Corpus-based extraction and identification of Portuguese Multiword Expressions

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

Universidade de Lisboa – Centro de Linguística
{amalia.mendes ; sandra.antunes ; fbacelar.nascimento ; luisa.alice, ptsa}@clul.ul.pt
miguel.casteleiro@zmail.pt

Résumé

Cet article présente la méthodologie suivie et les résultats obtenus dans le cadre d'un projet qui a pour objectif la construction d'une large base de données d'expressions multi-mots de la langue portugaise. Ces expressions multi-mots ont été automatiquement extraites d'un corpus équilibré de 50 millions de mots, interprétées statistiquement à l'aide de mesures d'association lexicales et ont été ensuite manuellement vérifiées. La base de données lexicales recouvre différent types d'expressions multi-mots avec différents degrés de cohésion, qui vont de la quasi totale fixité jusqu'aux groupes de mots qui se réalisent préférentiellement ensemble, comme les collocations. Le large ensemble de données de cette ressource permettra une révision des typologies d'unités multi-mots en portugais et l'évaluation de différentes mesures d'associations lexicales.

Mots-clés : expressions multi-mots, collocations, extraction d'information, base de données lexicales, mesures d'associations lexicales, typologies d'expressions multi-mots.

Abstract

This presentation reports on an on-going project aimed at building a large lexical database of corpus-extracted multiword (MW) expressions for the Portuguese language. MW expressions were automatically extracted from a balanced 50 million word corpus compiled for this project, furthermore these were statistically interpreted using lexical association measures, followed by a manual validation process. The lexical database covers different types of MW expressions, from named entities to lexical associations with different degrees of cohesion, ranging from totally frozen idioms to favoured co-occurring forms, such as collocations. We aim to achieve two main objectives with this resource. Firstly to build on the large set of data of different types of MW expressions, thus revising existing typologies of collocations and integrating them in a larger theory of MW units. Secondly, to use the extensive hand-checked data as training data to evaluate existing statistical lexical association measures.

Keywords: multiword expressions, collocations, information extraction, lexical database, lexical association measures, typology of multiword expressions.

1. Introduction

The advance of computer technologies allowed the development of corpus-based approaches which enable the identification and analysis of complex patterns of word associations (like compound nouns, technical terms, idioms or proper names), 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). These word associations uncovered with corpus analysis are not immediately identified when one only relies on intuition-based studies. And, since new word associations are regularly created, the attainment of frequency data, 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 collocation with a certain stability in the language.

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. As a result, the study of MW expressions became important for several areas, ranging from 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) or computational linguistics, where these expressions are of high importance for many applications in the field of natural language processing, namely information extraction and retrieval, language generation, machine translation, question-answering systems, and word sense disambiguation, since one still finds notorious difficulties (*e.g.*, overgeneration, idiomaticity and parsing problems (Sag *et al.*, 2002)), when the systems have to deal with MW units.

The work reported in this presentation aimed firstly at uncovering significant lexical collocations of European Portuguese, but the scope of the project was enlarged so as to cover not only collocations, but also more fixed expressions, like idioms, and also named entities. Portuguese MW expressions were automatically extracted from a 50 million words balanced written corpus and imported into a lexical database. The presentation will first focus on the corpus constitution and on the specifically designed MW unit's extraction tool (section 2); the lexical database design and implementation and the methodology adopted (section 3), and further possible developments (section 4).

2. Multiword expressions' corpus extraction tool

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).

The corpus balance is an important aspect to be considered, since a particular word may co-occur with different lexical units according to the genre type of the text in which it occurs. 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. We plan to later proceed with this project with the extraction of MW units from a spoken corpus of 1M words, previously compiled at CLUL. We decided not to include spoken register in the corpus for now since the strong discrepancy between the available written and spoken corpus makes it preferable to process the data separately.

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[1]	**277.586**
TOTAL	**50.866.984**

Table 1. Constitution of the corpus

The MW unit's extraction tool automatically extracts from the corpus groups of 2, 3, 4 and 5 tokens and gives the results presented in table 2 and exemplified by the MW unit *noite de consoada* 'Christmas eve'.

10 **noite de consoada** 1 eg(3) og(10) ic(8.588317) fg(10) fe(16971 2290575 52) N(50310890)

209764730	s da SIC -- que o transmitirá na	noite de consoada	-- tomam os se
209764737	Povinho» à droga, passando pela	noite de consoada,	a discoteca e
209764744	ulham presentes numa evocação da	noite de consoada.	À medida que
209764751	e vai continuar a trabalhar pela	noite de consoada	adentro. Texto
209764758	ezes, faltar alguma coisa para a	noite de consoada.	Ainda que o l
209764765	as. Saiu para a rua. Nem parecia	noite de consoada.	Aqui e ali, e
209764772	À memória vêm-lhe imagens de uma	noite de Consoada,	muito tradici
209764779	enor: ao falar, por telefone, na	noite de consoada,	no intervalo
209764786	a vida foi deslizando assim. Na	noite de Consoada,	porém, aconte
209764793	ário O ADEUS ÀS ARMAS Quando, na	noite de consoada,	se iniciou a

Table 2. Example of the MW unit noite de consoada 'Christmas eve'

In the results presented in table 2, several types of information are available for each group:

- 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, in table 2);
- Number of elements of the group ("eg" in table 2);
- Frequency of the group at a specific distance ("og" in table 2);
- Lexical association measure: 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). MI 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 and Hanks, 1989) ("ic" in table 2);
- Total frequency of the group in all occurring distances ("fg" in table 2);
- Frequency of each element of the group ("fe" in table 2);
- Total number of words in the corpus ("N" in table 2);
- Concordances lines (KWIC format) of the MW expression in the corpus, together with an index code pointing to the exact occurring position in the corpus.

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

[1] Parliament sessions are considered written data since the spoken sessions undergo extensive revision when transcribed.

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 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). The final candidate list obtained comprises 1,751,377 MW units, still a considerable number.

3. A database of MW units

A lexical database was designed in Access format so as to enable the representation of MW units and to offer a platform for user-friendly manual validation. The candidate list is loaded into the database together with all the associated fields. Manual validation relies strongly on those fields: statistical measure, frequency, distance, number of elements and concordance lines in KWIC format. The association between the database and the corpus text allows for a larger concordance context to be viewed during the hand-correction process. Concordances wrongly associated with the MW unit can also be eliminated and the total group frequency is automatically recounted in the Frequency field.

The first objective of the project was to establish the set of significant lexical collocations for Portuguese and to assign to each selected expression a numeric value that would correspond to a collocations typology based on cohesion, as well as on compositional meaning or not. However, the exact definition of a collocation and how it differs from other MW expressions is known as a challenging issue, since discrete categorization is difficult to apply to concepts defined in terms of degree of fixedness, compositionality and substitutability. The first experience of manual validation proved to be extremely difficult to establish such categorization of each group and a very time-consuming task to be performed at a first stage of the work, considering the large set of groups to be covered.

This led us to select expressions that presented some syntactic and semantic cohesion, without attempting to follow any prior typology. Different types of MW expressions are thus covered, with different degrees of cohesion, ranging from:

- frozen groups, such as proverbs or idioms, that are fully lexicalized and that do not undergo neither morphosyntactic variation nor internal modification (*e.g.*, *patrão fora, dia santo na loja* 'while the cat is away, the mice will play');
- semi-frozen groups where the meaning of the expression can not be predicted by the meaning of the parts (*e.g.*, *esticar o pernil* 'kick the bucket'), that are not subject to syntactical variability (*e.g.*, internal modification **esticar o grande pernil* 'kick the big bucket' or passivization **o pernil foi esticado* 'the bucket was kicked') but allow inflectional variation (*e.g.*, *esticaram o pernil* 'kicked the bucket');
- semi-frozen groups that can be either compositional or semantically idiosyncratic and that allow for the substitution of one of the collocates by other words associated through a synonym or hyperonymy/hyponym relation (*e.g.*, *onda/maré/vaga de assaltos* 'wave of robberies'; *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 satistically idiosyncratic and they are observed with much higher frequency than any other alternative lexicalization of the same concept, which may reveal that they may be in their way to a possible fixedness (*e.g.*, *instaurar um processo* 'to bring an action'; *erros e imprecisões* 'mistakes and imprecisions').

This first broad selection will afterwards allow an easier elaboration of a more precise typology of MW expressions.

Some data organization was however attempted through the use of numeric values assigned to the MW expressions and that correspond to 4 main purposes:

- to identify MW expressions that may or may not occur with an hyphen (*e.g.*, *casa de banho* 'bathroom'; *fato de banho* 'swimming suit');
- to identify MW expressions that refer to named entities (*e.g.*, *União Europeia* 'European Union', *Presidente da República* 'President of the Republic');
- to identify expressions that constitute verbal phrases (*e.g.*, *respirar fundo* 'to breathe deeply'; *perder os sentidos* 'to loose consciousness') and expressions that constitute other phrases, like nominal phrases (*e.g.*, *ar puro* 'fresh air'; *armas de destruição massiva* 'weapons of mass destruction') or adjectival phrases (*e.g.*, *absolutamente indispensável* 'absolutely indispensable');
- to identify 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.

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

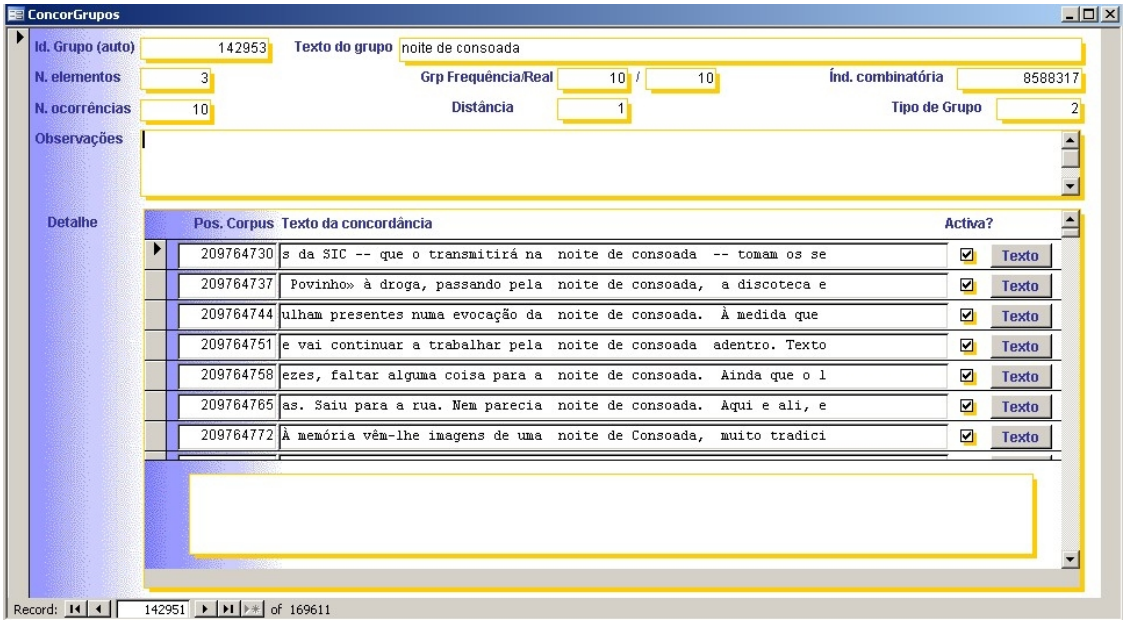

Figure 1. Record for the collocation noite de consoada *'christmas eve' in the database*

A list of all the words presented in the selected MW units is automatically created. Manual validation can also be processed through the list of all word forms in the candidate list, since each inflected form is associated with a list of all MW expression it enters in.

Due to the high number of data to analyse, the manual validation of the list of MW expressions was first applied on a selection of the results, based on the values of the association measure Mutual Information. Previous studies of automatic extraction and evaluation of MW units undertaken by the team (Bacelar do Nascimento, 2000; Pereira and Mendes, 2002) showed that the higher results of the MI measure were not the most significant ones, and that MI values around 7-11 presented a higher number of good MW expressions. A

similar conclusion was reached in a comparative study of lexical association measures (Evert and Krenn, 2001), showing that MI, however giving lower results than other measures on the first sections of the candidate list, present similar results in the remaining sections. A general manual survey of several frequency spans 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 led us to first select MW expressions with MI values between 8 and 10, giving a total of 170,000 candidate units out of the total list of 1,7M.

From these 170,000 hand-checked units, 30,966 were categorically selected as a MW unit while 1,637 belong to the expressions that will later require further attention. From a statistically point of view, in the manual validation task, a numeric value was assigned to 19.1 % of the candidate units, a very low percentage. We plan to compare the MI results with other association measures like t-test and log-likelihood.

A list of all the words that were part of at least one MW unit selected as significant was then elaborated so as to manually evaluate all the groups this word enters in. This will produce a list of lexical elements associated with all the MW expressions that contain the word and that were considered as significant units.

The next step is to organize data in terms of lemmatization. It is a fact that 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 in the corpus and that were statistically interpreted, making it necessary to reorganize our list of possibly significant groups by lemmatizing the set of MW expressions and to associate groups with each lemma. In some cases, only one of the group elements will accept variation, like *não valer a pena* 'not to be worthwhile', where only the verb *valer* can vary (being a 3rd person impersonal construction, only time and number variation is admitted).

But other cases are more complex since all of the group elements can experience word forms variation. For example, in the expression *estar atento a* 'to be attentive to' the verb can vary in person, number and time, the adjective can vary in gender and number and the prepositional element can be contracted with different articles and pronouns, giving a large set of possibilities (*e.g.*, *estou atento à* 'I'm attentive to_the[fem, sg]', *estamos atentos ao* 'we are attentive to_the[masc, sg]', *estivemos atentos àquela* 'we were attentive to_that_one[fem]' – contracted elements are connected in the English translation). To cover all possible realizations of the MW expression lemma *estar atento a* implies recovering and organizing all different word forms occurrences of this group elements, a large task.

The high number of inflected variants of a MW expression in Portuguese is a factor that makes automatic evaluation of significant units even more difficult: the frequency of most MW units is scattered through several variants that are each evaluated separately with statistical measures that do not cover the group lemma. But processing the lexical measures over the lemmatized MW expressions would also be misleading since it would obliterate the fact that some inflected variants are, in a high number of cases, clearly preferable or even the only possible realizations.

In some cases, none of the word forms realizations of one of the group elements is frequent enough to make the group automatically recognized as a possible significant expression. This is true for many expressions with a verbal element, like *esfregar as mãos de contentamento* 'to rub ones hands with satisfaction', where the different word forms of the verb *esfregar* 'rub' have very low frequency (frequency 1, 2 or 3 maximum). Since a minimum frequency

was established during the tool running process, none of the group realization is recovered (*esfregou_3rdp_sg as mãos de contentamento*, *esfreguei_1stp_sg as mãos de contentamento*, *esfregavam_3rdp_pl as mãos de contentamento*,...). The MW expression is identified from a smaller group *as mãos de contentamento*, and it is the visualization of the concordance lines that points to the existence of a larger group, as can be seen in figure 2, below. The lemmatization of the selected MW expressions is under development.

The syntactic and semantic analysis of the selected list of units will be later the basis for proposing a typology of MW expressions that will build on the large set of real-occurring data from the corpus.

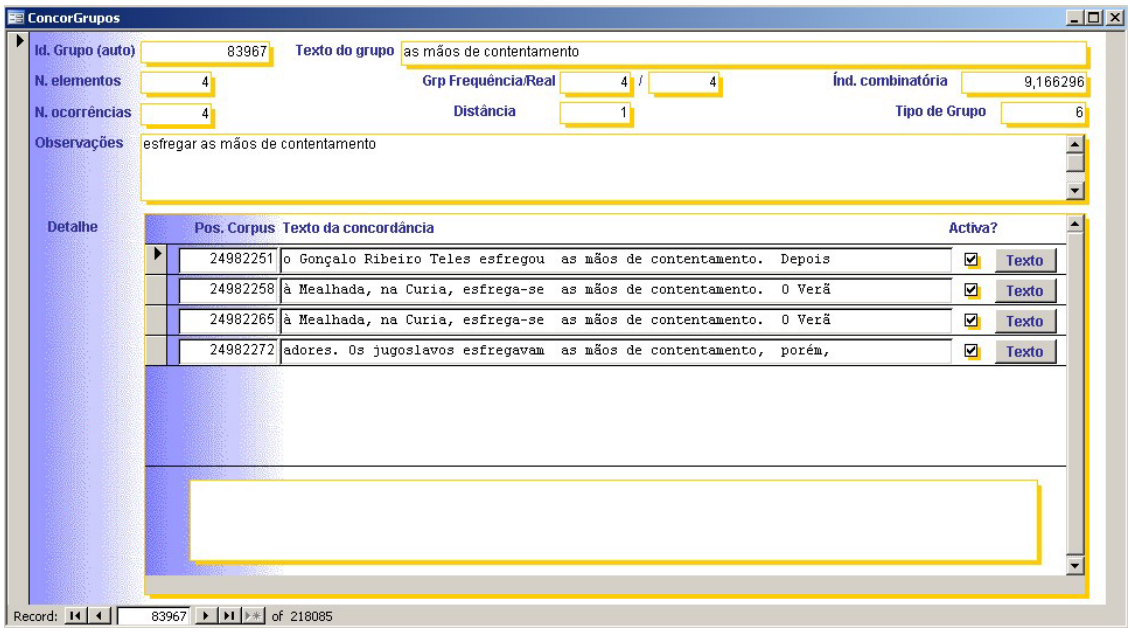

Figure 2. Record for the collocation as mãos de contentamento *'hands with satisfaction' in the database*

4. Further Developments

Besides the issues on fixedness degree and compositional meaning, the study of these MW expressions allows to identify associative patterns that characterizes 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 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).

This resource will be of extreme importance for several areas, like psycholinguistics, lexicography and computational linguistics, helping to develop and evaluate language processing tools able of dealing with MW expressions specific issues, like automatic unit recognition, lexical association measures for validation of significant MW units, as well as tagging and parsing problems.

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

5. Acknowledgments

The work described in this presentation results from the project Word Combinations in Portuguese Language, undertaken at the Centre of Linguistics of the University of Lisbon (http://www.clul.ul.pt), under a research grant of the Portuguese Ministry of Science (POCTI/LIN/48465/2002).

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

References

BACELAR DO NASCIMENTO M.F. (2000). "Exemples de combinatoires lexicales établis pour l'écrit et l'oral à Lisbonne". In M. Bilger (ed.). *Corpus, Méthodologie et Applications Linguistiques*. Champion et Presses Universitaires de Perpignan, Paris-Perpignan: 237-261.

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

BRAASCH A., OLSEN S. (2000). "Toward a Strategy for a Representation of Collocations – Extending the Danish PAROLE-lexicon". In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens: 1009-1016.

BUTLER C.S. (1998). "Collocational Frameworks in Spanish". In *International Journal of Corpus Linguistics* 3(1): 1-32.

CALZOLARI N., FILLMORE C.J., GRISHMAN R., IDE N., LENCI A., MACLEOD C., ZAMPOLLI A. (2002). "Towards Best Practice for Multiword Expressions in Computational Lexicons". In *Proceedings of the Third International Conference on Language Resources and Evaluation*. Las Palmas: 1934-1940.

CHURCH K.W., HANKS P. (1990). "Word association norms, mutual information, and lexicography". In *Computational Linguistics* 16 (1): 22-29.

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

DEANE P. (2005). "A Nonparametric Method for Extraction of Candidate Phrasal Terms". In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL'05). University of Michigan.

EVERT S., KRENN B. (2001). "Methods for the Qualitative Evaluation of Lexical Association Measures". In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*: 188-195.

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

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

HAUSMANN K.W (1979). "Un dictionnaire des collocations est-il possible ?" In *Travaux de Linguistique et de Littérature* XVII (1): 187-195.

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

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

KRENN B. (2000a). "CDB – A Database of Lexical Collocations". In *Proceedings of the Second International Conference on Language Resources and Evaluation*. Athens: 1003-1008.

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

LIN D. (1998). "Extracting Collocations from Text Corpora". In *First Workshop on Computational Terminology*. Montréal.

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

MEL'CUK I. (1984). *Dictionnaire explicatif et combinatoire du français contemporain*. Les Presses de l'Université de Montréal, Montréal.

PANTEL P., LIN D. (2001). "A Statistical Corpus-Based Term Extracted". In *Proceedings of the Biennial Conference of the Canadian Society on Computational Studies of Intelligence: Advances in Artificial Intelligence*: 36-46.

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

PECINA P. (2005). "An Extensive Empirical Study of Collocation Extraction Methods". In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics* (ACL'05). University of Michigan.

PEREIRA L.A.S., MENDES A. (2002). "An Electronic Dictionary of Collocations for European Portuguese: Methodology, Results and Applications". In A. Braasch and C. Povlsen (eds), *Preceedings of the 10th EURALEX International Congress*. Copenhagen: 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.

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

SINCLAIR J., RENOUF A. (1991). "Collocational Frameworks In English". In K. Aijmer and B. Altenberg (eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. Longman, Harlow: 128-143.

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