Multiword Identification

Ricardo Portela[1,2], Nuno Mamede[1,2], and Jorge Baptista[2,3]

[1]Instituto Superior Técnico, Portugal
[2]Spoken Language Laboratory, INESC-ID Lisboa, Portugal
[3]Universidade do Algarve, Portugal

Instituto Superior Técnico, Portugal
rjrp@ist.utl.pt, nuno.mamede@inesc-id.pt, jbaptis@ualg.pt

Abstract. This paper deals with the identification of multiwords in Portuguese texts. The validation is based in a comparison between statistical measures and the use of algorithms that do not need the use of thresholds. Syntactic criteria used to identify multiwords are also explored. Results are based on data taken from a large-sized *corpus*. The statistical comparison is done with the aid of GRID computing, as well as scheduling and parallel programming software for information retrieval.

1 Introduction

In this paper a *word* is a lexical unit consisting of a single token (or graphic unit). A *multiword* is a lexical unit formed by two or more words to yield a new concept, different from the composition of the meaning of its elements. For example, *chapéu de chuva* (umbrella) is a multiword because its meaning (denoting a particular object) is different from the composition of the meanings of *chapéu* (hat) and *chuva* (rain) when used separately.

Multiwords may present different morpho-syntactic structures; such as , for example:

- Noun + *de* (of) + Noun: *lua-de-mel* 'moon-of-honey' (honeymoon)
- Noun + Adjective: *mesa redonda* 'table round' (round table)
- Adjective + Noun: *bela-sombra* 'beautiful shadow' (tree, Amarilidaceae)
- Verb + Noun: *mata-moscas* 'kill-flies' (fly-trap)
- Noun + Noun: *ano-luz* 'year-light' (light year)

Multiwords can also be formed by other morpho-syntactic structures (for a comprehensive typology, see [1]) , but in this paper, only Noun + Adjective situations - which is the most common type of compounding - will be considered.

Since multiwords are formed exactly like ordinary noun phrases, it is necessary to spot them in texts, for they constitute meaning units by themselves, essential to any message understanding tasks. This process is even more complex because in some cases, depending on context, the same word sequence can function both as a phrase, with compositional meaning, and as a multiword,

hence, semantically opaque. This is the case, for example, of *braço direito* 'right arm', which could denote the limb of a person or it could also have a global, non-compositional meaning (wing man).

The identification process was divided in two stages. In the first stage, morphosyntactic patterns are identified in the corpus and their ocurrences are counted, as well as the occurences of their individual elements; in the second stage, these counts are processed with statistical algorithms.

2 Corpus Processing

For the identification of multiwords the following tools were used: the $STRING$ processing chain ; the GRID; the Condor[2] and the Hadoop[3] systems. The $STRING$ (Statistical and Rule-Based Natural Language Processing chain) [4], developed by the Spoken Language Laboratory, L^2F at INESC-ID Lisboa [1], transforms the corpus in to XML files, showing all the lexical and syntactical structures that are needed for extracting multiword patterns. This Natural Language Processing (NLP) chain consists of a series of modules that tokenize, POS-tag (*i.e.* mark words with their appropriate part-of-speech tags: noun, verb adject, etc.) and POS-disambiguate the raw text; the XIP parser [5] is used to chunk the text and extract the syntactic dependencies between the constituents of sentences. The L^2F GRID is a network of computers that is used to accelerate the processing of the corpus through the $STRING$. Condor manages the scheduling of the GRID use during the corpus processing, by queuing and prioritizing processes. Hadoop is used to perform the statistical algorithms for the identification of multiwords on large data, while maintaining a high-throughput data access. The *corpus* here used is the CETEMPúblico[6], a 190 million words, journalistic text drawn from the daily newspaper *Público*[2] and provided by *Linguateca*[3].

Firstly, the corpus was processed using the $STRING$ chain in the GRID using the Condor system in order to obtain the syntactical information needed to identify multiword patterns. To do this, this large-sized corpus should be processed in a manageable time (under 24 hours). Besides, as the NLP chain is in constant upgrade, reprocessing the corpus would have to be done several times, and therefore it should not be an excessively time-consuming process. To solve this problem, the corpus was divided into successively smaller files until the largest file size, that caused no memory problems, was found, with approximately with 280kB. In this way, the GRID is able to process of 60MB of information under 21 minutes and the entire corpus in about 3 hours, well under the desired 24 hour delay.

In order to automatize the processing of the corpus, a series of scripts were made to run all the necessary commands:

[1] http://www.l2f.inesc-id.pt/
[2] http://www.publico.pt/
[3] http://www.linguateca.pt/

- condor-submit.sh: Submits the results to the Hadoop file system;
- run-xip.sh: Runs the xip-runner jar in an input file;
- run-xip.condor: Sets the parameters for running condor (such as the machines to be used);
- xip-runner.jar: A jar application that runs the STRING chain.

The results end up in a specific folder in the Hadoop file system. The xml results can then be easily ran and accessed through the Hadoop structure, which allows the use of programs in the jar format.

3 Statistical Measures

This section describes the statistical measures used by the algorithms to identify multiwords. These algorithms will be described in section 4. The measures here present result from an review of the literature on multiword identification [7].

3.1 Dice coefficient

The Dice coefficient [8], [9] consists in measuring the cohesiveness between two words of a bi-gram $[w_1 p_{12} w_2]$, as defined in equation (1).

$$Dice([w_1 p_{12} w_2]) = \frac{2 \times f([w_1 p_{12} w_2])}{f([w_1]) + f([w_2])} \quad (1)$$

Where $f([w_1 p_{12} w_2])$, $f([w_1])$ e $f([w_2])$ are the respective frequencies of the bi-gram $[w_1 p_{12} w_2]$ and the uni-grams $[w_1]$ and $[w_2]$; p_{12} represents the distance between the words w_1 and w_2.

3.2 Specific Mutual Information

The method *Specific Mutual Information* [10] is used to measure the overlap between two ocurrences, contributing to the measure of cohesiveness between the two words of a bigram, as defined in equation (2).

$$MI([w_1 p_{12} w_2]) = \log_2 \frac{N \times f([w_1 p_{12} w_2])}{f([w_1]) \times f([w_2])} \quad (2)$$

The terms in this equation are the same as in the previous one, while N represents the total number of words in the *corpus*. This method is particularly prone to overestimate data with low frequencies.

3.3 ϕ^2

The ϕ^2 [11] is based in the Pearson's χ^2 test for 2 x 2 contingency tables (see also section 3.6, below), and it is concerned with testing the null hypothesis that two random variables are independent. The null hypotheses of statistical

independence is represented by: $H_0 : p(w_i p_{ij} w_j) = p(w_i) \times p(w_j)$. If ϕ^2 is minimal, the null hypothesis H_0 is true and the words that are being analyzed are independent. Otherwise, one can say that the words were highly related with a certain degree of freedom. This method is defined in equation (3).

$$\phi^2([w_1 p_{12} w_2]) = \frac{(N \times f([w_1 p_{12} w_2]) - f([w_1]) \times f([w_2]))^2}{f([w_1]) \times (N - f([w_1])) \times f([w_2]) \times (N - f([w_2]))} \tag{3}$$

The terms in this equation are the same as in the previous one.

3.4 Simpson Similarity coefficient

The *Simpson Similarity coefficient* [12] evaluates the association between two words by calculating the coefficient of the intersection of two words and the smallest of the two, so as not to underestimate sets in which one of the words presents a much higher frequency than the other, thus yielding very low value for this set. This method is defined in equation (4).

$$SIMPSON([w_1 p_{12} w_2]) = \frac{2 \times f([w_1 p_{12} w_2])}{\min(f([w_1]), f([w_2]))} \tag{4}$$

The terms in this equation are the same as in the first one.

3.5 Symmetrical Conditional Probability

The *Symmetrical Conditional Probability* [13] measures the cohesiveness between two words in a bi-gram, as is defined by equation (5).

$$SCP([x, y]) = p(x|y) \cdot p(y|x) = \frac{p([x, y])^2}{p([x]) \cdot p([y])} \tag{5}$$

p(x,y), p(x) e p(y) are the respective probabilities of the ocurrence of the bigram [x,y] and of the unigrams [x] and [y] in the *corpus*; $p(x|y)$ is the conditional probability of x appearing in the first position of the bigram and the y in the second position of the bigram; in a similar way, $p(y|x)$ is the conditional probability of y appearing in the first position of the bi-gram and the x in the second position of the bigram.

3.6 Pearson's χ^2

The *Pearson's* χ^2 [14], it is based in the comparison of the observed frequencies and the expected frequencies against the null hypothesis. For example, from a sample of 100 balls in which there is an equal number of red and black balls, the expected frequency is 50% for the black balls and 50% for the red balls. This method is defined by equation (6).

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i} \tag{6}$$

The O_i is the observed frequency, E_i the expected frequency and n the number of possible outputs for every event.

3.7 Log-likelihood Ratio

The *Log-likelihood Ratio* [15], just like the ϕ^2, tests the null hypothesis that two random variables are independent. The null hypothesis of statistical independence is stated by $H_0 : p(w_i p_{ij}|w_j) = p(w_i p_{ij}|\overline{w_j})$ thus setting the independence paradigm between two rows of a contingency table. This method is defined by equation (7)

$$\begin{aligned} Loglike([w_1 p_{12} w_2]) = -2\log\lambda = \\ 2\text{x}(\log\theta_1^{s_1}(1-\theta_1)^{n_1-s_1} + \log\theta_2^{n_2-s_2} \\ -\log\theta^{s_1}(1-\theta)^{n_1-s_1} - \log\theta^{s_2}(1-\theta)^{n_2-s_2} \end{aligned} \tag{7}$$

Where

- $s_1 = f([w_1 p_{12} w_2])$
- $s_2 = f([w_2]) - f([w_1 p_{12} w_2])$
- $n_1 = f([w_1])$
- $n_2 = N - f([w_2])$
- $\theta_1 = \frac{s_1}{n_1}$
- $\theta_2 = \frac{s_2}{n_2}$
- $\theta = \frac{f([w_2])}{N}$
- $f([w_1 p_{12} w_2])$, $f([w_1])$ e $f([w_2])$ are the respective frequencies of the bigram $[w_1 p_{12} w_2]$ and the unigrams $[w_1]$ e $[w_2]$
- N is the total number of words in the *corpus*
- p_{12} is the distance between the words w_1 and w_2

3.8 Mutual Expectation

The *Mutual Expectation* [16,17] is based on the concept of the *Normalized Expectation (NE)*, which consists in measuring the cost, in terms of cohesion, of the loss of a word in a N-gram, this is, the probability of a word occuring in a certain position, knowing the ocurrence of the other words and their positions. One of the most efficient criteria for multiword identification is frequency, knowing this one can deduce that between two N-grams with the same *NE*, the most frequent N-gram is more likely to be a multiword. This method can be described by equation (8).

$$\begin{aligned} ME([w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) = p([w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ \text{x} NE([w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \end{aligned} \tag{8}$$

A N-gram is defined algebraically by the vector of words $[w_i \cdot p_{1i} w_i \cdot p_{1n} w_n]$, w_i is a word in the N-gram, p_{1i} is the distance in which separates the word w_1 and the word w_i, $p()$ is the frequency and $NE()$ the *Normalized Expectation*.

4 Algorithms

This section describes the algorithms studied and used for the identification of multiwords. These algorithms where chosen because all of them have a similar point, they all identify if the candidate is a local maxima, this concept will be describe later in the algorithms.

4.1 *LocalMaxs* Algorithm

The *LocalMaxs* algorithm [18] identifies multiwords from a list of N-gram based on two assumptions. First, association measures show us that the more cohesive a group of words is, the higher are the values of the association measures, thus allowing for the multiword identification, and second, multiwords are group of words that are highly associated to each other. From this assumptions, an N-gram W, is a multiword if its association measure, $g(W)$, is the local maxima. The algorithm can be defined by equation (9).

$$\begin{gathered} \forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1} W \text{ is a multiword if} \\ (length(W) = 2 \wedge g(W) > g(y)) \\ \vee \\ (length(W) > 2 \wedge g(x) \leq g(W) \wedge g(W) > g(y)) \end{gathered} \tag{9}$$

Where Ω_{n-1} is the set of association measures of all the (N-1)-grams contained in the N-gram W, and Ω_{n+1} the set of association measures of all the (N+1)-grams containing the N-gram W.

This algorithm does not depend on threshold values, for it focuses only in the identification of local variations of associations measures.

For this work, the following association measures where used: the Dice coefficient, Specific Mutual Information (MI), ϕ^2, Symmetric Conditional Probability(SCP), Log-likelihood Ratio, and Mutual Expectation.

4.2 HELAS

The Helas [19] is a hybrid system, which extracts multiwords using information about the cohesiveness of the group of words and the cohesiveness of the part-of-speech tags. The idea of this system is evaluate the cohesion of the word-tag association, which means that the more a group of words is cohesive and the more the part-of-speech tags are cohesive, the more likely it is that the group of words is a multiword. For this purpose, the global cohesion is evaluated by the combination of the mutual expectation of the words and the mutual expectation of the part-of-speech. This is measured by the *Combined Association Measure* (CAM) which is defined in equation (10).

$$\begin{gathered} CAM([p_{11}u_1t_1 \ldots p_{1i}u_it_i \ldots p_{1n}u_nt_n]) = \\ ME([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n])^{\alpha} \\ \times \\ ME([p_{11}t_1 \ldots p_{1i}t_i \ldots p_{1n}t_n])^{1-\alpha} \end{gathered} \tag{10}$$

Where α is a parameter that defines the focus of more relevancy.

The process of selection is made by the GenLocalMaxs algorithm, which is similar to the algorithm describe before but it uses the results of the CAM's. This algorithm is described in equation (11).

$$\begin{gathered} \forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1}, W \text{ is a multiword if} \\ (sizeof(W) = 2 \wedge CAM(W) > CAM(y)) \\ \vee \\ (sizeof(W) \neq 2 \wedge CAM(W) \geq CAM(x) \wedge CAM(W) > CAM(y)) \end{gathered} \tag{11}$$

Where W is a positional N-gram word-tag, Ω_{n-1} is the set of all the (N-1)-grams contained in the N-gram W, and Ω_{n+1} is the set of all the (N+1)-grams containing the N-gram W and *sizeof()* is a function that returns the number of words of a positional N-gram word-tag.

5 Syntactic criteria for multiword identification

Since noun + adjective multiwords are structurally identical to ordinary noun phrases, their identification is based in the different syntactical properties, namely the restrictions or constraints on the properties that formally identical, ordinary noun phrase would show. The following criteria, based on G.Gross (1988)[20], and adapted to Portuguese by Baptista (1994) [1], were used to manually identify noun + adjective multiwords.

- The adjective stops accepting a post-copula (predicative) context;
- The adjective stops allowing degree variation or another quantification mechanism;
- The adjective can not be coordinated with another adjective when it is part of the multiword;
- The adjective can not be zeroed;
- There is a distributional rupture in the paradigmatic variation of the noun and of the adjective in the multiword;
- The multiword can not vary in number (and/or gender) even if the individual component words allow it;

In the application of these criteria, a candidate sequence is considered a multiword if at least one of the constraints is observed. The more constraints a sequence presents, the higher the degree of cohesion of the multiword.

6 Noun + Adjective Structure Identification

This process is divided in two steps. First a filter is applied to the xml trees resulting from running the CETEMPúblico corpus through the STRING processing chain, thus identifying the xml tags that represent a noun token followed immediately by an adjective token. This results in a file with all the candidates found, followed by the number of occurrences. The processing chain automatically identifies approximately 22,000 different multiword patterns. Other information is also retrieved, namely the (N+1)-grams that contains the identified candidate, bigrams and unigrams of words and their respective part-of-speech tags. Secondly, the data retrieved in the first step is processed to obtain the association measures and to apply the algorithms described earlier in this paper. The measures presented in this paper are used for the calculation of bigrams, with the exception of the Mutual Expectation. In order to process the data, some of measures had to be normalized to calculate trigrams. The normalized measures were the Dice coefficient, the Specific Mutual Information, the ϕ^2, the Symmetric Conditional Probability (SCP), and the Log-likelihood Ratio.

Only the candidates that occur in the data more than five times were processed, because the methods used in this work become unreliable when dealing with rare events [7].

7 Evaluation

The process of evaluation applied both program verification and result validation, to discover program errors in the pattern retrieval and to verify the correct application of the statistical methods and algorithms.

For the first problem, a list of 100 sentences where randomly chosen and processed by the NLP Chain. The resulting output was manually scanned for the desired patterns and then it was automatically scanned by the filter. The results where compared and they matched. A total of 110 noun + adjective (109 different) structures where found.

To validate the results, a list of unigrams, bigrams, multiword candidates, (N+1)-grams and their respective part-of-speech tags where manually produced and the statistical measures and algorithms where manually calculated. The test only consisted in 11 patterns, with a total of 6 different patterns because of the difficulty of manually calculating a large quantity of data. The automatically obtained results matched those that were manually calculated.

8 Results

This section presents the most relevant results from the methods here used. Firstly, the results from the *LocalMaxs* algorithm and the *Helas* system are presented and then some preliminary results from the application of syntactic criteria to multiword identification are provided.

8.1 *LocalMaxs* algorithm and *Helas* system

For the *LocalMaxs* algorithm, 6 different association measures where used. Table 1 shows the number of different multiwords found for the structure Noun + Adjective for each different measure and the total number of instances.

Table 1. Results for *LocalMaxs*

Association measure	Nr. of multiwords	Nr. of instances
Dice coefficient	116,565	2,981,983
Specific Mutual Information(MI)	12,917	630,767
ϕ^2	21,319	12,51948
Symmetric Conditional Probability(SCP)	22,967	1,527,815
Log-likelihood Ratio	116,036	182,9301
Mutual Expectation	139,701	3,273,087

The results where crossed to determine the number of multiwords that are identified by all association measures. The results were also crossed but successively removing an association measure at a time, in order to determine which measure could be less significant for the identification. The results are presented in table 2.

Table 2. Number of different multiwords and number of instances in the crossed results

	Nr. of multiwords	Nr. of instances
All measures crossed	4,368	91,788
crossing without Dice coefficient	4,374	91,840
crossing without MI	14,577	498,115
crossing without ϕ^2	4,516	106,290
crossing without SCP	4,439	94,498
crossing without Log-likelihood Ratio	6,031	345,611
crossing without Mutual Expectation	4,368	91,788

The discrepancy of the results is significant when the Specific Mutual Information association measure is not crossed and a small difference is noted when the Log-likelihood Ratio is not crossed.

In the Helas system 11 different values for α were used: $\alpha \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$, where the total focus in the words is $\alpha = 1.0$ and the total focus of the part-of-speech tags is $\alpha = 0.0$. The results where equal for all values of α, identifying a total of 139,701 different multiwords and 3,273,087 instances of the multiwords.

The same result should be expected for $\alpha = 1.0$ and the result of the *LocalMaxs* using the Mutual Expectation association measure because without using information of part-of-speech tags the two algorithms are exactly the same. As

for the other results, this system does not have information about the 22,000 most common multiwords, this problem induces the loss of significant information for multiword identification.

Precision was measured from manually checking a randomly chosen sample of 908 candidates, retrieved from the crossing of all measures: 29,5% of the candidates were marked as multiwords, and one of them was marked as ambiguous. It should be noticed, however, that the list of multiword candidates do not include the compound words already identified by the STRING lexicons, which consist of more than 35,000 entries, the large majority on this particular noun+adjective type.

8.2 Syntactic criteria for multiword identification

A preliminary attempt was also made to apply the syntactic criteria for multiword identification using the data resulting from the STRING chain processing of the corpus, particularly by exploring the set of dependencies produced by the XIP parser. Two criteria were tested: the adjective loss of predicativity and restrictions on number variation of the multiword. Since these criteria were devised for human/manual identification of multiwords, their application to the data can only be interpreted as an approximation. For the first criterion, the data was searched for patterns where the adjective modifies the noun of the multiword candidate by way of a copula verb, *e.g. mesa redonda*: *A mesa é redonda* (round table: The table is round). The search pattern was also extended to included relative clauses having the noun as the antecedent of the relative pronoun, *e.g. A mesa que é redonda* (The table that is round). If the pattern is found the candidate is excluded. Secondly, number variation was tested by comparing the singular/plural ratio of the candidate multiword against the singular/plural ratio of the noun without any modifying adjectives. The ratio is calculated by dividing the number of occurrences in singular or plural by the total number of occurences (singular + plural), if the ratio of the candidate multiword is below 0.9 this candidate is discarded. If the difference between the ratio of the multiword and the ration of the noun alone is below a given threshold the multiword candidate is discarded. Different thresholds were tested for the sample, until the best performing value of 0.2 was found.

Table 3 shows the results from the application of this syntactic criteria to data from the above mentioned sample of 908 multiword candidates.

Table 3. Syntactic criteria

Words classified with	Percentage of classified words
Predicative context	11.67%
Variation in Number	29.52%

The results from the predicative context are very low compared with the results of the crossing of all measures described previously, although the results

from the variation in number are very similar to them, not being remarkably better.

9 Conclusions and Future Work

The values of precision and the number of candidates retrieved by the systems here used may have resulted from not counting the multiwords previously identified by the STRING processing chain. A comparison with the data obtained without using the large-sized lexicons available should provide a better estimate of the systems' performance.

Because of the large data retrieved and the number of extracted multiword candidates, it is not feasible to manually calculate the precision over the entire data set. Hence, only a small random sample was used in this paper, to gauge precision. To solve this problem, a larger, statistically acceptable sample will be retrieved and checked manually.

Results from using syntactic criteria, while still preliminary, are promising. In future work, the remaining syntactic criteria will be assessed in the identification of multiword candidates. This information will also be used with the genetic algorithm GALEMU [21], which is suitable for this type of data.

10 Supports

This work was supported by project CMU-PT/HuMach/0053/2008, sponsored by Fundação para a Ciência e a Tecnologia (FCT).

References

1. J. Baptista. Estabelecimento e Formalização de Classes de Nomes Compostos. Master's thesis, Faculdade de Letras da universidade de Lisboa, Lisboa, 1994.
2. Todd Tannenbaum, Derek Wright, Karen Miller, and Miron Livny. Condor – A Distributed Job Scheduler. In Thomas Sterling, editor, *Beowulf Cluster Computing with Linux*, chapter 15. MIT Press, Outubro 2001.
3. T. Luís. Parallelization of Natural Language Processing Algorithms on Distributed Systems. pages 9–11, Universidade Técnica de Lisboa, Portugal, 2008.
4. N. Mamede. String - A Cadeia de Processamento de Língua Natural do L^2F em Fevereiro de 2011 (Technical Report). INESC-ID Lisboa, Lisboa, 2011.
5. Aït-Mokhtar, Salah; Jean-Pierre Chanod, and Claude Roux. Robustness beyond shal- lowness: Incremental deep parsing. Natural Language Engineering, 8. pages 121–144, Cambridge University Press, New York, 2002.
6. D. Santos and P. Rocha. Evaluating CETEMPúblico, a free resource for Portuguese. In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 442–449, Toulouse, 2001.
7. P. Pecina and P. Schlesinger. Combining Association Measures for Collocation Extraction. In *ACL'06*, page 652, 2006.

8. Frank Smadja, Kathleen R. McKeown, and Vasileios Hatzivassiloglou. Translating collocations for bilingual lexicons: a statistical approach. *Comput. Linguist.*, 22(1):1–38, 1996.
9. L. Dice. Measures of the Amount of Ecologic Association Between Species. *Journal of Ecology*, 1945.
10. Kenneth Ward Church and Patrick Hanks. Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1):22–29, 1990.
11. W. Gale and K. Church. Concordances for Parallel Texts. *Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*, 1991.
12. F. Martínez-Santiago, M.C. Díaz-Galiano, M.T. Martín-Valdivia, V.M. Rivas-Santos, and L.A. Ure na Lopez. Using Neural Networks for Multiword Recognition in IR. In *Proceedings of Conference of International Society of Knowledge Organization (ISKO-02)*, pages 559–564, Granada, Espanha, 2002.
13. G. Lopes and J. Silva. A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In *Proceedings of the* 6^{th} *Meeting on the Mathematics of Language*, pages 369–381, 1999.
14. D. Hull and G. Grefenstette. Experiments in Multilingual Information Retrieval. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 9–6, 1996.
15. T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19:61–74, 1993.
16. G. Dias, S. Guilloré and J. Lopes. Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora. In *Proceedings of 6éme Conferérence Annuelle sur le Traitement Automatique des Langues Naturelles*, Cargése, 1999.
17. B. Daille. Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. In Judith Klavans and Philip Resnik, editors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*, pages 49–66. The MIT Press, Cambridge, Massachusetts, 1996.
18. S. Guilloré J. Silva, G. Dias and J. Lopes. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In *EPIA '99: Proceedings of the 9th Portuguese Conference on Artificial Intelligence*, pages 113–132, London, UK, 1999. Springer-Verlag.
19. G. Dias. Multiword unit hybrid extraction. In *Proceedings of the ACL 2003 workshop on Multiword expressions*, pages 41–48, Morristown, NJ, USA, 2003. Association for Computational Linguistics.
20. G. Gross. Degré de figement des noms composés. In *Languages 90*, pages 57–72, Paris: Larousse, 1988.
21. G. Dias and S. Nunes. Evaluation of Different Similarity Measures for the Extraction of Multiword Units in a Reinforcement Learning Environment. In *Proceedings of the 4th International Conference on Languages Resources and Evaluation*, pages 1717–1721, 2004.