Title: A Lexical Database of Multi-Word Expressions in Portuguese

Amália Mendes and Sandra Antunes
Centre for Linguistics at the University of Lisbon

WG1, and possibly WG2

In this poster, we present an overview of our work on multi-word expressions (MWE) in Portuguese. We discuss the methodology followed to extract a lexicon of MWE from a 50 million words corpus, as well as a typology of the MWE encountered in our data. Finally, we briefly sketch a proposal for the annotation of MWE in running text using this lexicon.

1. As it is widely known, the lexicon does not consist mainly of simple lexical items but appears to be populated with numerous chunks, more or less predictable, though not fixed. In fact, the development of computer technologies and corpus-based approaches has enabled the identification of complex patterns of word associations, and showed that great part of a speaker's lexicon is composed by these word associations.

In contrast to languages for which there is a wide range of studies regarding MWE both from a linguistic and a computational point of view (e.g, Sag et al, 2002; Fellbaum et al., 2006; Calzolari et al, 2002; Baldwin and Kim, 2010;), for Portuguese most studies have paid more attention to idiomatic expressions and compound nouns in general, relegating the analysis of other types of expressions to the morpho-syntactic properties of its elements (cf. Antunes and Mendes 2012).

Using a balanced written corpus of 50 million words, we compiled a lexicon of MWE that includes idiomatic expressions, but also collocations: expressions of frequently cooccurring words that do not show syntactic or semantic fixedness. This lexicon is implemented on a MySQL relational database. The MWE are organized under canonical forms and inflectional variations of these canonical forms are recorded. In total, the lexicon contains 14,153 canonical forms and 48,154 MWEs variations. For each of those several examples are collected from the corpus. Each MWE entry is also assigned to one or multiple word lemmas, of a total number of 1180 single word lemmas. The MWE were selected from a sorted list of n-grams based on the Mutual Information measure (Church and Hanks, 1990) and validated manually (Mendes et al., 2006). Several criteria were applied, on which the definition of a MWE usually relies:

a) lexical and syntactic fixedness that can be observed through the possibility of replacing elements, inserting modifiers, changing the syntagmatic structure or gender/number features;
b) total or partial loss of compositional meaning, which means that the meaning of the expressions can not be predicted by the meaning of the parts;
c) frequency of occurrence, which means that the expressions may be semantically compositional but occur with high frequency, revealing sets of favoured co-occurring forms, known as collocations.

2. Considering the existence of different types of MWE with different degrees of syntactic and semantic cohesion, our analysis tries to categorize these expressions taking into account their lexical, syntactic, semantic and pragmatic properties. From a semantic standpoint, three major classes were considered: (i) expressions with compositional meaning (*pão de centeio* 'rye bread'); (ii) expressions with partial idiomatic meaning, i.e., at least one of the elements keeps its literal meaning (*vontade de ferro* 'iron will'); (iii) expressions with total idiomatic meaning (*pés de galinha* 'crow's feet').

Note, however, that one may find notorious difficulties regarding the evaluation of the meaning of certain expressions that seems to be linked to two major factors: (i) the polysemous nature of the words (it is necessary to establish a boundary between compositional and figurative meanings. If we consider the literal meaning to be the first prototypical meaning of a word, this restrictive definition will trigger us to consider a large number of MWE as idiomatic); (ii) the awareness of the semantic motivation that had led to the idiomatic meanings, which depends on cultural and social factors. Although we tried to accentuate the different degrees of lexicalization of this type of expressions, we are acutely aware that drawing this dividing line neither is easy nor allows for accurate definitions and divisions.

Within each of these three semantic categories, the expressions are also analyzed according to their grammatical category and lexical and syntactic fixedness. Regarding the latest aspect, the expressions may be: (i) fixed (no variation); (ii) semi-fixed (nominal/verbal inflection); (iii) with variation: lexical

(permutation, replacement of elements, insertion of modifiers) and/or syntactic (constructions with passives, relatives, pronouns, extraction, adjectival vs. prepositional modifiers).

This information will allow us to enrich the lexicon. Our purpose is to label each MWE entry in the lexicon regarding: (i) canonical form of the expression; (ii) definition of idiomatic expressions through synonyms or literal paraphrases; (iii) grammatical category of both the expression and its elements; (iv) idiomatic property and additional meanings; (v) possible variation; (vi) function of the internal elements of MWE (e.g., obligatory, optional, free).

3. The lexicon is envisaged as the basis for the annotation of idiomatic MWE in running text. Each MWE encountered in the corpus would be annotated with a link to the corresponding entry in the lexicon. Linking each MWE to its canonical form would allow for an easier detection of all occurrences of one particular MWE and check its variation in the corpus. The annotation process would combine automatic retrieval with manual validation in order to better account for variable expressions (Hendrickx et al., 2010).

When studying the MWE lexicon, we noticed different properties of MWEs according to their syntactic patterns. Consequently, we propose to divide our annotation guidelines according to each syntactic pattern. At the sentence level, MWEs such as proverbs or aphorisms (e.g. *água mole em pedra dura tanto bate até que fura* lit. 'water in hard rock beats so long that it finally breaks') do not accept any possible syntactic changes nor inflectional variation, but may accept lexical variation. On the contrary, MWEs which are verb phrases will admit much more morpho-syntactic variation. We will discuss the problematic cases for annotation and proposed solutions, focusing on the variational properties of MWEs. Following Moon (1998), we will also assume that, in most of the cases, these expressions "have fixed or canonical forms and that variations are to some extent derivative or deviant". The canonical forms of (variable) expressions are listed in the MWE lexicon. Without doubt, the corpus would contain many MWE that were not yet listed in the lexicon. Therefore, each sentence would need to be checked manually for new MWE and the newly discovered expression would be manually added to the lexicon.

4. We believe that the lexical database of MWE is an important source of data to evaluate the automatic identification and extraction of MWE and to establish a typology of word associations in Portuguese, and that it could further be used for MWE annotation in running texts.

References

S. Antunes and A. Mendes. 2013. "MWE in Portuguese: proposal for a typology for annotation in running text", *The 9th Workshop on Multiword Expressions (MWE 2013)*, Workshop at NAACL 2013, Atlanta, Georgia, USA, June 13/14, 2013.

T. Baldwin and S. Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, Second Edition. CRC Press, Taylor and Francis Group, Boca Raton, FL. ISBN 978-1420085921.

N. Calzolari, C. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod, and A. Zampolli. 2002. Towards best practice for multiword expressions in computational lexicon. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Spain, pp. 1934–1940.

K.W. Church and P. Hanks. 1990. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1): 22–29.

C. Fellbaum, A. Geyken, A. Herold, F. Koerner, and G. Neumann. 2006. Corpus-based studies of German idioms and light verbs. International Journal of Lexicography, 19(4): 349-360.

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, M. F. Bacelar do Nascimento, S. Antunes, and L. Pereira. 2006. COMBINA-PT: a large corpus-extracted and hand-checked lexical database of portuguese multiword expressions. In *Proceedings of LREC 2006*, Genoa, Italy, pp. 1900–1905.

R. Moon. 1998. *Fixed expressions and idioms in English: A corpus-based approach*. In Oxford Studies in Lexicography and Lexicology. Clarendon Press, Oxford.

I. Sag, T. Baldwin, F. Bond, A. Copestake, and D. Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. In *Proceedings of CICLING-2002*.