A Lexical Database for the Analysis of Portuguese MWEs

Sandra Cristina dos Santos Antunes
Center of Linguistics of the University of Lisbon, Lisbon, Portugal
sandra.antunes@gmail.com

Abstract. We present an Access database that will help to analyze and classify Portuguese MWEs according to their linguistic properties. We used a subset of an existing lexicon of Portuguese MWEs that was extracted from a corpus using Mutual Information as a statistical measure, followed by manual validation. The lexicon is organized in a three-level structure: the main lemmas, the group lemmas and the variants of those groups. This subset was imported to the database together with linguistic information about the MWEs and their variants (morphosyntactic structure, syntactic category, frequency, MI, grammatical function, discursive function, etc.). A semantic and syntactic fine-grained typology was established, and, by selecting a particular MWE, we can classify it with respect to its degree of semantic decomposability and syntactic transformation. The database is highly customizable and enables the addition/deletion of semantic or syntactic categories considered important throughout the analysis. In the end, all the information of the database will be exported to a XML format, resulting in a lexicon enriched with linguistic information.

Keywords: Access database, classification of MWE, enriched lexicon.

1 Introduction

This paper presents an on-going work that aims at the analysis, classification and annotation of Portuguese lexical multiword expressions (MWEs). This term is understood as an umbrella term, since it includes any sequence of two or more graphic words that present a high syntactic and/or semantic cohesion, embracing, therefore, different types of word combinations (collocations, compounds, formulae, light-verb constructions, idioms, similes, sayings, clichés, etc.).

It is widely known that these expressions play a crucial role in language ([7], [12], [16], [18]). But, despite the fact that their analysis is being intensely carried out in several linguistic areas, they still pose problems regarding their accurate identification. It is not easy to distinguish different types of expressions, which has been proven by the large quantity of typologies in the literature ([2], [5], [9], [10], [15], [18], [22]).

Considering the Portuguese language, there is a major gap regarding the proper classification of these expressions. In order to fill that gap, we developed a database in Access format that imports a subset of a Portuguese MWEs lexicon (COMBINA-PT) and enables the classification of each expression according to its linguistic properties. In the application, each MWE is described according to a comprehensive set of
semantic and syntactic categories. The available set of search functions, using filters, enables the extraction of linguistic regularities.

This paper will briefly present the compilation of the lexicon and the methodology adopted for the MWEs selection (section 2), the database and the criteria used for the MWEs classification (section 3), and the practical applications of the tool (section 4).

2 The lexicon

For this work, we used COMBINA-PT, which is a lexicon of significant lexical word combinations of European Portuguese\(^1\). The MWEs were extracted from a balanced\(^2\) 50 million word written corpus\(^3\) using Mutual Information (MI) \([4]\) as a statistical association measure\(^4\), followed by manual validation. Table 1 presents the corpus constitution from which the MWEs were extracted.

<table>
<thead>
<tr>
<th>Table 1. Corpus constitution.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Newspapers</td>
</tr>
<tr>
<td>Books</td>
</tr>
<tr>
<td>Magazines</td>
</tr>
<tr>
<td>Miscellaneous</td>
</tr>
<tr>
<td>Leaflets</td>
</tr>
<tr>
<td>Supreme Court Verdicts</td>
</tr>
<tr>
<td>Parliament Sessions</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
</tr>
</tbody>
</table>

As described in \([17]\), n-grams of 2, 3, 4 and 5 tokens were extracted from the corpus. Sequences constituted by 3 to 5-grams are contiguous, while 2-grams sequences can be either contiguous or separated by a maximum of 3 tokens.

Considering the large candidate list extracted from the corpus (1.7 million MWEs), it was necessary to hand-check only a subpart of the groups. Following previous studies ([6], [19]), the team firstly selected groups with MI values between 8 and 10, since there is a higher concentration of good candidates around those values. Throughout manual validation, we followed several criteria upon which usually relies the definition of a MWE: lexical and syntactic fixedness, semantic cohesion, frequency of occurrence (which reveals sets of favoured co-occurring forms) and grammatical constituency (only completed constituents were selected).

\(^1\) The lexicon is available at Meta-Share repository: http://www.meta-net.eu/meta-share.

\(^2\) The corpus covers a wide range of textual genres representing, in a proportional way, the language usage.

\(^3\) This corpus was extracted from the Reference Corpus of Contemporary Portuguese, a written and spoken monitor corpus, in a total of 311M words: http://www.clul.ulisboa.pt/en/10-research/713-crpc-reference-corpus-of-contemporary-portuguese

\(^4\) The choice of MI relied on the fact that it is reported to differentiate between MWE and non-MWE \([28]\).
The MWEs in the lexicon are organized in order to identify: (i) a main lemma, from which the MWE was selected (e.g., mistério ‘mystery’); (ii) a group lemma, i.e., the neutral expression that corresponds to the canonical form (e.g., revelar o mistério ‘unravel the mystery’); (iii) all the variants that occurred in the corpus (e.g., revelar este mistério ‘unravel this mystery; o mistério foi revelado ‘the mystery was unravelled’). Concordances lines for each MWE are also available in KWIC format.

In all, the lexicon comprises 1,180 main lemmas, 14,153 group lemmas and 48,154 variants.

3 The database

In contrast to languages for which there is a wide range of studies regarding MWEs, for Portuguese little work has been done so far. Most of the studies pay particular attention to idiomatic expressions ([3], [14], [28]), compounds ([11], [21], [26]) or light verbs [11], while other types of expressions are classified taking into account their morphosyntactic structure [20].

Given the existence of different types of MWEs (with different degrees of syntactic and semantic cohesion), as well as the difficulty in distinguishing between certain expressions (such as free combinations, collocations and compounds), we created a database that will help to analyze and classify these groups taking into account their lexical, grammatical, syntactic, semantic, pragmatic and discursive properties.

Due to the large amount of COMBINA-PT data, we decided to firstly analyze only a part of the lexicon. At the moment, the database contains 169 main lemmas, 3,230 group lemmas and 1,503 variants. The database contains all the information considered important for the analysis of the MWEs: (i) the morphosyntactic structure; (ii) the syntactic category; (iii) the variants (if any) and their morphosyntactic structure; (v) the frequency of the expression and the variants; (vi) the MI; (vii) the grammatical function; (viii) the discursive function [18]; (ix) the definition (important for idiomatic expressions); (x) a concordance line; (xi) indication of presence/absence of the expressions in a Portuguese reference dictionary [30] (since Portuguese lexicographers do not make full use of corpora (they only use quotations from literary texts), it could be interesting to see to what extent the dictionary entries meet corpus data).

Lemmas that pertain to different grammatical classes are registered in different entries with different numbers: frio1N, frio2ADJ ‘cold’. The same is true for polysemous MWE, i.e., that can have several senses (such as sinal verde ‘green light’, which can be either a traffic sign or an indication of approval): sinal verde1, sinal verde2.

Figure 1, below, shows how the database is structured in a three-level structure (on the left): main lemma, MWEs for each lemma and variants of each MWE. The selection of a main lemma displays the set of related MWE; by clicking on the “+” sign of a MWE, all the variants of that expression will be displayed.

By selecting a particular MWE, all the information regarding that expression will be disclosed (on the right). For the semantic and syntactic analysis, we established a typology, presented in a tree diagram format, with checkboxes that must be selected according to their degree of semantic decomposability and syntactic transformation.
Regarding the semantic analysis, and following [2], [5], [10], [15] and [22], we firstly considered a scale of idiomaticity, composed by three major classes: (i) compositional meaning; (ii) partially idiomatic meaning (at least one of the elements keeps its literal meaning); (iii) total idiomatic meaning. Within each of these three semantic categories, the MWEs are also classified regarding their type: collocations, compounds, formulae, light verbs, idioms, similes, sayings and clichés. A fine-grained distinction can also be made inside some of these types: collocations are subdivided into favoured co-occurring forms and expressions with restricted collocability. This helps to visualize which type of MWE may occur in different semantic categories (collocations, compounds, similes and sayings are spread amongst compositional and idiomatic levels), highlighting the process of lexicalization of some expressions. Although we are trying to draw this dividing line, we are aware that the semantic evaluation of certain expressions may be difficult, not allowing accurate divisions.

From the syntactic standpoint, the observation of the variants of each MWE will help its analysis regarding the lexical and syntactic fixedness [18]. Each expression can be classified as: (i) fixed (no variation); (ii) semi-fixed (nominal/verbal inflection); (iii) with variation. The variation field is, in turn, subdivided into lexical (permutation, replacement of elements, reduction, etc.), syntactic (passivization, relativization, pronominalization, nominalization, insertion of elements, etc.) and structural (free realizations, exploitations, etc.). It is also known that corpus data may not cover all the possibilities of the language system [24]. For that reason, it is possible to add variants of a MWE that were found in other corpora (duly identified), the internet or through introspection, confirming their lexical-syntactic variation/fixedness.

The database is highly customizable and enables the addition/deletion of semantic or syntactic categories that we may consider important throughout the analysis.

Finally, all the information of the database will be exported to a readable format (XML), resulting in a lexicon enriched with linguistic information ([8], [13], [27]).

4 Applications

This new database will facilitate the encoding of the linguistic properties of MWE, providing a lexicon with several layers of information. As an Access database, it is possible to apply filters and to create reports that enumerate all the criteria that match for any category of MWEs. Filters can be applied to any field and all the combinations are possible. We believe that the resulting lexicon will constitute a value for areas such as: (i) lexicography (the lexicographer can consult the MWEs actually produced by native speakers, and easily observe that different collocates may point towards different meanings of a word [25]); (ii) foreign language acquisition (allowing the production of natural-sounding speech and writing); (iii) natural language processing (helping the development of automatic identification systems); (iv) theoretical linguistics (expediting semantic and syntactic proposals).

Since Portuguese is a highly inflectional language, practically all the verbs and nouns that occur in MWEs show inflection marks.
Fig. 1. MWE for the lemma *ferro* ‘iron’ and classification of *panela de ferro* ‘iron cauldron’.

References