Workflow of a Portuguese dictionary of multiword expressions

Amália Mendes, Sandra Antunes, Luísa Pereira, Fernanda Bacelar do Nascimento
Centre for Linguistics at the University of Lisbon
E-mail: {amalia.mendes, sandra.antunes, luisa.alice.sp}@clul.ul.pt

1 Introduction

The project Word Combinations in Portuguese Language (COMBINA-PT) consisted of a dictionary of Portuguese multiword expressions (MWE) automatically extracted through the analysis of a balanced 50 million word corpus, statistically interpreted with lexical association measures and validated by hand (Mendes et al, 2006; Antunes et al, 2006). The availability of large amounts of textual data and the development of corpus-based approaches enable the identification and analysis of complex patterns of word associations, proving that 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), Sinclair (1991)). We discuss in section 2 the workflow associated to the development of this resource, following the different stages identified in Klosa (2013), and in section 3 how this workflow reflects in the time span of the project.

2 Lexicographical workflow

2.1 Preparation

The preparation phase started with the elaboration of the submission for external funding and extended during the first two months of the project, taking up a total of three months. During this preparation phase, we established what would be our approach to the extraction of MWE.

The project relied on the availability of the Reference Corpus of Contemporary Portuguese (CRPC), a large corpus of different varieties of Portuguese, covering different genres (now 312 million words). A balanced subcorpus of CRPC had to be designed (cf. 2.2) and this primary source would be the only source of data in the project (no dictionaries or lexica were used).

We decided to follow a corpus-driven perspective and not to establish specific syntactic patterns (e.g. noun-preposition-noun) for our query. These patterns would then emerge from the set of selected units. This would lead to a wide coverage of types of MWE: totally frozen groups, semi-frozen or just sets of favoured co-occurring forms, that constitute however syntactic dependencies, idioms, named entities.

Since the exact definition of a collocation and how it differs from other MW expressions is known as a challenging issue (discrete categorization is difficult to apply to concepts defined in terms of degree of fixedness, compositionality, substitutability, etc.), we decided that, at a first stage of the work, we would select all the expressions that presented some syntactic and semantic cohesion, without attempting to follow any prior typology.

At this stage, we also defined that Mutual Information would be the lexical association measure to apply to our candidate set (Church & Hanks 1990) and that we would consider n-grams from 2 to 5 grams: groups of 2, 3, 4 and 5 tokens (groups of 2 tokens can be contiguous or be separated by a maximum of 3 tokens, while groups of more than 2 tokens are contiguous). Contexts would be retrieved for each n-gram.

We also decided to roughly label the selected MW according to 4 main classes: (i) MW expressions that may or may not occur with an hyphen (e.g., *chapéu de chuva* ‘umbrella’; *casa de banho* ‘bathroom’); (ii) MW expressions that refer to named entities, such as institutions, functions, etc. (e.g., *União Europeia* ‘European Union’, *Presidente da República* ‘President of the Republic’); (iii) expressions that constitute verbal phrases (e.g., *respirar fundo* ‘to breathe deeply’) nominal phrases (e.g. *ar puro* ‘fresh air’) or adjectival phrases (e.g. *absolutamente indispensável* ‘absolutely indispensable’); iv) MW expressions that require further attention, either because they are seen as doubtful cases or either because the expression exceeds 5 tokens (the limit extracted by the tool) and needs to be correctly identified during the lemmatization process.

Beside the decision about what to extract and what to consider as a MWE, we took decisions regarding the tools that were required and the computational support.

2.2 Data acquisition

The COMBINA-PT corpus is a balanced 50,8M word written corpus extracted from the Reference Corpus of Contemporary Portuguese (CRPC). CRPC is a written and spoken monitor corpus (cf. Sinclair, 1991), compiled at CLUL since 1988 and that comprises all the national and regional varieties of Portuguese, in a total of 311 million words[1].

Given the importance of corpus balance (particular word may co-occur with different lexical units according to the type of discourse in which it occurs), the corpus includes different text types, as showed in table 1. The compilation of this corpus out of the larger CRPC started during the preparation stage (design) and continued afterwards (see timeline in section 3).

[1] http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc

CORPUS CONSTITUTION			
Newspapers			**30.000.000**
Books	Fiction	6.237.551	
	Technical	3.827.551	
	Didactic	852.787	**10.818.719**
Magazines	Informative	5.709.061	
	Technical	1.790.939	**7.500.000**
Miscellaneous			**1.851.828**
Leaflets			**104.889**
Supreme court verdicts			**313.962**
Parliament sessions			**277.586**
TOTAL			**50.866.984**

Table 1: Constitution of the Corpus

2.3 Computerisation

A computer scientist was hired at the beginning of the project to implement an n-grams extractor tool, to run over the COMBINA corpus. The tool provided the following information for each n-gram:

- Distance (first number after the MW unit in bold);
- Number of elements of the group;
- Frequency of the group at a specific distance;
- Lexical association measure;
- Total frequency of the group in all occurring distances;
- Frequency of each element of the group;
- Total number of words in the corpus;
- Concordances lines (KWIC format) of the MWE in the corpus.

We then hired a second computer scientist to design database in MySQL for storing this information and provide an interface for the linguist task: selection of significant MWE, selection of valid concordance lines, organization of MWE to abstract from inflection and other variations.

The hiring of a computational linguist for the full time of the project would certainly have improved our performance and helped us to achieve a larger coverage in our resource.

2.4 Data processing

When analysing the first batch of data, and considering the large candidate list extracted from the corpus and the need of effective ways to reduce noise, several cut-off options were implemented to allow for the elimination of: (i) groups with internal punctuation; (ii) word pairs with first or final grammatical word using a stop-list (in case one wishes to rule out non-lexical associations); (iii) groups under a selected total minimum frequency. All of these options were then applied when running the extraction processing tool and a minimum frequency was established (3 for groups of 3 to 5 tokens; 10 for 2-token groups). Another aspect that required some attention was the presence in the corpus of some characters that disturbed the extraction of the n-grams.

These questions were considered during the preparation phase, but required the analysis of the data, after assuring the primary source, so that we could fully understand their implications. The first months of the project involved in fact an overlap of the preparation, computerisation and data processing phases (cf. Tiberius and Schoonheim (to Appear)).

2.5 Data analysis

Due to the large amount of data, we decided to proceed with the analysis of a set of lemmas (1180). These lemmas were selected based on their occurrence in MWE that had MI values between 8 and 10 (a manual survey of our total list of candidates showed that there was a higher concentration of good candidates in every frequency span around medium MI values of 7-12). This approach reduced our candidate list to 170,000 units. One linguist did the manual inspection and selection of significant MWE for each of these lemmas during one year.

A large set of MW expressions occur only in a specific word form, like the case of the nominal phrase *reparação de danos* ‘damage repair’, that do not occur in the plural form **reparações de danos* ‘damage repairs’. However, since Portuguese is a highly inflectional language, there is still a large number of MW expressions, specially those involving a verb, that do occur in different inflected forms, making it necessary to organize inflection variants under a single MWE canonical form. In cases of MWE with one free element, a pronoun (*someone*, *something*) is used to fill this place in the canonical form of the MWE. The same linguist, simultaneously to the selection phase, performed this task and also the selection of the valid concordance lines.

2.6 Preparation for online release

We exported the data of the MYSQL database as a report in html, that can be read online. A manual with all our criteria for the selection and organization into canonical forms is made available together with the data.

2.7 Afterlife

Our work selecting the MWE has pointed some drawbacks of the MYSQL database. We have gone back to the computerisation phase, although with different goals: a new version of the database was designed and implemented, and the list of MWE selected has been uploaded. This version will enable a detailed description of syntactic and semantic properties of each unit.

3 Time span of the different phases

The time span of the workflow is presented in Table 3. We see how several phases overlap at the beginning of the project and how we come back to a computerization task at the end.

Phase	**1**				**5**					**10**					**15**					**20**					**25**					**30**
Preparation	▨	▨	▨																											
Data acquisition	▨	▨	▨	▨																										
Computerisation		▨	▨	▨	▨	▨	▨	▨	▨																				▨	▨
Data processing						▨	▨	▨	▨	▨	▨	▨																		
Data analysis										▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨	▨				
Preparation for online release																											▨	▨	▨	▨
Afterlife																														

Table 3 Process phases of our project

4 References

Antunes, Sandra, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro, Amália Mendes, Luísa Pereira, Tiago Sá (2006) “A Lexical Database of Portuguese Multiword Expressions” in VIEIRA, R. et al. (2006) PROPOR 2006, LNAI 3960, Berlin, Springer-Verlag, pp. 238-243.

Church, K. W. & P. Hanks (1990) “Word association norms, mutual information, and lexicography”, *Computational Linguistics*, 16 (1), pp. 22-29.

Firth, J. (1955) "Modes of meaning", *Papers in Linguistics 1934-1951*, London, Oxford University Press, pp. 190-215.

Klosa, Annette (2013) The lexicographical process (with special focus on online dictionaries). In: Gouws, Rufus H./Heid, Ulrich/Schweickard, Wolfgang/Wiegand, Herberst Ernst (Hgg.): Dictionaries. An international Encyclopedia of Lexicography. Supplement Volume: Recent Developments with Focus on Electronic and Computational Lexicography. Berlin, Boston: de Gruyter, S. 517-524. (Handbücher zur Sprach- und Kommunikationswissenschaft; 5.4).

Mendes, Amália, Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro, Luísa Pereira, Tiago 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, Génova, 22-28 de Maio de 2006, pp. 1900-1905.

Sinclair, J. (1991) *Corpus, Concordance, Collocation*, Oxford, Oxford University Press.

Tiberius, C. and T. Schoonheim (To Appear) The *Algemeen Nederlands Woordenboek* (ANW) and its Lexicographical Process In: Vera Hildenbrandt (Ed.): Der lexikografische Prozess bei Internetwörterbüchern. 4. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie“. Mannheim: Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik X/XXXX).