MWE in Portuguese: Proposal for a Typology for Annotation in Running Text

Sandra Antunes and **Amália Mendes**
Centro de Linguística da Universidade de Lisboa
Av. Prof. Gama Pinto, 2
1649-003 Lisboa, Portugal
{sandra.antunes, amalia.mendes}@clul.ul.pt

Abstract

Based on a lexicon of Portuguese MWE, this presentation focuses on an ongoing work that aims at the creation of a typology that describes these expressions taking into account their semantic, syntactic and pragmatic properties. We also plan to annotate each MWE-entry in the mentioned lexicon according to the information obtained from that typology. Our objective is to create a valuable resource, which will allow for the automatic identification MWE in running text and for a deeper understanding of these expressions in their context.

1 Introduction

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 (Firth, 1955). In fact, the development of computer technologies and corpus-based approaches has enabled the identification of complex patterns of word associations, proving that the speakers use a large number of preconstructed phrases that constitute single choices (Sinclair, 1991:110). Several studies have also shown that great part of a speaker's lexicon is composed by these word associations (Jackendoff, 1997; Fellbaum, 1998). These multiword expressions (MWE)[1] appear in every kind of spoken and written discourse and, despite the fact that they don't pose any problems from the speaker's point of view (we easily recognize that they function as a single unit that may have a specific meaning), natural language processing (NLP) applications, on the other hand, find notorious difficulties when dealing with them (Sag et al., 2000).

Bearing in mind the extreme importance of the study of this linguistic phenomenon for the improvement of NLP systems, this paper will address an ongoing analysis that aims to create a typology for MWE in Portuguese (based on a MWE lexicon previously extracted from a 50 million word written corpus) that will be used to enrich that lexicon with extensive information regarding these expressions. This annotated lexicon will be a resource that will allow for the annotation of these expressions in running text (Hendrickx et al., 2010a).

This presentation will briefly discuss compilation of the lexicon and the methodology adopted for MWE selection and organization (section 2), the typology based on syntactic, semantic and statistic criteria (section 3), the annotation proposal of the lexicon (section 4) and applications of the work (section 5).

2 MWE: Corpus and Lexicon

The work we are going to present used the lexicon of word combinations[2] that was created within the scope of the project COMBINA-PT – Word Combinations in Portuguese Language[3]. The corpus used for their extraction was 50 million word writ-

[1] The term multiword expression will be used to refer to any sequence of words that act as a single unit, embracing all different types of word combinations (collocations, compound nouns, light verbs, institutionalized phrases, idioms, etc.).

[2] The lexicon is available at Meta-Share repository: http://www.meta-net.eu/meta-share.

[3] https://www.clul.ul.pt/en/research-teams/187-combina-pt-word-combinations-in-portuguese-language

Proceedings of the 9th Workshop on Multiword Expressions (MWE 2013), pages 87–92,
Atlanta, Georgia, 13-14 June 2013. ©2013 Association for Computational Linguistics

ten corpus extracted from the Reference Corpus of Contemporary Portuguese[4], and has the constitution presented in Table 1 (Mendes et al., 2006):

CORPUS CONSTITUTION	
Newspapers	**30.000.000**
Books	**10.917.889**
Magazines	**7.500.000**
Miscellaneous	**1.851.828**
Leaflets	**104.889**
Supreme court verdicts	**313.962**
Parliament sessions	**277.586**
TOTAL	**50.966.154**

Table 1. Constitution of the corpus

The MWE in the lexicon are organized in order to identify a main lemma (from which the MWE was selected) and a group lemma, which corresponds to the canonical form of the MWE and covers all the variants that occurred in the corpus. Concordances lines for each MWE are also available in KIWIC format. Table 2 illustrates some MWE that were identified when analyzing the lemma *fogo* ‘fire’.

Main Lemma	**Group Lemma**	**Concordances**
fogo ‘fire’	*arma de fogo* ‘firearm’	*uma arma de fogo relativamente leve* ‘a relatively light firearm’ *800 mil portugueses possuem armas de fogo* ‘800 thousand Portuguese have firearms’
	baptismo de fogo ‘baptism of fire’	*teve o seu baptismo de fogo no assalto* ‘he had his baptism of fire in a robbery’
	fogo cruzado ‘crossfire’	*civis apanhados no fogo cruzado entre o exército* ‘civilians were caught in a crossfire between the army’
	fogo de artifício ‘firework’	*espectáculos de fogo de artifício* ‘firework shows’ *à 1 hora haverá fogos de artifício* ‘there will be fireworks at 1:00 a.m.’

Table 2. Example of MWE for the lemma *fogo* ‘fire’

In all, the lexicon comprises 1.180 main lemmas, 14.153 group lemmas and 48.154 word combinations.

Mendes et al. (2006) describe the criteria used for MWE selection: following the results of previous studies (Evert and Krenn, 2001; Pereira and Mendes, 2002), the authors first selected groups with MI[5] values between 8 and 10, and, throughout manual validation, applied several criteria upon which usually relies the definition of a MWE:

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, which could tell that they may be in their way to a possible fixedness.

3 Data Analysis: Towards a Typology

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, for Portuguese little work has been done so far. Great part of the existing studies had paid more attention to idiomatic expressions and compound nouns in general, relegating the analysis of other types of expressions to the morphossyntactic properties of its elements (Macário Lopes, 1992; Chacoto, 1994; Baptista, 1994; Vilela, 2002; Ranchhod, 2003)[6].

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. Thus, 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

[4] CRPC is a monitor corpus of 311 million words, constituted by sampling from several types of written and spoken text and comprising all the national and regional varieties of Portuguese (https://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc).

[5] Statistical association measure (Church and Hanks, 1990).

[6] Some research has been carried out regarding the identification and annotation of Complex Predicates, usually called in the literature Light Verb Constructions or Support Verb Constructions (Hendrickx et al., 2010b; Duran et al., 2011; Zeller and Padó, 2012).

(*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.

This semantic criterion implies that the same type of MWE may occur in different classes. It is the case with compound nouns. 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)[7]; (iii) with variation: lexical (permutation, replacement of elements, insertion of modifiers) and/or syntactic (constructions with passives, relatives, pronouns, extraction, adjectival vs. prepositional modifiers).

Our typology relies, then, on several categories, some of which we will briefly present.

Expressions with Compositional Meaning

➢ Favoured co-occurring forms – expressions that occurred with high frequency in the corpus, revealing a tendency to co-occur in certain contexts (*pão seco* 'dry bread', *desvendar o mistério* 'unravel the mystery'). Expressions with full lexical and syntactic variation[8].

➢ Compound nouns – expressions that represent a single concept (*noite de núpcias* 'wedding night', *cama de casal* 'double bed', *cavalo alazão*[9] 'chestnut horse', *Idade do Ferro* 'Iron Age'). Usually, these expressions are semi-fixed. However, we also observed that some combinations may occur in a small distributional paradigm (*cama de solteiro* 'single bed') that allows for predicative constructions (*a cama é de solteiro* lit. 'the bed is single'). Entities are fixed.

➢ Institutionalized expressions – expressions observed with higher frequency than any alternative lexicalization of the same concept (*lufada de ar fresco* 'breath of fresh air', *condenar ao fracasso* 'doomed to failure', *abrir um precedente* 'set a precedent'). Apart from inflection, since there are alternative expressions, we also observed lexical variation, such as substitution (*rajada de ar fresco* 'rush of fresh air'), insertion of modifiers (*condenar este projecto ao fracasso* lit. 'to doom this project to failure') and change in the syntagmatic structure (*o precedente foi aberto* 'a precedent has been set', *abertura de um precedente* lit. 'the opening of a precedent').

➢ Ligh verb constructions – expressions where the noun is used in a normal sense and the verb meaning appears to be bleached (*dar um passeio* 'take a walk'). Expressions with lexical and syntactic variation (substitution, insertion of modifiers, change in the syntagmatic structure).

➢ proverbs (*no poupar é que está o ganho* 'profit is in saving'). Despite our conception of proverbs as frozen expressions, the fact is that speakers' lexical creativity may result in the production of expressions such as *no anunciar/atacar/descontar/esperar/comparar é que está o ganho* 'profit is in announcing/attacking/discounting/waiting/comparing'.

Expressions with Partial Idiomatic Meaning

➢ Expressions with an additional meaning that can not be derived from the meaning of its parts[10], (*cinturão negro* 'black belt' + martial arts expert, *abrir a boca* 'open the mouth' + to speak/to yawn, *deitar as mãos à cabeça* lit. 'throw the hands in the head' (throw one's hands up) + despair). Nominal expressions are semi-fixed while verbal expressions may undergo inflection and lexical variation, such as substitution (*levar/lançar as mãos à cabeça* lit. 'put/lay the hands in the head') and insertion of modifiers (*deitou logo as mãos à cabeça* lit. 'put immediately his hands in his head').

[7] Since Portuguese is a highly inflectional language, practically all the verbs and nouns that occur in MWE inflect.

[8] More examples of variation will be included in Section 4.

[9] "Lexikalische Solidaritäten" (Coseriu, 1967).

[10] Quasi-phrasemes or quasi-idioms (Mel'cuk, 1998).

➢ Compound nouns: (i) the meaning does not occur in any other combination (*sorriso amarelo* lit. 'yellow smile' → yellow = wry); (ii) the meaning may occur in different combinations (*café fresco* 'fresh coffe', *pão fresco* 'fresh bread' → fresh = recent); (iii) periphrastic nouns[11] (*continente negro* 'black continent' = Africa); (iv) entities (*dama de ferro* 'iron lady'). Apart from inflection, some expressions are subject to lexical and syntactic variation, namely insertion of modifiers (*sorriso muito amarelo* lit. 'smile very yellow'), alternation between simple elements and elements with suffixes (*sorrisinho amarelo* lit. 'little yellow smile') and alternation between adjectival and prepositional modifiers (*silêncio mortal* 'deadly silence', *silêncio de morte* 'silence of death'). Entities are fixed.

Expressions with Total Idiomatic Meaning

➢ Expressions transposed to another semantic field by metaphoric process (*balde de água fria* 'cold shower', *faca de dois gumes* 'double-edge knife', *esticar o pernil* 'kick the bucket', *deitar água na fervura* 'pour oil on troubled waters', *a sangue frio* 'in cold blood'). Adverbial expressions are fixed. Some of the nominal and verbal structures may undergo lexical and syntactic variation, such as substitution (*arma/espada/pau de dois gumes* 'double-edge weapon/sword/stick'), insertion of modifiers (*deitar mais água na fervura* 'pour more oil on troubled waters'), permutation (*estar de mãos e pés atados* 'bound hand and foot', *estar de pés e mãos atados* 'bound foot and hand' (helpless)) and occurrence both in negative and affirmative sentences (*ter olhos na cara* lit. 'have eyes in the face' (put things in perspective), *não ter olhos na cara* lit. 'do not have eyes in the face').

➢ Compound nouns (*flor de estufa* 'greenhouse plant' (delicate person); *mão de ferro* 'iron fist'). Apart from inflection, we observed alternation between simple elements and elements with suffixes.

➢ Proverbs (*grão a grão enche a galinha o papo* lit. 'grain by grain the hen fills its belly' (little strokes fell great oaks)). As in compositional proverbs, we also observed lexical variation (*grão a grão enche muita gente o papo* lit. 'grain by grain lots of people fill their bellies').

In what idiomatic expressions are concerned, it is important to note the fact that the transposition of an expression to another semantic field is a synchronic process that usually implies that at some point in time (including the present day) the expressions may simultaneously present compositional and idiomatic meanings (*porto de abrigo* 'harbor'; 'safe haven'). Curiously, from a statistical point of view, our study showed that the idiomatic meaning is the one that usually presents high frequency of occurrence. This information, together with the interpretation of the context, may help the automatic systems to decide whether they face a compositional or idiomatic expression.

In a sweeping look at the data, we observed that MWE show particular properties according to their syntactic pattern. Thus, at the sentence level (proverbs and aphorisms), MWE usually do not accept syntactic changes (the possible change seems to be lexical, when speakers substitute one or more elements), while verb phrases admit much more morphossyntactic variation. Noun phrases, on the other hand, raise specific issues. Compositional groups can behave as idiomatic ones and it is not always easy to distinguish them. The modifiers of the noun can express different semantic relations (part of, made of, used for) that may interact with the meaning (literal or idiomatic) of the noun.

4 Annotation of the Lexicon

The information presented on our typology will allow us to enrich the lexicon mentioned in Section 2. Our purpose is to have each MWE entry in the lexicon labeled 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 MWE parts (e.g., obligatory, optional, free).

As we have seen before, MWE have different types of variation for which we have to account for. We will briefly discuss our proposal for handling the annotation of some cases of lexical and syntactic variation in the lexicon.

Lexical Variation

➢ Insertion of modifiers – lexical elements (usually with an emphatic function) that do not belong to the canonical form are not part of the MWE and

[11] Cf. Sanromán, 2000.

are not labeled (*sorriso muito amarelo* lit. 'smile very yellow').

➢ Lexical substitution – This variation is restricted to limited set of alternatives. This set is recorded in the MWE lexicon as '*obligatory parts of the MWE and member of a set list*' (*comer/vender/comprar/impingir/levar gato por lebre* lit. 'eat/sell/buy/impose/take a cat instead of a hare' (buy a pig in a poke)).

➢ Free lexical elements – 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').

There are also cases where parts of the MWE may freely vary, while other parts remain fixed (*a educação é a mãe de todas as civilizações* 'education is the mother of all civilizations', *a liberdade é a mãe de todas as virtudes* 'liberty is the mother of all virtues'). These cases are treated likewise (*ALGO é a mãe de todas as NOUN-PL* 'something is the mother of all NOUN-PL')

Also, since creative use of language can lead to MWEs that only partly match the canonical MWE (cf. proverbs), we label these parts as '*different from canonical form*'.

Syntactic Variation

➢ Pronouns/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 mãos dele/estar nas suas mãos* 'to be in the hands of him'/'to be in his hands').

➢ From active to passive voice – Auxiliary verbs are not labeled as part of the MWE (*passar ALGO a pente fino/ALGO foi passado a pente fino* lit. 'pass something with a fine toothcomb'/'something was passed with a fine toothcomb' (to scrutinize)).

According to Hendrickx et al. (2010a), this annotated lexicon could be the basis for the annotation of idiomatic MWE in running text[12]. 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. 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.

5 Conclusion

This paper has shown the ongoing research that aims to describe, as detailed as possible, the syntactic and semantic properties of different types of Portuguese MWE. During our analysis, we encountered two major problems: (i) the evaluation of the meaning of certain expressions (compositional or idiomatic); (ii) the attempt to account for all possible lexical and syntactic variation. The information obtained from the typology will be used to annotate a MWE lexicon. Having a resource with such information (that includes additional meanings, possible variation that accounts for obligatory and optional elements, etc.) will be of extreme value for the development and evaluation of automatic MWE identification systems.

References

Baptista Jorge. 1994. *Estabelecimento e Formalização de Classes de Nomes Compostos*. MA Thesis, Faculdade de Letras da Universidade de Lisboa, Lisbon.

Chacoto Luísa. 1994. *Estudo e Formalização das Propriedades Léxico-Sintácticas das Expressões Fixas Proverbiais*. MA Thesis, Faculdade de Letras da Universidade de Lisboa, Lisboa.

Church Kenneth and Patrick Hanks. 1990. Word Association Norms, Mutual Information and Lexicography. *Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics*. Vancouver, Canada, pp. 76-83.

Coseriu Eugenio. 1967. Lexikalische Solidaritäten. *Poetica 1*. pp. 293-303.

Duran Magali Sanches, Carlos Ramish, Sandra Maria Aluísio and Aline Villavicencio. 2011. Identifying and Analyzing Brazilian Portuguese Complex Predicates. *Proceedings of the Workshop on Multiword Expressions*. Association for Computational Linguistics. Portland, Oregon, USA, pp. 74-82.

[12] The authors' approach is to annotate CINTIL corpus, a 1M word corpus of both spoken and written data from different sources that has been previously annotated with linguistic information such as part-of-speech, lemma, inflection, proper names, etc. (http://cintil.ul.pt/pt/).

Evert Stephan and Brigitte Krenn. 2001. Methods for the Qualitative Evaluation of Lexical Association Measures. *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*. Toulouse, France, pp. 188-195.

Fellbaum Christiane. 1998. *An WordNet Electronic Lexical Database*. The MIT Press, Cambridge, MA.

Firth R. John. 1955. Modes of meaning. *Papers in Linguistics 1934-1951*. London, Oxford University Press, pp. 190-215.

Hendricks Iris, Amália Mendes and Sandra Antunes. 2010a. Proposal for Multi-word Expression Annotation in Running Text. *Proceedings of the fourth Linguistic Annotation Workshop.* Association for Computational Linguistics. Uppsala, Sweden, pp. 152-156.

Hendricks Iris, Amália Mendes, Sílvia Pereira, Anabela Gonçalves and Inês Duarte. 2010b. Complex Predicates annotation in a corpus of Portuguese. *Proceedings of the fourth Linguistic Annotation Workshop.* Association for Computational Linguistics. Uppsala, Sweden, pp 100-108.

Jackendoff Ray. 1997. *The Architecture of the Language Faculty*. The MIT Press, Cambridge, MA.

Macário Lopes Ana Cristina. 1992. *Texto Proverbial Português: elementos para uma análise semântica e pragmática*. PhD Dissertation, Universidade de Coimbra, Coimbra.

Mel'čuk Igor. 1998. Collocations and Lexical Functions. Cowie, A. P. (ed.), *Phraseology. Theory, Analysis, and Applications*. Oxford University Press, Oxford, pp. 23-53.

Mendes Amália, Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João M. Casteleiro, Luísa Pereira and Tiago Sá. 2006. COMBINA-PT: A Large Corpus-extracted and Hand-checked Lexical Database of Portuguese Multiword Expressions. *Proceedings of the Fifth International Conference on Language Resources and Evaluation*. Genoa, Italy, pp. 1900-1905.

Pereira Luísa and Amália Mendes. 2002. An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications. Braasch, A. and C. Povlsen (eds.), *Proceedings of the 10th International Congress of the European Association for Lexicography*. Copenhagen, Denmark, vol. II, pp. 841-849.

Ranchhod Elisabete. 2003. O Lugar das Expressões 'Fixas' na Gramática do Português. Castro, I. and I. Duarte (eds.), *Razões e Emoção. Miscelânea de Estudos oferecida a Maria Helena Mira Mateus*. Imprensa Nacional Casa da Moeda, Lisboa, pp. 239-254.

Sag Ivan, Timothy Baldwin, Francis Bond, Ann Copestake and Dan Flickinger. 2002. Multiword Expressions: A Pain in the Neck for NLP. Gelbukh A. (ed.), *Proceedings of the Third International Conference on Intelligent Text Processing and Computational Linguistics*. Mexico City, Mexico, pp. 1-15.

Sanromán A. Iriarte. 2000. *A Unidade Lexicográfica. Palavras, Colocações, Frasemas, Pragmatemas*. PhD Dissertation, Universidade do Minho, Braga.

Sinclair John. 1991. *Corpus, Concordance and Collocation*. Oxford University Press, Oxford.

Vilela Mário. 2002. *Metáforas do Nosso Tempo*. Almedina, Coimbra.

Zeller Britta and Sebastian Padó. 2012. Corpus-Based Acquisition of Support Verb Constructions for Portuguese. *Proceedings of the 10th International Conference on Computational Processing of the Portuguese Language*. Coimbra, Portugal, pp. 73-84.