus, like auxiliary verbs or light verbs, that lower the MI value)

MWE	MI	Freq
ter força 'to have strenght'	2.2	306
ganhar tempo 'to save time'	3.1	83
estar em forma 'to be in good shape'	2.9	82

➢ MWE with low MI and low frequency (they are all intuitively recognized as MWE by native speakers, but no such correspondence is to be found in quantitative criteria)

MWE	MI	Freq
fonte de vida 'source of life'	2.7	5
de última geração 'state-of-the-art'	3.4	4
folha de serviço 'track record'	4.5	5

4 – Annotation of MWE in running text

We plan to use the MWE lexicon to automatically annotate a corpus of 1M tokens, of both spoken and written data, tagged with POS information.

Each MWE in the CINTIL corpus will be annotated with a link to the MWE-entry in the lexicon. The lexicon contains the typical properties of MWE, such as:
(i) canonical form;
(ii) synonyms (or literal paraphrases);
(iii) typology labels (MWE PoS category and the PoS categories of its elements, its fixed or semi-fixed nature, its idiomatic property and possible additional meanings);
(iv) functions of MWE parts (e.g. obligatory / optional / free part).

We propose to divide our annotation guidelines according to syntactic patterns, since MWE reveal different properties:

- ➢ sentence level MWE (proverbs and aphorisms) usually do not accept syntactic changes, like passivization, relativization or inflectional variation, while verb phrases admit much more morpho-syntactic variation;
- ➢ fixed NPs can behave as compound nouns and the modifiers of the noun can express different semantic relations (part of, made of, used for, etc.) that may interact with the meaning (literal or idiomatic) of the noun.

Regarding MWE internal variation, mapping MWE occurrences in the corpus to their canonical form will depend on their lexical, syntactic and structural variation:

➢ Free lexical realizations – These elements are marked in the lexicon with, e. g. a pronoun (ALGUÉM 'someone', ALGUM 'something') or a particular phrase (NP, PP).

estar nas mãos de ALGUÉM [*to be in the hands of someone*]

➢ restricted variation – Variation is restricted to limited set of alternatives that is recorded in the MWE lexicon as '*obligatory parts of the MWE and member of a set list*'.

comer / *vender* / *comprar* / *tomar* / *impingir* / *levar gato por lebre* (verb alternation)

[*to eat / to sell / to buy / to receive / to impose / to take a cat instead of a hare* = to buy a pig in a poke]

➢ Insertion – Lexical elements that do not belong to the canonical form are not part of the MWE and are not labelled (usually, they have an emphatic function).

dizer cobras e lagartos / dizer sempre cobras e lagartos

[*to say snakes and lizards / to say always snakes and lizards* = to speak ill of someone]

➢Pronominalization / Possessives – These elements will be marked up as part of the MWE, but will have an additional label to signal that they are optional.

estar nas suas mãos / estar nas mãos dele [*to be in his hands / to be in the hands of him*]

➢Truncation – We do not label explicitly this phenomenon. The occurring part is marked with a reference link to the MWE in the lexicon.

mais vale um pássaro na mão (*do que dois a voar*) [*a bird in the hand is worth* (*two in the bush*)]

➢ Idioms manipulation – Creative use of language can lead to MWEs that only partly match the canonical MWE. We label these parts as '*different from canonical form*' .

no poupar / anunciar / atacar / descontar / esperar/ comparar / cooperar é que está o ganho

[*in the saving / announcing / attacking / discounting / waiting / comparing / cooperating is the profit*]

Publications

S. Antunes, M. F. Bacelar do Nascimento, J. M. Casterleiro, A. Mendes, L. Pereira, and T. Sá, 2006. A Lexical Database of Portuguese Multiword Expressions. In LNAI, volume 3960, pages 238–243. Springer- Verlag, Berlin (PROPOR 2006).

M. F. Bacelar do Nascimento, A. Mendes, and S. Antunes, 2006. Spoken Language Corpus and Linguistic Informatics, chapter Typologies of MultiWord Expressions Revisited: A Corpus-driven Approach, pages 227–244. Coll. Usage-Based Linguistic Informatics, Vol.V. John Benjamins.

I. Hendrickx, A. Mendes and S. Antunes, 2010. Proposal for Multi-word Expression annotation in running text. In: *Proceedings of the fourth Linguistic Annotation Workshop (LAW IV)*, Association for Computational Linguistics, Uppsala, Sweden, pp 152-156.

A. Mendes, S. Antunes, M. F. Bacelar do Nascimento, J. M. Casteleiro, L. Pereira, T. Sá, 2006. COMBINA-PT: a Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. *Proceedings of the V International Conference on Language Resources and Evaluation - LREC2006*, Genoa, May 22-28, 2006.

Contactos

Av. Professor Gama Pinto, 2, 1649-003 Lisboa, Portugal

Tel.: +351 21 790 47 00 | Fax: +351 21 796 56 22

URL: http://www.clul.ul.pt

CLUL Centro de Linguística da Universidade de Lisboa