A Lexical Database of Portuguese Multiword Expressions

Sandra Antunes, Maria Fernanda Bacelar do Nascimento, João Miguel Casteleiro,
Amália Mendes, Luísa Pereira and Tiago Sá

Centro de Linguística da Universidade de Lisboa (CLUL)
Av. Prof. Gama Pinto, 2, 1649-003 Lisboa, Portugal
{sandra.antunes, fbacelar.nascimento, amalia.mendes, luisa.alice, ptsa}@clul.ul.pt
miguel.casteleiro@zmail.pt

Abstract. This presentation focuses on an ongoing project which aims at the creation of a large lexical database of Portuguese multiword (MW) units, automatically extracted through the analysis of a balanced 50 million word corpus, statistically interpreted with lexical association measures and validated by hand. This database covers different types of MW units, like named entities, and lexical associations ranging from sets of favoured co-occurring forms with high corpus frequency and low cohesion to strongly lexicalized expressions with no, or minimum, variation. This new resource has a two-fold objective: to be an important research tool which supports the development of collocation typologies and their integration in a larger theory of MW units; to be of major help in developing and evaluating language processing tools able of dealing with MW expressions.

1. Introduction

Firth (1955) described a collocation as the characterization of a word according to the words that typically co-occur with it, showing that the meaning of a word is closely related to the set of co-occurring words and 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. And once they start to be frequently repeated, collocations tend to correspond to a conventional way of saying things, turning out to be an important aspect in the lexical structure of the language. However, these word associations are not immediately identified when one only relies on intuition-based studies. But the availability of large amount of textual data and the advance of computer technologies allowed the development of corpus-based approaches which enable the identification and analysis of complex patterns of word associations, proving that, in fact, natural languages follow regular patterns at syntagmatic level.

The identification and analysis of these associative patterns provide important information about the meaning of a word and its actual uses (Sinclair, 1991) and constitute an important resource for several areas, such as psycholinguistics (development of hypothesis about the representation of the individual mental lexicon, semantic memory and cognitive processes in general), lexicography (improvement of their coverage in modern dictionaries) and computational linguistics (helping to avoid overgeneration, idiomaticity and parsing problems (Sag et alii, 2002) that usually occur in NLP applications, such as machine translation, information extraction and retrieval, question-answering systems, language generation and word sense disambiguation).

Aiming to contribute to the study of MW expressions, this presentation will address an ongoing project, Word Combinations in Portuguese Language (COMBINA-PT), developed in order to account for the most significant MW units in Portuguese, extracted from a 50 million word balanced written corpus and imported into a lexical database. Specific issues will be addressed, like the corpus constitution (section 2), the MW unit's extraction tool (section 3), the process of selection of those units (section 4) and further developments (section 5).

2. Constitution of the corpus

The extraction of significant word associations requires a large corpus of real-occurring data. The attainment of frequency data, also allowed through corpus-based studies, is absolutely important to determine (altogether with other criteria of linguistic analysis) if a particular group of words may be

considered a MW unit with a certain stability in the language (which is important when one faces groups that recently started to appear).

The corpus used for MW unit's extraction is a balanced 50,8M word written corpus extracted from the Reference Corpus of Contemporary Portuguese, a monitor corpus of 330 million words, constituted by sampling from several types of written and spoken text and comprising all the national and regional varieties of Portuguese (http://www.clul.ul.pt/english/sectores/projecto_crpc.html). In the near future, we plan to enlarge our results by extracting the MW units of a Portuguese spoken corpus of 1M words, previously compiled at CLUL. However, the data will be processed separately due to the strong discrepancy between the available amount of written and spoken corpus.

Since a particular word may co-occur with different lexical units according to the type of discourse in which they occur, the corpus balance is an important aspect to be considered. In this way, it is essential that the different types of discourse have a balanced dimension in order to properly describe every different patterns of co-occurrence of a lexical unit.

According to these criteria, the corpus has the following constitution:

CORPUS CONSTITUTION			
NEWSPAPERS			**30.000.000**
BOOKS	Fiction	6.237.551	
	Technical	3.827.551	
	Didactic	852.787	**10.818.719**
MAGAZINES AND JOURNALS	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

3. Extraction of MW units

The first step consisted on the extraction, from the corpus, of all the groups of 2, 3, 4 and 5 tokens with a minimum frequency of 3 for groups of 3 to 5 tokens and 10 for 2-token groups. This task was performed using a software developed at CLUL. The groups automatically extracted are statistically analysed using a selected association measure and are afterwards sorted. The tool allows the user to select which measure to apply, and was first run with Mutual Information (MI), that calculates the frequency of each group in the corpus and crosses this frequency with the isolated frequency of each word of the group, also in the corpus (Church & Hanks 1990).

In order to reduce noise and only extract as much relevant MW units as possible, several cut-off options were implemented when running the extraction tool: (i) excluding combinations separated by punctuation; (ii) excluding two-word groups with initial or ending grammatical words using a stop list; (iii) excluding groups under the selected total minimum frequency. The final candidate list obtained comprises 1,751,377 MW units, still a considerable number.

The results of the application of the MW unit's extraction tool are presented in table 2, exemplified by the MW unit *espécies selvagens* 'wild species'.

15 **espécies selvagens** 1 eg(2) og(15) ic(9.845638) fg(15) fe(2066 397) N(50310890)

110751299	No topo da tabela de animais de	espécies selvagens	que morreram
110751306	odo selectivo, acaba por afectar	espécies selvagens.	Por isso, é
110751313	ameaça à conservação de algumas	espécies selvagens,	como o abutr
110751320	qualidade do habitat das nossas	espécies selvagens	seja também u
110751327	carretar "enormes ganhos para as	espécies selvagens".	Dizem que a
110751334	tenas de milhar de exemplares de	espécies selvagens,	tanto cinegé
110751341	vido verdadeiras carnificinas de	espécies selvagens	protegidas em
110751348	presentados actualmente por duas	espécies selvagens	e por raças d
110751355	nternacional sobre o Comércio de	espécies selvagens	(CITES). Uma
110751362	existe uma grande quantidade de	espécies selvagens.	Na calma mad
110751369	passar (muito facilmente!) para	espécies selvagens	semelhantes.
110751376	Os transgenes que passam para as	espécies selvagens	não podem dep
110751383	mbientes, com inúmeros habitats,	espécies selvagens,	recursos nat
110751390	izadas Base de Dados relativas a	espécies selvagens	de Flora e Fa
110751397	um homem suspeito de tráfico de	espécies selvagens.	Para apreend

Table 2. Example of the extraction of the MW unit *espécies selvagens* 'wild species'

In the results presented in table 2, the tool automatically extracts several types of information:

- Distance: 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 (first number after the MW unit in bold);
- Number of elements of the group ("eg")
- Frequency of the group at a specific distance ("og");
- Lexical association measure (Mutual Information) ("ic") ;
- Total frequency of the group in all occurring distances ("fg");
- Frequency of each element of the group ("fe");
- Total number of words in the corpus ("N");
- Concordances lines (in KWIC format) of the MW expression in the corpus, together with reference code.

4. Selection of MW units

In order to enable the representation of MW units and to offer a platform for user-friendly manual validation a lexical database was designed in Access format. The candidate list is loaded into the database together with the associated fields: statistical measure, frequency, distance, number of elements and concordance lines in KWIC format. The manual revision process consists of MW expressions validation as well as concordance lines validation since some contexts are wrongly associated with specific MW expressions. Doubtful cases of significant concordance lines can be solved by viewing a larger concordance context since the database is associated with the corpus through an index file. When concordance lines are eliminated, the total group frequency is automatically recounted in the Frequency field.

When selected as a significant MW unit, the group is attributed a numeric value. The first objective was to establish a correspondence between numeric values and a MW units typology that would be based on cohesion, compositionality (or not), substitutability, etc.. However, first experience of MW units validation proved to be extremely difficult to establish the degree of fixedness of each group and a very time-consuming task to be performed at a first stage of the work. In order to accomplish this task in the established time and with the maximum accuracy, the decision was taken to only select the groups that presented some syntactic and semantic cohesion. This filtration will afterwards allow an easier elaboration of a more precise typology of MW expressions.

According to these criteria, when a MW unit is selected as a valid one it receives a value for the attribute "Type of multiword unit", that covers the following types:

— groups forming a lexical category (e.g., *chapéu de chuva* 'umbrella'; *casa de banho* 'bathroom'; *fim de semana* 'weekend'; *fato de banho* 'swimming suit' – these are examples of expressions that may or may not occur with an hyphen);

— groups forming a phrase, for example a nominal or adjectival phrase (e.g., *senso comum* 'common sense'; *arte contemporânea* 'contemporary art'; *manteiga rançosa* 'rancid butter'; *extremamente importante* 'extremely important');

— groups that comprise a sentence (e.g., *olhar de lado* 'to leer at'; *lançar um ultimato* 'to make an ultimatum'; *pôr a mesa* 'to set the table'; *recuperar de uma lesão* 'to recover from a injury);

— groups that specify named entities, such as institutions, functions, etc. (e.g., *União Europeia* 'European Union', *Presidente da República* 'President of the Republic'; *Dia Mundial da Paz* 'International/World Day of Peace; *Tratado de Amesterdão* 'Amsterdam treaty');

— cases where MW units have more than 5 tokens (maximum of tokens extracted from the corpus) and that will be recovered after the implementation of the lemmatization (e.g.; *não há amor como o primeiro* 'there's no love like the first love'; *pôr/colocar o carro à frente dos bois* 'to put the cart before the oxen');

— cases of MW expressions that require further attention.

An example of a record represented in the database is presented in figure 1.

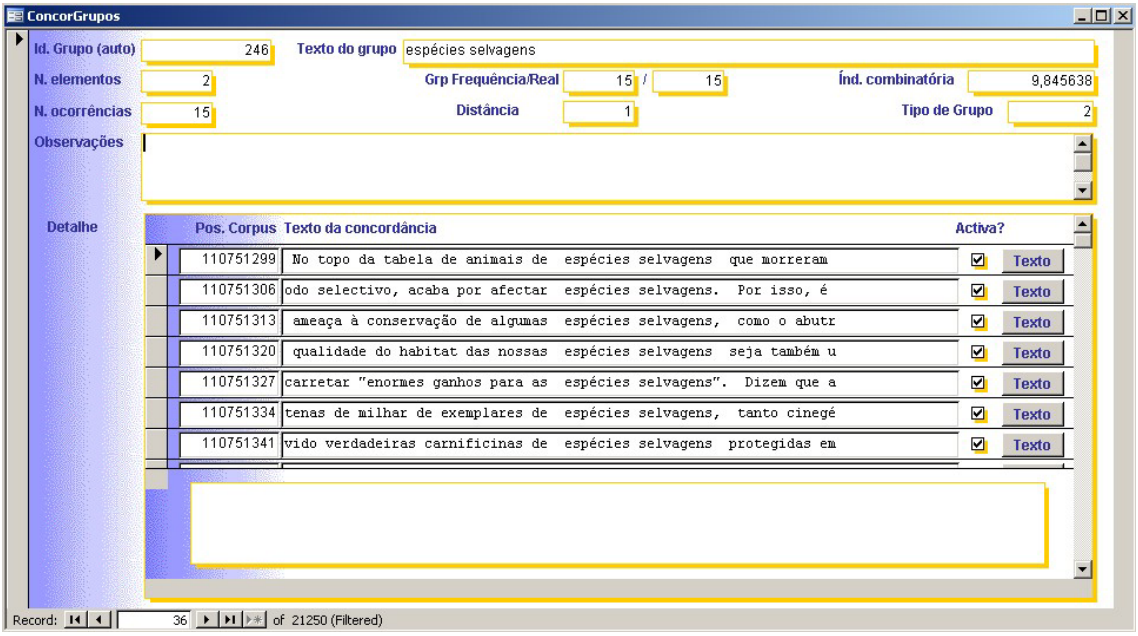

Fig. 1. Record for the collocation *espécies selvagens* 'wild species' in the database

The large candidate list of 1,7 million units made it impossible to assure manual validation of the whole data in the two-year time of the project, making it necessary to hand-check only a subpart of the groups automatically extracted. The selection of this subpart relied on the association measure applied to the set of candidate list, namely Mutual Information (MI), and on the results obtained with the list ordering. Our previous work on corpus extracted MW expressions (Bacelar do Nascimento (2000) and Pereira & Mendes (2002)) using MI had showed the strong tendency of this measure to present the best results with medium values instead of presenting the most significant MW units at the top of the results and new observation of specific lemma confirmed this evaluation of MI measure (similar to the conclusions attained by evaluative studies of several word association measures like Evert & Krenn (2001)). The total candidate list presented MI values between -5 and 33 and data observation showed that the most significant MW units received a MI value between 7 and 11. Having in mind the time-span available for manual validation, we selected a first subpart of the candidate list covering MI values between 8 and 10, in a total of 170,000 units, i.e., 10% of the initial candidate list, which would be hand-checked. From these 170,000 units, we selected about 31,000 as being significant MW units and other 1,637 groups were considered doubtful and will be further evaluated.

A list of all the word forms present in a selected MW unit is automatically created enabling the evaluation of all the groups a word enters in and producing a list of lexical elements associated with all the MW expressions that contain that word and that were considered significant units. Manual validation can also be processed through the list of all inflected forms in the candidate list, since each inflected form is associated with a list of all MW expression it enters in.

Even without following a previously established typology for the selection of MW units, a brief analysis of the already selected units showed that different degrees of cohesion are covered, ranging from:

— frozen groups (such as proverbs or idioms) that have a non-compositional meaning but are fully lexicalized and do not undergo neither morphosyntactic variation nor internal modification (e.g., *um dia é da caça outro do caçador* 'every dog has his day');

— semi-frozen groups that have a non-compositional meaning but are partially lexicalized (e.g., *fazer tábua rasa* 'to wipe off the slate'). These expressions are not subject to syntactic variability (e.g., internal modification **fazer a tábua muito rasa* 'to wipe off the slate a lot' or passivization **a tábua foi feita rasa* 'the slate was wiped off'), but they can undergo inflectional variation (e.g., *fizeram tábua rasa* 'they wiped off the slate');

— semi-frozen groups that can be either compositional or semantically idiosyncratic and that allow for the substitution of one of the collocates by another one associated through a synonymy or hyperonymy/hyponymy relation (*onda/maré/vaga de assaltos* 'wave of robberies; *fogo/lume brando* 'gentle/soft fire'; *países/estados membros* 'member states');

— sets of favoured co-occurring forms, that constitute however syntactic dependencies. These expressions are semantically and syntactically compositional but they are statistically idiosyncratic, which means that they are observed with much higher frequency than any other alternative lexicalization of the same concept, revealing that they may be in their way to a possible fixedness (*prática corrente* 'daily practice'; *possível e imaginário* 'possible and imaginary').

All the word forms that are part of the MW expressions that were manually validated as significant ones are compiled into a list of lemma and the hand-checking process will next proceed by covering all MW expressions of those lemma in order to achieve lexical association coverage for specific lexical elements.

5. Further Developments

The program will, in a latter stage, be run with other lexical association measures like t-test and log-likelihood (Dunning, 1993), using the set of already manually validated MW expressions as an important source of data for results evaluation and association measures comparison.

The syntactic and semantic analysis of the selected list of units will be the basis for proposing a typology of MW expressions that will build on the large set of real-occurring data from the corpus. Besides the issues on fixedness degree and compositional meaning, the study of these MW expressions will allow the identification of associative patterns that characterize a word according to: (i) co-occurrence patterns (systematic co-occurrence with particular lexical items in a contiguous or non-contiguous form); (ii) grammatical patterns (systematic co-occurrence with a certain verb class, with a specific temporal verb forms or with certain syntactic constructions); (iii) paradigmatic patterns (hyperonymy, homonymy, synonymy or antonymy phenomena); (iv) discursive patterns (strong associations in one language register can be a weak association in another register).

In a latter stage, we also intend to run the program on several other resources previously compiled, namely specialized corpora and corpora of Portuguese varieties (Brazilian and African), and evaluate and compare the results.

This will be an important resource for the Portuguese language that will make available important data for the study of MW expressions, from the point of view of lexicography and lexicology, the lexicon-syntax interface and natural language processing.

The Lexical Database of manually revised MW units will be available for online query at the project site (http://www.clul.ul.pt/english/sectores/projecto_combina.html).

6. Acknowledgments

The work described in this presentation has been undertaken under the project Word Combinations in Portuguese Language (COMBINA-PT), sponsored by the Portuguese Ministry of Science (POCTI/LIN/48465/2002) and developed at the Center of Linguistics of the University of Lisbon.

The authors would like to thank the reviewers for all the helpful comments.

7. References

Bacelar do Nascimento, M. F. (2000) "Exemples de combinatoires lexicales établis pour l'écrit et l'oral à Lisbonne", in Bilger, M. (ed.), *Corpus, Méthodologie et Applications Linguistiques*, Paris: H. Champion et Presses Universitaires de Perpignam 2000, pp. 237-261.

Bahns, J. (1993) "Lexical collocations: a contrastive view", *ELT Journal*, 47:1, pp. 56-63.

Braasch, A. & S. Olsen (2000) "Toward a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon", *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1009-1016.

Butler, C. S. (1998) "Collocational Frameworks in Spanish", *International Journal of Corpus Linguistics*, vol. 3(1), pp. 1-32.

Calzolari, N., C. J. Fillmore, R. Grishman, N. Ide, A. Lenci, C. MacLeod & A. Zampolli (2002) "Towards Best Practice for Multiword Expressions in Computational Lexicons", *Proceedings of the Second International Conference on Language Resources and Evaluation*, Las Palmas, Canary Islands; Spain, 29 May – 31 May 2002, pp. 1934-1940.

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

Clear, J. (1993) "From Firth principles: Computational tools for the study of collocation", in Baker, M., G. Francis and E. Tognini-Bonelli (eds.) *Text and technology: In honour of John Sinclair*, Amsterdam, John Benjamins.

Dunning, T. (1993) "Accurate methods for the statistics of surprise and coincidence", *Computational Linguistics*, 19(1), pp. 61-74.

Evert, S. & B. Krenn (2001) "Methods for the Qualitative Evaluation of Lexical Association Measures", *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pp. 188-195.

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

Firth, J. (1957) "A Synopsis of Linguistics Theory, 1930-1955", *Studies in Linguistic Analysis.* Oxford Philological Society; reprinted in Palmer, F. (ed.) (1988) *Selected Papers of J. R. Firth*, Harlow, Longman.

Hausmann, K. W (1979) "Un dictionnaire des collocations est-il possible?", in *Travaux de Linguistique et de Littérature XVII*, 1.

Heid, U. (1998) "Towards a corpus-based dictionary of German noun-verb collocations", *Euralex 98 Proceedings*, Université de Liège.

Kjellmer, G. A. (1994) *Dictionary of English Collocations*, Oxford, Oxford University Press.

Krenn, B. (2000a) "CDB - A Database of Lexical Collocations", *Proceedings of the Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June 2000, vol. II, pp. 1003-1008.

Krenn, B. (2000b) "Collocation Mining: Exploiting Corpora for Collocation Identification and Representation", *Proceedings of KONVENCS 2000*, Ilmenau, Deutschland.

Mackin, R. (1978) "On collocations: Words shall be known by the company they keep", in *Honour of A. S. Hornby*, Oxford, Oxford University Press, pp. 149-165.

Mel'cuk, I. (1984) *Dictionnaire explicatif et combinatoire du français contemporain*, Les Presses de L'Université de Montréal, Montréal, Canada.

Pearce, D. (2002) "A Comparative Evaluation of Collocation Extraction Techniques", *Proceedings of the Third International Conference on Language Resources and Evaluation*, Las Palmas, Spain, pp. 13-18.

Pereira, L. A. S. & A. Mendes (2002) "An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications", in Braasch, A. & C. Povlsen (eds.), *Preceedings of the 10th EURALEX International Congress*, Copenhagen, Denmark, vol. II, pp. 841-849.

Pereira, L. A. S. (1994) *Como se combinam as palavras? Contributo para um Dicionário de Combinatórias do Português*, M.A. Thesis, Faculty of Letters, University of Lisbon, ms.

Sag, I., T. Baldwin, F. Bond, A. Copestake & D. Flickinger (2002), "Multiword Expressions: A Pain in the Neck for NLP", in Gelbukh, A. (ed.) *Proceedings of CICLing-2002*, Mexico City, Mexico.

Sinclair, J. & A. Renouf (1991) "Collocational Frameworks In English", in Aijmer, Karin, and Bengt Altenberg (eds.) *English Corpus Linguistics: Studies in Honour of Jan Svartvik*, Longman, Harlow, pp. 128-143.

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