Proposal for Multi-Word Expression Annotation in Running Text

Iris Hendrickx, Amália Mendes and Sandra Antunes
Centro de Linguística da Universidade de Lisboa, Lisboa, Portugal
{iris, amalia.mendes, sandra.antunes}@clul.ul.pt

Abstract

We present a proposal for the annotation of multi-word expressions in a 1M corpus of contemporary portuguese. Our aim is to create a resource that allows us to study multi-word expressions (MWEs) in their context. The corpus will be a valuable additional resource next to the already existing MWE lexicon that was based on a much larger corpus of 50M words. In this paper we discuss the problematic cases for annotation and proposed solutions, focusing on the variational properties of MWEs.

1 Introduction

Given the widespread studies of co-occurring words phenomenon, the term 'multi-word expression' (MWE) usually refers to a sequence of words that act as a single unit, embracing all different types of word combinations. Their study is of extreme importance for computational linguistics, where applications find notorious difficulties when dealing with them (Sag et al., 2002).

Having a well-balanced corpus annotated with multi-word expressions offers the possibility to analyze the behavior of MWEs as they appear in running text. Such corpus will contain a rich and diversified set of MWE and also be an excellent resource to evaluate automatic MWE identification systems. Here we propose our approach to the manual annotation of the CINTIL corpus (Barreto et al., 2006) with MWE information. This Portuguese corpus of 1M tokens is a balanced corpus of both spoken and written data from different sources and has been previously annotated with linguistic information such as part-of-speech and lemma and inflection.

As the starting point for our annotation project, we want to use a Portuguese MWE lexicon containing approximately 14,000 entries. The lexicon contains besides idiomatic expressions, also many collocations: expressions of frequently co-occurring words that do not show syntactic or semantic fixedness. We are mostly interested in the idiomatic expressions and will only mark up these in the corpus.

2 Related Work

There is already quite some work about the creation and representation of MWE lexicons (Baldwin and Kim, 2010). Most of the currently available corpora annotated with MWE information consist of a collection of extracted sentences containing a MWE (for example the data sets in the MWE 2008 shared task[1]). Fellbaum et al. (2006) report on a larger German example corpus consisting of MWEs with their surrounding sentences. There are also data sets specifically designed for automatic MWE identification, in which part of the sentences contains an idiomatic expression and the other part expresses a literal meaning (e.g. (Sporleder and Li, 2009)). An example of a balanced corpus fully annotated with MWEs is the Prague Treebank which is enriched with a diverse set of MWE annotations (Böhmová et al., 2005).

3 MWE Lexicon

Our annotation proposal uses information from a lexicon of MWE for Portuguese (available online[2]). This lexicon is implemented on a MySQL relational database. The MWEs were extracted from a 50M words balanced corpus of Portuguese. The MWE are organized under canonical forms. Also inflectional variations of the 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

[1] More infomation at: http://multiword.sourceforge.net/

[2] MWE lexicon: http://www.clul.ul.pt/sectores/linguistica_de_corpus/manual_combinatorias_online.php

Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010, pages 152–156,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

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; Antunes et al., 2006; Bacelar do Nascimento et al., 2006).

4 Proposed annotation

In this section we discuss our approach to the annotation of MWEs in the corpus.

4.1 Typology

We want to classify each idiomatic MWE occurring in the CINTIL corpus according to a typology that expresses the typical properties of the MWE. Although the lexicon of MWEs covers a wide range of units, from idiomatic expressions to collocations, we decided to restrict our annotation of the corpus to cases of idiomatic MWEs because those are the problematic ones for any task of semantic annotation and disambiguation. The MWE lexicon does not provide labels for idiomatic vs. compositional expressions, so this information will have to be added during the annotation task. Identifying idiomatic MWEs is not a simple task. For clear cases of idiomatic units, the global meaning can not be recovered by the sum of the individual meanings of the elements that compose the expression.

In other cases, only part of the MWE has an idiomatic meaning, while one or more of the elements are used in their literal meaning (e.g *saúde de ferro* ‘iron health’). Deciding if one of the elements of the MWE is literal or not depends in fact of our definition of literal: if we consider it to be the first prototypical meaning of a word, this very restrictive definition will trigger us to label a large number of MWEs as idiomatic. Other MWEs are compositional but receive an additional meaning, like *cartão vermelho* in football, which is literally a red card but has an additional meaning of punishment.

We want to cover these different cases in our annotation, and to establish a typology that takes into account morpho-syntactic and semantic aspects of the MWE: its functional part-of-speech (PoS) category, the PoS categories of its internal elements, its fixed or semi-fixed nature, its global or partial idiomatic property and motivation, and possible additional meanings.

4.2 Division by syntactic category

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 and to establish different properties that enables us to distinguish literal from idiomatic usage. 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’) have specific properties: they do not accept any possible syntactic changes like passivization or relativization, they do not accept any inflectional variation, the only possible change is lexical (when speakers substitute one or more elements, like we will discuss in section 4.4). However fixed, the meaning of this example is clearly motivated and compositional in the sense that it is recovered by the meaning of the individual elements. On the contrary, MWEs which are verb phrases will admit much more morpho-syntactic variation. Moreover, noun phrases raise specific issues: the most syntactically fixed units will be very close or identical to compound nouns. For example, the meaning of the prepositional modifier of the noun can be literal but the overall expression will still be used as a compound and will denote a very specific entity, frequently from domain-specific languages (*projecto de lei* ‘project of legislation’, *contrato de compra e venda* ‘sell contract’). Moreover, the prepositional and adjectival modifiers of the noun will express many different semantic relationships (part of, made of, used for) which interact with the meaning (literal or idiomatic) of the noun (Calzolari et al., 2002). Establishing specific guidelines for these different types of MWEs will enable a more accurate annotation. To decide upon the difficult cases of idiomatic and non-idiomatic usage, we plan to use the intuitions of different annotators.

4.3 Linking to MWE lexicon

We will annotate each encountered MWE in the corpus with a link to the MWE-entry in the lexicon, instead of labelling each MWE with its typology. This way we link each MWE to its canonical form and other additional information. Moreover, we can easily gather all occurrences of one particular canonical MWE and check its variation in the corpus. It will also allow us to work with a more detailed typology and will give us the possibility to revise it during the annotation process. It might be difficult to establish beforehand very precise guidelines that will apply to all the MWEs and even to all the MWEs of a specific subtype. Often, guidelines are constantly in need of revision as we encounter slightly different contexts who challenges decisions previously taken.

The corpus annotation will enable us to extend the information in the MWE lexicon with typology labels regarding the whole expression (function, idiomatic meaning) but also regarding individual words of the expression as to whether they are obligatory or not.

We plan to add a meaning to idiomatic expressions using a dictionary. We expect that MWEs will be unambiguous: they have the same meaning each time they are used. In some cases, the synonym or paraphrase proposed for the MWE might not be able to replace the MWE in the corpus context. For example, the MWE *às mãos cheias* means *em grande quantidade* 'in large quantity', but this meaning can not always replace the MWE in context.

The annotation process of fully fixed expressions could be retrieved automatically. For the variable expressions we will combine automatic retrieval with manual validation, Here the automatic retrieval step will aim for a high recall and select all sentences that contain the lemmas of the MWE. Without doubt our corpus will contain many MWEs that are not yet listed in the MWE lexicon. Therefore each sentence will need to be checked manually for MWEs. We can create the links between the lexicon and MWEs in the corpus automatically, but again, as not all MWEs will occur in the lexicon, we will need to do a manual validation of the automatic labelling and also add newly discovered MWEs to the lexicon.

4.4 MWE Variation

Corpus analysis clearly shows that MWEs have different types of internal variation. 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. Mapping MWE occurrences in the corpus to their canonical form can be a hard task depending on the flexibility of the MWE. In the next part we discuss our proposal how to handle the annotation of several types of variation in MWEs: lexical, syntactic and structural variation, lexical insertions and truncation of MWEs.

4.4.1 Lexical diversity

MWEs have a wide range of lexical variation and it can apply to any type of grammatical category, although we do notice that verb variation is the commonest type. Studying the lexicon showed us that there is a group of cases in which a word in a MWE can only be replaced by another word from a very limited set (usually not larger than 10 words) of synonyms or antonyms. For these cases this set is already recorded in the MWE lexicon. We mark these variable words as: 'obligatory parts of the MWE and member of a specified list'. In 1 we show an example: the canonical form followed by a sentence containing this MWE and the English translations.

Many MWEs also contain parts that are almost lexically free or only restricted to a semantic class such as person or named entity. These elements are represented in the MWE lexicon with a pronoun (e.g. *alguém*, *algum* ('someone', 'something')) or the tag *NOUN* (with possible gender/number restrictions) when a pronoun cannot substitute the free part. When marking up these elements in the corpus, we will label them with a reference to the pronoun used in the canonical form (example 2).

(1) **dizer/ sair** *da boca para fora*
(to say / to get out from the mouth outside)
Arrependeu-se com o que lhe **saiu** *da boca para fora*
'She regretted her slip of the tongue'

(2) *estar nas mãos de* **ALGUÉM**
A nossa vida *está nas mãos de* **Deus**
'Our life is in the hands of God'

MWEs are not always contiguous: it is frequent to encounter insertion of lexical elements which do not belong to the canonical form of the MWE. Often, the function of the inserted elements is adverbial, quantificational or emphatic. Or the MWE occurs in a negative context, by the insertion of the adverb *não*. Such inserted elements that are not part of the MWE are not labelled. This is the case of the quantifier `muitas` in (3), which is not part of the canonical form of the MWE *dar voltas à cabeça* 'to think'.

(3) *Dei* `muitas` *voltas à cabeça* para encontrar uma solução.
'I've been thinking a lot to find a solution.'

Another type of MWE variation is truncation: only a part of the full expression is lexically realized. This phenomenon usually occurs with proverbs and sayings. For example in 4 the bracketed part was not realized in the sentence, but it is part of the canonical form in the MWE lexicon. When marking up such truncated expressions we do not label explicitly this phenomenon, we just mark up the occurring part with a reference link to MWEs in the lexicon.

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

4.4.2 Syntactic variation

An obvious form of syntactic variation is inflection of verbs and nouns. Since Portuguese is a highly inflectional language, practically all the verbs that occur in MWEs inflect, except for some fixed sayings. Also shifting from active to passive voice leads to syntactic variation. We do not label auxiliary verbs as part of the MWE.

Several MWEs that have a free part such as example 2 do not only exhibit lexical variation but also syntactic variation: pronominalization (*estar nas mãos dele*) or with a possessive form (*estar nas suas mãos*). In such cases we will mark up possessives as part of the MWE but give them an additional label to signal that they are optional elements. However, possessives are not always optional, sometimes it is an obligatory part of the canonical form and we will annotate it normally (e.g. *o leão mostra a sua raça.* 'the lion shows what he's made off').

Also permutations of the MWE can occur (ex.5). We do not signal this phenomenon in our annotation as this can easily be detected when comparing to the canonical form.

(5) *estar de mãos e pés atados / estar de pés e mãos atados*
'to be tied hand and foot/ foot and hand'

4.4.3 Structural variation

True idioms are both semantically and syntactically fixed. However, language use is creative and can lead to MWEs that only partly match the 'real' MWE as listed in the MWE lexicon. For these cases we mark up the different part with an extra label to clarify which part exactly varies. For example 6.

(6) *no poupar é que está o ganho*
in the saving is the profit
no **esperar / provar / comparar** *é que está o ganho*
in waiting / proving / comparing is the profit

(7) já *dei voltas* **e voltas** *à cabeça*
'thoughs went on and on in my mind'

(8) **ALGO** *é a mãe de todas* **NOUN-PL**
'something is the mother of all x'
a educação *é a mãe de todas* **as civilizações**
a liberdade *é a mãe de todas* **as virtudes**
'education is the mother of all civilizations'
'freedom is the mother of all virtues'

Another interesting case is shown in example 7 in wich a part of the MWE is duplicated for emphasis. This should be treated differently than the example in 3. In these cases we will label the duplicated part as 'part of the MWE but optional' (similar to possessives).

There are cases in which part of the MWE may vary without any apparent limits, while the other part remains fixed. An example can be found in 8. These are actually just an extension of ones we already discussed (see example 2) and we treat them in the same matter.

5 Conclusion

In sum, we propose to split the annotation of MWEs to develop separate annotation guidelines for the grammatical categories, as we have observed that e.g. nominal MWEs behave differently than verbal MWEs. Each MWE in the running text will be linked to its canonical form in the lexicon. The lexicon itself will be enhanced with additional information such as typology information and MWE meaning. Special elements of the MWE such as optional or variable parts will be explicitly marked as such both in the lexicon and in the annotation of the MWE in the corpus. We are convinced that the implementation of our proposal will lead to a rich new resource that can help us study the behavior of MWE in more depth. We also plan to use this resource for the development and evaluation of automatic MWE identification systems.

References

S. Antunes, M. F. Bacelar do Nascimento, J. M. Casteleiro, 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.

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.

F. Barreto, A. Branco, E. Ferreira, A. Mendes, M. F. P. Bacelar do Nascimento, F. Nunes, and J. Silva. 2006. Open resources and tools for the shallow processing of portuguese. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006)*, Genoa, Italy.

A. Böhmová, S. Cinková, and E. Hajičová. 2005. A Manual for Tectogrammatical Layer Annotation of the Prague Dependency Treebank (English translation). Technical report, ÚFAL MFF UK, Prague, Czech Republic.

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)*, pages 1934–1940, Las Palmas, Spain.

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.

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*, pages 1900–1905, Genoa, Italy.

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*.

C. Sporleder and L. Li. 2009. Unsupervised recognition of literal and non-literal use of idiomatic expressions. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 754–762, Athens, Greece, March. Association for Computational Linguistics.