Using *LocalMaxs* Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units

Joaquim Ferreira da Silva[1], Gaël Dias[1], Sylvie Guilloré[2] & José Gabriel Pereira Lopes[1]

[1] Universidade Nova de Lisboa
Faculdade de Ciencias e Tecnologia, Departamento de Informática
Quinta da Torre, 2725, Monte da Caparica, Portugal
{jfs,ddg,gpl}@di.fct.unl.pt
[2] Université d'Orléans
Laboratoire d'Informatique Fondamentale d'Orléans
BP 6102 - 45061, Orléans Cédex 2, France
sylvie.guillore@lifo.univ-orleans.fr

Abstract. The availability of contiguous and non-contiguous multiword lexical units (MWUs) in Natural Language Processing (NLP) lexica enhances parsing precision, helps attachment decisions, improves indexing in information retrieval (IR) systems, reinforces information extraction (IE) and text mining, among other applications. Unfortunately, their acquisition has long been a significant problem in NLP, IR and IE. In this paper we propose two new association measures, the Symmetric Conditional Probability (SCP) and the Mutual Expectation (ME) for the extraction of contiguous and non-contiguous MWUs. Both measures are used by a new algorithm, the *LocalMaxs*, that requires neither empirically obtained thresholds nor complex linguistic filters. We assess the results obtained by both measures by comparing them with reference association measures (Specific Mutual Information, ϕ^2, Dice and Log-Likelihood coefficients) over a multilingual parallel *corpus*. An additional experiment has been carried out over a part-of-speech tagged Portuguese *corpus* for extracting contiguous compound verbs.

1 Introduction

The acquisition of MWUs has long been a significant problem in NLP, being relegated to the borders of lexicographic treatment. The access to large-scale text *corpora* in machine-readable formats has recently originated a new interest in phraseology. The evolution from rule based formalisms towards lexicalization, that is the evolution from "general" grammar rules towards rules specifying the usage of words on a case-by-case basis, has been followed by a great deal of studies and proposals for the treatment of compound and frozen expressions. Studies presented in [1] and [18] postulate that MWUs embody general grammatical rules and obey to flexibility constraints.

The automatic extraction of multiword lexical units from specialized language *corpora* is an important issue. However, most of these units are not listed in current dictionaries. Multiword lexical units are compound nouns (*Zimbabwean minister of foreign affairs, Fonds Social Européen -the French expression for 'European Social Fund'-, bacalhau à braz -a Portuguese dish-*), frozen phrases (*raining cats and dogs,*

plus ou moins -the French phrase for 'more or less'-, dando que se recebe -a Brazilian expression that might be translated as 'by giving one may receive'), compound verbs (*to take into account, mettre au point –'to fix' in French-, pôr em causa –'to doubt' in Portuguese-*), prepositional locutions (*as a result of, en raison de –'because of' in French-, a partir de –'after' or 'since' in Portuguese*), adverbial locutions *(from time to time, dès que possible –the French expression for 'as soon as possible'-, por exemplo –the Portuguese phrase for 'for instances'-*). It is clear that such units should automatically be extracted from *corpora*, in order to enable their rapid incorporation into NLP specialized lexica. Such dynamic lexical databases would enable parsers to be more effective and efficient. Moreover, MWUs and relevant expressions may be used for refining information retrieval searches [25], enhancing precision, recall and the naturalness of the resulting interaction with the user.

Besides, information about the structure of MWUs should also be available in the NLP lexica. Indeed, one should not only find contiguous MWUs (i.e. uninterrupted sequences of words) but also non-contiguous MWUs (i.e. fixed sequences of words interrupted by one or several gaps filled in by interchangeable words that usually are synonyms). Non-contiguous MWUs may be exemplified by the following sequences: *a total* _____ *of* where the gap may be fullfilled by nouns like *cost* or *population*, *fournir* ____ _____ *sur* (i.e. a French compound verb for '*to give something about someone'*) where the gaps may be filled in with possible morpho-syntactic sequences Article+Noun such as *des informations* (i.e. *some informations*) and *um* _____ *número de* (i.e. a Portuguese noun phrase for '*a number of*') where the gap may be instantiated by occurrences of Adjectives like *determinado* or *certo* (which would result in the English expression '*a determined number of*' or '*a certain number of*'). This kind of information, if it was available in lexica, it would greatly help on attachment decision and as a consequence it would increase the precision of parsers.

The research community has adopted four distinct policies in order to retrieve MWUs. Some approaches only extract contiguous multiword lexical units and require language-dependent information such as part-of-speech tags and base their analysis on syntactical regularities or linguistic resources such as dictionaries ([11], [7] and [3]). In order to scale up the acquisition process, other language-dependent approaches combine shalow morpho-syntactic information with statistics in order to identify syntactical regularities and then select the most probable candidate sequences of words ([16], [21] and [12]). Some other language-dependent systems prefer to use in a first stage statistical techniques to calculate how correlated (associated, aggregated) are the words of a bigram and then apply frequency or/and correlation thresholds ([28] and [10]) in order to extract candidate units. The candidates are then pruned by using morpho-syntactic information. Finally, some purely statistical approaches propose language-independent techniques for the extraction of contiguous and non-contiguous multiword lexical units. They evidence regularities by means of association measure values that evaluate the mutual attraction or "glue" that stands between words in a sequence ([9], [29], [8], and [23]). However, the systems presented so far in the literature rely on ad hoc establishment of frequency or/and association measure thresholds that are prone to error. Indeed, thresholds pose important empirical problems related to their value that depends on the *corpus* size and other factors introduced by the researcher [29]. Besides, the proposed statistical

measures are usually not applied to generic *n*-grams[1] ($n \geq 2$) as they are limited to bigrams.

In this paper, we propose two systems based exclusively on statistical methodologies that retrieve from naturally occurring text, contiguous and non-contiguous MWUs. In order to extract the MWUs, two new association measures, the Symmetric Conditional Probability (SCP) and the Mutual Expectation (ME) are used by a new multiword lexical unit acquisition process based on the *LocalMaxs* algorithm [24]. The proposed approaches cope with two major problems evidenced by all previous works in the literature: the definition of ad hoc frequency and/or association measure thresholds used to select MWUs among word groups and the limited application of the association measures (considering the length of *n*-gram). The introduction of the *LocalMaxs* algorithm that relies on local maxima for the association measure (or "glue") of every *n*-gram, avoids the classical problem of the definition of global thresholds. So, our methodology does not require the definition of any threshold. Moreover, two normalization processes are introduced in order to accommodate the MWU length factor. So, both approaches measure not only the "glue" within each bigram but also within every *n*-gram, with $n > 2$.

In order to extract contiguous MWUs we used the SCP measure and the *LocalMaxs* algorithm since the SCP measure has shown appropriate for capturing contiguous compound nouns, proper names, and other compound sequences recognized as “natural” lexical units. For the case of the non-contiguous MWUs, we used the ME measure and the *LocalMaxs* algorithm, since this measure shows great ability to capture collocations and other non-contiguous multiword lexical units. In the next sections, we will use one of these measures depending on the kind of MWUs we want to extract (i.e. contiguous or non-contiguous MWUs), and we compare their performances with other well-known statistics that are previously normalized: the Specific Mutual Information [9], the ϕ^2 [17], the Dice coefficient [27] and the Log-Likelihood ratio [15].

In the second section, we present the *LocalMaxs* algorithm for the election of MWUs. In the sections 3 and 4 we expand on the SCP and ME measures and include the normalization used by each measure. Using multilingual parallel *corpora* of political debates[2] and Portuguese *corpora*, in the fifth and sixth sections, we respectively show the results for the contiguous and non-contiguous MWUs and compare both measures with the association measures mentionned above (Specific Mutual Information, ϕ^2, Dice coefficient and the Log-Likelihood ratio). In the seventh section we make the assessment of related work. Finally, in the eighth section we present conclusions and future work.

---

[1] An n-gram is a group of words in the *corpus*. We use the notation $[w_1 \ldots w_n]$ or $w_1 \ldots w_n$ to refer the *n*-gram of length *n*.

[2] The corpus has been extracted from the European Parliament multilingual debate collection which has been purchased from the European Language Resources Association (ELRA) - http://www.icp.grenet.fr/ELRA/home.html.

2 The *LocalMaxs* Algorithm

Most of the approaches proposed for the extraction of multiword lexical units are based on association measure thresholds ([9], [12], [23] and [27]). This is defined by the underlying concept that there exists a limit association measure that allows one to decide whether an n-gram is a MWU or not. But, these thresholds can only be justified experimentally and so are prone to error. Moreover, the thresholds may vary with the type, the size and the language of the document and vary obviously with the association measure. The *LocalMaxs* algorithm [24] proposes a more robust, flexible and fine tuned approach for the election of MWUs.

The *LocalMaxs* algorithm works based on the idea that each *n*-gram has a kind of "glue" sticking the words together within the *n*-gram. Different *n*-grams usually have different "glues". As a matter of fact one can intuitively accept that there is a strong "glue" within the bigram [*Margaret, Thatcher*] i.e. between the words *Margaret* and *Thatcher*. On the other hand, one can not say that there is a strong "glue" for example within the bigram [*if, every*] or within the bigram [*of, two*]. So, let us suppose we have a function $g(.)$[3] that measures the "glue" of each *n*-gram. The *LocalMaxs* is an algorithm that works with a *corpus* as input and automatically produces MWUs from that *corpus.*

The *LocalMaxs* algorithm elects the multiword lexical units from the set of all the cohesiveness-valued *n*-grams based on two assumptions. First, the association measures show that the more cohesive a group of words is, the higher its score[4] will be. Second, MWUs are highly associated localized groups of words. As a consequence, an *n*-gram, *W*, is a MWU if its association measure value, *g(W)*, is a local maximum. Let's define the set of the association measure values of all the (*n-1*)-gram contained in the *n*-gram *W*, by Ω_{n-1} and the set of the association measure values of all (*n+1)*-grams containing the *n*-gram *W*, by Ω_{n+1}. The *LocalMaxs* algorithm is defined as follows:

Algorithm 1: The *LocalMaxs*

$$\forall x \in \Omega_{n-1}\ ,\ \forall y \in \Omega_{n+1} \qquad W \text{ is a MWU if}$$
$$(\text{length}(W) = 2 \text{ and } g(W) > g(y)) \text{ or}$$
$$(\text{length}(W) > 2 \text{ and } g(x) \leq g(W) \text{ and } g(W) > g(y))$$

So, an *n*-gram will be a MWU if its $g(.)$ value under that association measure corresponds to a local maximum, as it is shown in Fig. 1.

The reader will notice that, for the contiguous case, the Ω_{n-1} set is reduced to the association measure values of the following two (*n-1*)-grams: $[w_1 \ldots w_{n-1}]$ and $[w_2 \ldots w_n]$. And, the Ω_{n+1} set is reduced to the association measure values of all the

[3] We will write $g(W)$ for the $g(.)$ value of the generic *n*-gram *W* and $g([w_1 \ldots w_n])$ for the $g(.)$ value of the *n*-gram $[w_1 \ldots w_n]$ once we want to keep $g(.)$ as a one-argument function. We will instantiate this generic function by using various *n*-gram word association functions, namely *SCP(.), ME(.),* that will be one-argument functions too. So, we can write for example *ME(W),* $SCP([w_1,w_2,w_3])$, $SCP([w_1 \ldots w_n])$, etc…

[4] The entropy measure used by Shimohata [23] is one of the exceptions.

contiguous (*n+1*)-grams that contain the contiguous *n*-gram *W*. For the non-contiguous case, there are no such restrictions for the Ω_{n-1} and the Ω_{n+1}. All the possible combinations of (*n-1*)-grams and (*n+1*)-grams related with *W* are taken into account.

The *LocalMaxs* algorithm avoids the ad hoc definition of any global association measure threshold and focuses on the identification of local variations of the association measure values. This methodology overcomes the problems of reliability and portability of the previously proposed approaches. Indeed, any association measure that shares the first assumption (i.e. the more cohesive a group of words is, the higher its score will be) can be tested on this algorithm. For the purpose of our study, we applied the *LocalMaxs* algorithm to the Symmetrical Conditional Probability, the Mutual Expectation, the Specific Mutual Information, the ϕ^2, the Dice coefficient and the Log-Likelihood ratio.

One other interesting property of the *LocalMaxs* algorithm is the fact that it elects multiword lexical units on a localized basis allowing the extraction of MWUs formed by the justaposition of MWUs.[5]

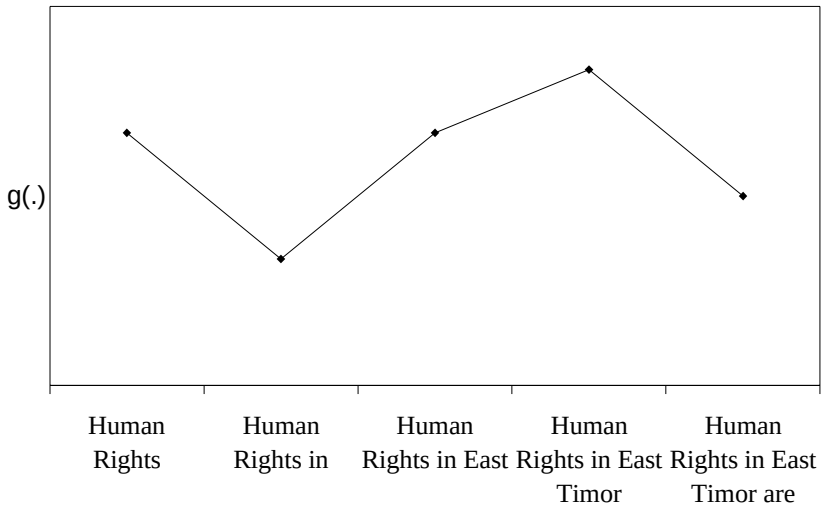

Fig1. The "glue" values of the *n*-grams

For example, the algorithm will elect as MWUs the *n*-grams *Human Rights* and *Human Rights in East Timor* as they are linked to local maxima, as shown in Fig. 1. Roughly exemplifying, the *g(.)* value of *Human Rights* is higher than the *g(.)* of *Human Rights in*, since in current text many unigrams could follow the bigram *Human Rights* (not only the unigram *in*) and many bigrams may precede the unigram *in*. The *g(.)* value for the *4*-gram *Human Rights in East* is higher than the *g(.)* of the *3*-

[5] This property points at a partial solution of the problem of the overcomposition by justaposition ilustrated by [12].

gram *Human Rights in*, however it will not be elected as MWU because the *5*-gram *Human Rights in East Timor* that contains the previous *4*-gram has a higher *g(.)* value. This *5*-gram will be elected since there are neither *4*-grams contained in that *5*-gram nor *6*-grams containing the same *5*-gram with a higher *g(.)* value. Although it is not mentioned here, the *n*-gram *East Timor* is also elected.

3 Extracting Contiguous MWUs from *Corpora*

We have used three tools [24] that work together in order to extract contiguous MWUs from any *corpus*:

-The *LocalMaxs* algorithm
-The Symmetric Conditional Probability (SCP) statistical measure
-The Fair Dispersion Point Normalization

3.1 The Symmetrical Conditional Probability Measure

Let's consider the bigram *[x,y]*. We say that the "glue" value of the bigram *[x,y]* measured by *SCP(.)* is:

$$SCP([x,y]) = p(x \mid y).p(y \mid x) = \frac{p(x,y)^2}{p(x) \cdot p(y)} \quad . \tag{3.1}$$

where *p(x,y)*, *p(x)* and *p(y)* are respectively the probabilities of occurrence of the bigram *[x,y]* and the unigrams *[x]* and *[y]* in the *corpus*; *p(x|y)* stands for the conditional probability of occurrence of *x* in the first (left) position of a bigram given that *y* appears in the second (right) position of the same bigram. Similarly *p(y|x)* stands for the probability of occurrence of *y* in the second (right) position of a bigram given that *x* appears in the first (left) position of the same bigram.

3.2 The Fair Dispersion Point Normalisation

Considering the denominator of the equation (3.1), we can think about any *n*-gram as a "pseudo-bigram" having a left part *[x]* and a right part *[y]*. The Fair Dispersion Point Normalization or simply Fair Dispersion "transforms" any *n*-gram of any size in a "pseudo-bigram" and embodies all the possibilities to have two adjacent groups of words from the whole original *n*-gram. Thus, applying the Fair Dispersion Point Normalisation to *SCP(.)* in order to measure the "glue" of the *n*-gram $[w_1 \ldots w_n]$, we substitute the denominator of the equation (3.1) by *Avp* defined in Equation (3.2):

$$Avp = \frac{1}{n-1} \sum_{i=1}^{i=n-1} p(w_1 \ldots w_i) \cdot p(w_{i+1} \ldots w_n) \, . \tag{3.2}$$

So, we have the normalized SCP defined in Equation (3.3):

$$SCP\_f([w_1 \ldots w_n]) = \frac{p(w_1 \ldots w_n)^2}{Avp} \, . \tag{3.3}$$

As notation matters, in *SCP_f(.)*, we have added *"_f"* for "fair" (from Fair Dispersion*)* to *SCP(.).* As it has shown in [24], the Fair Dispersion Point Normalization concept can be applied to other statistical measures in order to obtain a "fair" measure of the association or "glue" of any *n*-gram of size longer than 2.

4 Extracting Non-contiguous MWUs from *Corpora*

We have used four tools ([24] and [13]) that work together in order to extract non-contiguous MWUs from any *corpus*:

-The *LocalMaxs* algorithm
-The Normalized Expectation measure
-The Fair Point of Expectation
-The Mutual Expectation (ME) statistical measure

4.1 The Normalized Expectation Measure

We define the normalized expectation existing between *n* words as the average expectation of the occurrence of one word in a given position knowing the occurrence of the other *n-1* words also constrained by their positions. The basic idea of the normalized expectation is to evaluate the cost, in terms of cohesiveness, of the possible loss of one word in an *n*-gram. The more cohesive a word group is, that is the less it accepts the loss of one of its components, the higher its normalized expectation will be.

The underlying concept of the normalized expectation is based on the conditional probability defined in Equation (4.1). The conditional probability measures the expectation of the occurrence of the event $X=x$ knowing that the event $Y=y$ stands. $p(X=x,Y=y)$ is the joint discrete density function between the two random variables X, Y and $p(Y=y)$ is the marginal discrete density function of the variable Y.

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}. \tag{4.1}$$

The Fair Point of Expectation. Naturally, an *n*-gram is associated to *n* possible conditional probabilities. It is clear that the conditional probability definition needs to be normalized in order to take into account all the conditional probabilities involved in an *n*-gram.

Let's take the *n*-gram $[w_1\ p_{12}\ w_2\ p_{13}\ w_3\ ...\ p_{1i}\ w_i\ ...\ p_{1n}\ w_n]$ where p_{1i}, for $i=2,...,n$, denotes the signed distance that separates word w_i from word w_1[6]. It is convenient to consider an *n*-gram as the composition of *n sub*-(*n-1*)-grams, obtained by extracting one word at a time from the *n*-gram. This can be thought as giving rise to the occurrence of any of the *n* events illustrated in Table 1 where the underline denotes

[6] This *n*-gram is equivalent to $[w_1\ p_{12}\ w_2\ p_{23}\ w_3\ ...\ p_{2i}\ w_i\ ...\ p_{2n}\ w_n]$ where $p_{2i} = p_{1i} - p_{12}$ for $i=3,...,n$ and p_{2i} denotes the signed distance that separates word w_i from word w_2.

the missing word from the *n*-gram. So, each event is associated to a respective conditional probability. One of the principal intentions of the normalization process is to capture in just one measure all the *n* conditional probabilities. One way to do it, is to blueprint the general definition of the conditional probability and define an average event for its conditional part, that is an average event *Y=y*.

Table 1. Sub-(*n-1*)-grams and missing words

Sub-(n-1)-gram	Missing word
[_____ w_2 p_{23} w_3 ... p_{2i} w_I ... p_{2n} w_n]	w_1
[w_1 _____ p_{13} w_3 ... p_{1i} w_i ... p_{1n} w_n]	w_2
...	...
[w_1 p_{12} w_2 p_{13} w_3...$p_{1(i-1)}$ $w_{(i-1)}$ _____ $p_{1(i+1)}$ $w_{(i+1)}$...p_{1n} w_n]	w_I
...	...
[w_1 p_{12} w_2 p_{13} w_3 ... p_{1i} w_I ... $p_{1(n-1)}$ $w_{(n-1)}$ _____]	w_n

Indeed, only the *n* denominators of the *n* conditional probabilities vary and the *n* numerators remain unchanged from one probability to another. So, in order to perform a sharp normalization process, it is convenient to evaluate the gravity center of the denominators thus defining an average event called the fair point of expectation (FPE). Basically, the FPE is the arithmetic mean of the *n* joint probabilities[7] of the *n* (*n-1*)-grams contained in an *n*-gram. The fair point of expectation for an *n*-gram is defined in Equation (4.2).

$$FPE([w_1\ p_{12}\ w_2 \ldots p_{1i}\ \ w_i \ldots p_{1n}\ \ w_n]) = \frac{1}{n}\left(p([w_2 \ldots p_{2i}\ \ w_i \ldots p_{2n}\ w_n]) + \sum_{i=2}^{n} p\left(\left[w_1 \ldots \hat{p}_{1i}\ \ \hat{w}_i \ldots p_{1n}\ w_n \right]\right)\right). \tag{4.2}$$

$p([w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n])$, for $i=3,\ldots,n$, is the probability of the occurrence of the (*n-1*)-gram $[w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]$ and $p\left(\left[w_1 \ldots \hat{p}_{1i}\ \ \hat{w}_i \ldots p_{1n}\ w_n \right]\right)$ is the probability of the occurrence of one (*n-1*)-gram containing necessarily the first word w_1. The "^" corresponds to a convention frequently used in Algebra that consists in writing a "^" on the top of the omitted term of a given succession indexed from *1* to *n*.

Hence, the normalization of the conditional probability is achieved by introduction the FPE into the general definition of the conditional probability. The symmetric

[7] In the case of *n = 2*, the FPE is the arithmetic mean of the marginal probabilities.

resulting measure is called the normalized expectation and is proposed as a "fair" conditional probability. It is defined in Equation (4.3) [8].

$$NE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = \frac{p([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])}{FPE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n])}. \quad (4.3)$$

p([w₁...p₁ᵢwᵢ...p₁ₙwₙ]) is the probability of the *n*-gram *[w₁...p₁ᵢwᵢ... p₁ₙ wₙ]* occurring among all the other *n*-grams and *FPE([w₁... p₁ᵢ wᵢ ... p₁ₙ wₙ])* is the fair point of expectation defined in Equation (4.2).

4.2 The Mutual Expectation Measure

[23] shows that one effective criterion for multiword lexical unit identification is simple frequency. From this assumption, we deduce that between two *n*-grams with the same normalized expectation, that is with the same value measuring the possible loss of one word in an *n*-gram, the most frequent *n*-gram is more likely to be a multiword unit. So, the Mutual Expectation between *n* words is defined in Equation (4.4) based on the normalized expectation and the simple frequency. It is a weighted normalized expectation.

$$ME([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) = f([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]) \times NE([w_1 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]). \quad (4.4)$$

f([w₁ ... p₁ᵢ wᵢ ... p₁ₙ wₙ]) and *NE([w₁ ... p₁ᵢ wᵢ ... p₁ₙ wₙ])* are respectively the absolute frequency of the particular *n*-gram *[w₁ ... p₁ᵢ wᵢ ... p₁ₙ wₙ]* and its normalized expectation.

It should be stressed that, apart from representational differences related to our objectives –to extract contiguous and not necessarily contiguous *n*-grams- which gave rise to two representations –one with word positional information (section 4)- there are important differences between (3.3) and (4.4). The numerators are identical for contiguous *n*-grams. The denominators, (3.2) and (4.2), are different and are obtained assuming different smoothing strategies, due to initial research objectives.

5 Results for Contiguous MWUs

5.1 Comparing *SCP_f* with other Statistics-Based Measures

In order to assess the results given by the *SCP_f* measure, several measures were tested using the *Fair Dispersion Point Normalisation*[9] including: the Specific Mutual Information (SI) ([9], [10] and [2]), the SCP [24], the Dice coefficient [27], the

[8] The Normalized Expectation measure is different from the Dice coefficient introduced by [27] although they share the same expression for the case of bigrams.

[9] As a matter of fact, any *n*-gram can be divided in a left and a right part choosing any point between two adjacent words within the *n*-gram. In this way, one can measure the "glue" using some usual statistical measure without the *Fair Dispersion*, but the results are relatively poor. The enhancements obtained in Precision and Recall when the *Fair Dispersion* is introduced in several statistical measures are shown in [24].

Loglikelihood ratio [15], and the ϕ^2 coefficient [17]. Table 2 contains scores for these statistical measures working with the Fair Dispersion Point Normalisation and the *LocalMaxs* algorithm. We have used an average size *corpus* (919,253 words)[10].

The Evaluation Criterion. The *LocalMaxs* algorithm extracts *n*-grams, which are potential MWUs or relevant expressions. In order to decide if an extracted *n*-gram is a MWU or relevant expression or not, we considered as correct ones: proper names, such as *Yasser Arafat, Republica Centro Africana*[11]*, etc.*; compound names such as c*âmara municipal de Reguengos de Monsaraz (Reguengos de Monsaraz town hall), convenção dos Direitos Humanos (Human Rights convention)*, etc.; compound verbs such as levar a cabo *(to get, to carry out, to implement), ter em conta (to take into account)*, etc.; frozen forms such as *em todo o caso (anyway), segundo consta (according with what is said)*, etc., and other *n*-grams occurring relatively frequently and having strong "glue" among the component words of the *n*-gram such as *tanta e tão boa (so much and so good), afectadas pela guerra civil (afflicted by the civil war).*

The Results.

Table 2: Scores obtained by assigning several statistics-based association measures

Statistics-based measure: g(.)=	Precision (average)	Extracted MWUs (count)
SCP_f(.)	81.00%	24476
SI_f(.)	75.00%	20906
ϕ^2_f(.)	76.00%	24711
Dice_f(.)	58.00%	32381
LogLike_f(.)	53.00%	40602

The Precision column means the average percentage of correct MWUs obtained. It is not possible to calculate the exact number of MWUs in the *corpus*. So, we may measure how close to that number is the number of MWUs obtained by each statistical measure. As a matter of fact we are not facing the problem of counting very well defined objects like nouns or verbs of a *corpus*, but counting MWUs. So, the column Extracted MWUs, which gives the number of extracted MWUs by the considered measure[12], works as an indirect measure of Recall.

[10] This *corpus* corresponds to the news of some days in January 1994 from *Lusa* (the Portuguese News Agency).

[11] Note the spelling error in 'Republica' that should have been written as 'República'. However real *corpus* is like that and we can not escape from it as there are texts that may reproduce parts of other texts where the graphical form of words does not correspond to currently accepted way of writing.

[12] We have discarded hapaxes, every "MWU" or "relevant expression" that occurred just once.

Although there are very large MWUs, for example the *8*-gram *Presidente da câmara municipal de Reguengos de Monsaraz,* we have limited the MWUs produced by the *LocalMaxs* from *2*-grams to *7*-grams for reasons of processing time.

Dicussion of the Results. As we can see from Table 2, the *SCP_f* measure gets the best Precision and a comparatively a good value for Extracted MWUs. By using the *LocalMaxs* algorithm with any of the statistics-based measures *SCP_f*, *SI_f* or ϕ^2*_f*, a good Precision is obtained. However *SI_f* has a relative lower score for Extracted MWUs (count). The *Dice_f* and specially the *Loglike_f* measure showed not to be very selective. They extract many expressions (high values for MWUs (count)), but many of them are not relevant, they just have high frequency such as *dar ao (to give to the), dos outros (of the others),* etc... Moreover, as it is discussed in [26], *Dice_f* and *Loglike_f* measures do extract a lot of uninteresting units and fail to extract other interesting units that are selected by the other three word association measures. Thus, we have chosen the *SCP_f* measure to work with *LocalMaxs* algorithm in order to extract contiguous MWUs from *corpora.*

5.2 Extracting Contiguous MWUs from Different Languages

We have also tested the *LocalMaxs* and the *SCP_f* for different languages on non-annotated *corpora*, and we have obtained the following results:

Table 3. *LocalMaxs* and *SCP_f* scores for different languages

Language	Precision	Extracted MWUs (count)	*Corpus* size
English	77.00%	8017	493191
French	76.00%	8980	512031
German	75.00%	5190	454750
Medieval Portuguese	73.00%	5451	377724

The MWUs of Table 3 were obtained without any morpho-syntactic operation or linguistic filter. Although the Precision is not the same for the different languages in Table 3, we think this may be due to the different *corpus* sizes –remember that in the case of the Portuguese non-annotated *corpus* (See Table 2), the *corpus* size is 919,253 words and we have got 81% precision. Thus, we believe that for a larger *corpus*, similar precision measures may be attained for different languages.

5.3 Extracting Heavily Inflected Multiword Lexical Units

Verbs, in Portuguese and other Latin languages are heavily inflected. They vary in number, person, gender, mode, tense, and voice. So, in a *corpus* we can find phrases such as *ele teve em conta que... (he has taken into account that...), ele tem em conta o preço (he takes into account the price...), eles tinham em conta essas coisas (they took into account those things), isso foi tomado em conta porque (that was taken into*

account because), etc... As a consequence, due to the fact that the same verb can occur in different forms, it might be the case that we were not extracting every possible multiword verb phrase. So we needed to have every occurrence of any verb in a *corpus* lemmatized to its infinitive form.

In order to obtain that, we used an automatically tagged *corpus* from which we produced the equivalent text by changing just the verb forms to the corresponding infinitive forms. Acting this way, the infinitive verb forms get relevance in the *corpus,* avoiding the dispersion by several forms. This results in higher "glue" values for the *n*-grams containing verbs and words with a strong association among them. An existing neural network based tagger [22] has been used for tagging a superset of the previous Portuguese *corpus*[13] (i.e. a superset of the one containing 919,253 words used before). The tagger disambiguates the POS tags assigned to each word. Every word is tagged and its base form (singular for nouns, singular masculine for adjectives, infinitive for verbs, etc.) is also provided by the tagger. Then, in order to obtain the *corpus* we wanted, the verbs were changed to its infinitive forms except those in the past participle since they are usually used as adjectives. So, except for the verb forms that are not in the past participle, the rest of the words were kept as they were in the original *corpus.*

The Evaluation Criterion. In order to evaluate the results obtained by applying the *LocalMaxs* algorithm and the *SCP_f* word association measure to the transformed *corpus*, we need to remind that contiguous compound verbs may conform to a set of patterns. Generally these patterns have two or three words:

-Verb+Prep+Noun (*pôr em causa –to doubt-, levar a cabo –to get-, ter em atenção –to beware to-, entrar em vigor –to come into force-, ter por objectivo –to aim-, etc.)*

-Verb+Adv+Noun|Adv (*ter como lema –to follow-, ir mais longe –to reach farther-, etc.)*

-Verb+Contraction+Noun (*ter pela frente –to face-, subir ao poder –to reach the power-, etc.)*

-Verb+Prep+Verb (*estar para chegar –to be about arriving-, etc.*)

-Verb+Noun|Adv (*correr bem –to get succeed-, arredar pé –to leave-, marcar passo –to stay-, pôr cobro –to put a stop to- etc.*)

-Verb+Adj|Adv (*tornar possível –to enable-, tornar público –to divulg, tornar claro –to clarify-, etc.*

However, there are many statistically relevant expressions beginning with a verb that might or might not be considered contiguous compound verbs. In any case, NLP lexica should take them into account, since there is a strong co-occurrence between the verb and the rest of the *n*-gram. That would be the case of *angariar fundos –to get funding, to raise funds-, criar emprego –to create jobs-, efectuar contactos –to contact-, fazer sentir –to convince-, viver na miséria –to live in extreme poverty-, aceitar sem reservas –to accept without reserves-, aplicar a pena –to apply the*

[13] This *corpus* was made from an original *corpus* corresponding to the news of some days in January 1994 from *Lusa* (the Portuguese News Agency).

punishment-, etc.… For the purpose of evaluation we have also considered these kind of relevant expressions as correct contiguous compound verbs.

The Results.

Table 4. The scores for the contiguous compound verbs extractions

Form	Precision	Extracted compound verbs[14]
2-gram	81.00%	108
3-gram	73.00%	492

Discussion of the Results. Table 4 shows us respectable values for Precision. Once again, there is not a practical way to calculate the total number of the compound verbs existing in the *corpus*, so we can evaluate how close to that number is the number of *Extracted compound verbs* obtained by our approach (Recall). However, we would say that 600 (108 + 492) compound verbs extracted from a 1,194,206 words *corpus* is a good score, but we believe that a larger *corpus* will enable to extract an amount of compound verbs which must be closer to the number of compound verbs of the Portuguese language. Although the high performance of the *Neuronal Tagger* (about 85% Precision for verbs and 95% Recall[15]), the scores in Table 4 depend also on the tagger performance. Appendix A contains a sample of the compound verbs extracted by our approach.

6 Results for Non-contiguous MWUs

In this section, we first compare the results obtained by applying the *LocalMaxs* algorithm over a Portuguese *corpus* of political debates with approximately 300,000 words[16] with the Mutual Expectation (ME), the normalized Specific Mutual Information (*SI_n*)[17], the normalized ϕ^2 (ϕ^2*_n*)[18], the normalized Dice coefficient (*Dice_n*)[19] and the normalized Log-Likelihood ratio (*Loglike_n*)[20]. The results illustrate that the Mutual Expectation leads to very much improved results for the specific task of non-contiguous multiword lexical unit extraction as it is shown in Table 5.

[14] Remind the evaluation criterion explained before.

[15] Precision for POS-tagging is greater than these number sugests. As a matter of fact there are tags that are assigned correctly 100% of the times. Verbal tags, specially the past participle verbal tag is rather problematic as the corresponding word most of the times behave as an adjective. For a deeper analysis on this subject matter see [22].

[16] The authors are aware that the size of this corpus is relatively small. However, we must point at the fact that working with normalized events reduces the corpus length side effect factor.

[17] The *SI_n* is the result of the normalization process of the association ratio [9].

[18] The ϕ^2*_n* is the result of the normalization process of the Pearson's coefficient [17].

[19] The *Dice_n* is the result of the normalization process of the Dice coefficient [27].

[20] The *Loglike_n* is the result of the normalization process of the Log-likelihood [15].

6.1 The Evaluation Criterion

Table 5. Scores obtained by assigning several statistics-based word association measures

Statistics-based measure: g(.)=	Precision (average)	Extraction of correct MWUs (count)
ME(.)	90.00%	1214
SI_n(.)	61.00%	276
ϕ^2_n(.)	70.00%	294
Dice_n(.)	48.00%	474
LogLike_n(.)	49.00%	1044

We first built all the contiguous and non-contiguous *n*-grams (for *n=1* to *n=10*) from the Portuguese *corpus* and applied to each one its respective association measure value and finally ran the *LocalMaxs* algorithm on this data set. In the case of the extracted non-contiguous MWUs, we analyzed the results obtained for units containing exactly one gap leaving for further study the analysis of all the units containing two or more gaps. Indeed, the relevance of such units is difficult to judge and a case by case analysis is needed. However, the reader may retain the basic idea that the more gaps there exists in a non-contiguous MWU the less this unit is meaningful and the more it is likely to be an incorrect multiword lexical unit. Another important point concerning precision and recall rates has to be stressed before analysing the results. There is no consensus among the research community about how to evaluate the output of multiword lexical unit extraction systems. Indeed, the quality of the output strongly depends on the task being tackled. A lexicographer and a translator may not evaluate the same results in the same manner. A precision measure should surely be calculated in relation with a particular task. However, in order to define some "general" rule to measure the precision of the system, we propose the following two assumptions. Non-contiguous multiword lexical units are valid units if they are relevant structures such as *pela* _____ *vez* (which could be translated in English by 'for the ____ time') where the gap may be filled-in by occurrences of *primeira* (*first*), *segunda* (*second*) etc…; if they are collocations such as *tomar* _____ *decisão* (where the English equivalent would be '*to take* ____ *decision*') where the gap may be filled in with the articles *uma* (*a*) or *tal* (*such a*). Finally, a non-contiguous MWU is a valid unit if the gap corresponds to at least the occurrence of two different tokens in the *corpus*. For example, the following non-contiguous *n*-gram *Prémio Europeu* _____ *Literatura* does not satisfy our definition of precision as the only token that appears in the *corpus* at the gap position is the preposition *de*. Furthermore, the evaluation of extraction systems is usually performed with the well-established recall rate. However, we do not present the "classical" recall rate in this experiment due to the lack of a reference *corpus* where all the multiword lexical units are identified. Instead, we present the number of correctly extracted non-contiguous MWUs.

6.2 The Discussion of the Results

From the results of Table 2 and Table 5, one can acknowledge that the non-contiguous rigid multiword units are less expressive in this sub-language than are the contiguous multiword units. Nevertheless, their average frequency is very similar to the one of the extracted contiguous multiword units showing that they do not embody exceptions and that they reveal interesting phenomena of the sub-language. Some results are given in Appendix B.

The Mutual Expectation shows significant improvements in terms of Precision and Recall in relation with all the other measures. The most important drawback that we can express against all the measures presented by the four other authors is that they raise the typical problem of high frequency words as they highly depend on the marginal probabilities. Indeed, they underestimate the degree of cohesiveness when the marginal probability of one word is high. For instance, the *SI_n*, the *Dice_n*, the ϕ^2*_n* and the *Loglike_n* elect the non-contiguous multiword lexical unit *turcos _____ curdos* (*Turkish _____ Kurdish*) although the probability that the conjunction *e* (*and*) fills in the gap is one. In fact, the following *3*-gram [*turcos 1 e 2 curdos*] gets unjustifiably a lower value of cohesiveness than the *2*-gram [*turcos 2 curdos*]. Indeed, the high frequency of the conjunction *e* underestimates the cohesiveness value of the *3*-gram. On the opposite, as the Mutual Expectation does not depend on marginal probabilities except for the case of *2*-grams, it elects the longer MWU *refugiados políticos turcos e curdos*, correspondingly to the concordances output exemplified in Table 6. So, all the non-contiguous multiword lexical units extracted with the Mutual Expectation measure define correct units as the gaps correspond to the occurrences of at least two different tokens. The problem shown by the other measures is illustrated by low precision rates[21].

Table 6: Concordances for *turcos _____ curdos*

greve da fome sete	refugiados políticos turcos e curdos	na Grécia sete
greve da fome dos	refugiados políticos turcos e curdos	em protesto
dos sete	refugiados políticos turcos e curdos	que fazem greve de
na Grécia sete	refugiados políticos turcos e curdos	que estão detidos e
semanas onde sete	refugiados políticos turcos e curdos	estão presos em

6.3 Extracting Non-contiguous MWUs from Different Languages

A comparative study over a Portuguese, French, English and Italian parallel *corpus* has been carried out in [14] and has illustrated that the concept of multiword lexical unit embodies a great deal of cross-language regularities beyond just flexibility and grammatical rules, namely occurrence and length distribution consistencies. In all cases, the Mutual Expectation gave the most encouraging results.

[21] A more detailed analysis can be found in [13].

7 Assessment of Related Work

Several approaches are currently used in order to automatically extract relevant expressions for NLP lexica purposes and Information Retrieval. In this section, we discuss some related work focusing the extraction of these expressions and some of its applications.

In non statistical approaches, syntactic patterns of occurrence that generally enable the retrieval of adequate compounds are searched. Generally they do not go beyond *3*-grams. For example, [4] and [5] search for those patterns in partially parsed *corpora* (treebanks). However, Barkema recognizes that the occurrence of a pattern does not necessarily mean that compound terms have been found. [20] also use this kind of pattern matching and then generate variations by inflection or by derivation and check if those possible units do appear in the *corpus* used. More recent work can be founded in [6]. These approaches soon fall short of available POS-tagged *corpora* and the precision that their approaches enable are surely very low. They rely mostly on human selection.

The works proposed by Barkema and Jacquemin suffer from their language dependency requiring a specialized linguistic analysis to identify clues that isolate possible candidate terms. In order to scale up the acquisition process, [12] explores a method in which co-occurrences of interest are defined in terms of surface syntactic relationships and then are filtered out by means of the likelihood ratio statistics. In a first round, only base-terms (i.e. terms with length two) that match a list of previously determined syntactical patterns (noun phrase patterns) are extracted from a tagged *corpus*. Then, as the patterns that characterize base-terms can be expressed by regular expressions, a finite-automata is used to compute the frequency of each candidate base-term. In order to do so, each base-term is classified into a pair of lemma (i.e. two main items) and its frequency represents the number of times the two lemmas of the pair appear in one of the allowed morpho-syntactic patterns. Finally, the Log-Likelihood ratio statistics is applied as an additional statistical filter in order to isolate terms among the list of candidates. However, the attained precision and recall rates are not presented in Daille's work. Her approach requires a lot of morpho-syntactic work to extract any relevant expression, since the statistical part of it does not measure the correlation of the *n*-grams of length longer than 2. We believe that, by using statistics to measure the "glue" sticking together the whole *n*-gram, whatever the *n*-gram length is, rather than just measuring the "glue" for 2-grams, the same or better results could have been obtained in a more confortable way. Moreover, in Daille's work, the *Loglike* criterion is selected as the best. As Daille, we also think that this result is due to the fact that statistics were applied after the linguistic filters. Indeed, as we can infer from Tables 2 and 5, the precision attained by this criterion, after correction with Fair Dispersion Point Normalisation and the Fair Point of Expectation is rather low.

Linguistic approaches combined or not with statistical methods, present two major drawbacks. By reducing the searching space to groups of words that correspond uniquely to particular noun phrases structures, such systems do not deal with a great proportion of multiword lexical units such as compound verbs, adverbial locutions, prepositional locutions, conjunctive locutions and frozen forms. The study made by [19] shows that they represent 22.4% of the total number of the MWUs in their

specilized *corpus*. Furthermore, [18] points at the fact that multiword lexical units may embody specific grammatical regularities and specific flexibility constraints across domains. As a consequence, linguistic approaches need to be tuned for each new domain of application.

Smadja proposes [29] in the first part of a statistical method (XTRACT) to retrieve collocations by combining *2*-grams whose co-occurrences are greater than a given threshold. In the first stage, pairwise lexical relations are retrieved using only statistical information. Significant *2*-grams are extracted if the *z*-score of a word pair exceeds a threshold that has to be determined by the experimenter and that is dependent on the use of the retrieved collocations. In the second stage, multiple-word combinations and complex expressions are identified. For each *2*-grams identified at the previous stage, XTRACT examines all instances of appearance of the two words and analyzes the distributions of words and parts of speech tags in the surrounding positions. If the probability of the occurrence of a word or a part of speech tag around the *2*-gram being analyzed is superior to a given threshold, then the word or the part of speech is kept to form an *n*-gram. Although, Smadja's methodology is more flexible than the studies previously exposed it relies on ad hoc establishment of association measure thresholds that are prone to error and on association measures only defined for bigrams.

8 Conclusions and Future Work

By conjugating the *LocalMaxs* algorithm with *n*-gram-length normalized association measures, it is possible to extract automatically from raw texts relevant multiword lexical units of any size and thus populate lexica for applications in NLP and Information Retrieval, Information Extraction and Text Mining.

The different natures and structures between contiguous and non-contiguous multiword lexical units has lead us to elaborate two distinct methodologies in order to retrieve each particular kind of units. The studies and experiments showed that two skilled association measures (namely in their smoothing -normalization- techniques) were needed: the SCP with the Fair Dispersion Point Normalization for contiguous MWUs and the ME with the Fair Point of Expectation for non-contiguous MWUs.

The results obtained by comparing both the SCP and the ME with the Specific Mutual Information [9], the ϕ^2 [17], the Dice coefficient [27] and the Log-Likelihood ratio [15] allow us to point at the fact that these new introduced measures are specifically designed for the extraction of MWUs within the concept of local maxima embodied by the *LocalMaxs* algorithm. However, we believe that there are neither universal association measures nor absolute selection processes. Indeed, the sensibility we have acquired also makes us believe that for different applications one may consider different methodologies.

The statistical nature of our approaches confirms that the *LocalMaxs* algorithm is a more robust approach reducing the commitments associated to the complexity of the language. Besides, the local maximum criterion avoids the definition of frequency and/or association measure global thresholds. We had not the time to compare the two smoothing strategies proposed in this paper. So, as future work, we intend to improve our methodologies by cross-studying the results of each association measures and smoothing strategies in order to find the most fruitful combinations.

Appendix A: A Sample of Contiguous Compound Verbs Extraction

Ter em atenção (to beware to)
ter em conta (to have in mind)
tomar uma decisão (to take a decision)
travar de razões (to discuss)
usar a força (to use the police force)
atingir o limite (to attain the limit)
cantar as janeiras (to sing the janeiras)
causar a morte (to cause the death)
estar para durar (to be to last)
entrar em vigor (to come into force)
Levar a cabo (to get)
Levar a efeito (to get)
Pôr em causa (to doubt)
Dar explicações (to explain)
Dar prioridade (to give priority)
Marcar passo (to stay)
Pôr termo (to make an end of)
Tirar conclusões (to conclude)
Tornar público (to divulge)
Trocar impressões (to exchange ideas)

Appendix B: A Sample of Non-contiguous MWUs Extraction

Greve _____ fome (hunger strike)[22]
Progressos _____ registados (_____ regitered improvements)
Tomar _____ decisão (to take _____ decision)
Distorções _____ concorrência (concurrency distorsions; distorsion of the concurrency)
Presença _____ observadores (observators presence; presence of the observators)
Taxas _____ IVA (VAT rates, rates of the VAT)
Um _____ número de (a _____ number of)
Uma lista _____ projectos (projects list; list of the projects)
Transporte de _____ perigosas (transport of dangerous _____)
Proposta de _____ do Conselho (proposal of _____ of the Council)

Acknowledgements

This work was funded by the PhD grant PRAXIS 4/4.1/BD/3895 and the projects "DIXIT – Multilingual Intentional Dialog Systems", Ref. PRAXIS XXI 2/2.1/TIT/1670/95 and "PGR – Acesso Selectivo aos Pareceres da Procuradoria Geral da República" Ref. LO59-P31B-02/97, funded by Fundação para a Ciência e Tecnologia. The information about this projects can be accessed by http://kholosso.di.fct.unl.pt/~di/people.phtml?it=CENTRIA&ch=gpl.

We want to thank the reviewers and the Program Committee for having forced us to merge two papers into this single one. We hope our effort has achieved the results aimed by the Program Committee and reviewers.

References

1. Abeille, A.: Les nouvelles syntaxes: Grammaires d'unification et Analyse du Français, Armand Colin, Paris (1993)
2. Bahl, L., & Brown, P., Sousa, P., Mercer, R.: Maximum Mutual Information of Hidden Markov Model Parameters for Speech Recognition. In Proceedings, International

[22] There is no existing gap in the English translation as the unit is the result of two different ways of writing the same concept: *Greve da fome* and *greve de fome*.

Conference on Acoustics, Speech, and Signal Processing Society, Institute of Electronics and Communication Engineers of Japan, and Acoustical Society of Japan (1986)
3. Blank, I.: Computer-Aided Analysis of Multilingual Patent Documentation, First LREC, (1998) 765-771
4. Barkema, H.: Determining the Syntactic Flexibility of Idioms, in Fries U., Tottie G., Shneider P. (eds.): Creating and Using English Language Corpora, Rodopi, Amsterdam, (1994), 39-52
5. Barkema, H.: Idiomaticy in English Nps, in Aarts J., de Haan P., Oostdijk N. (eds.): English Language Corpora: Design, Analysis and Exploitation, Rodopi, Amsterdam, (1993), 257-278
6. Bourigault, D., Jacquemin, C.: Term Extraction and Term Clustering: an Integrated Platform for Computer Aided Terminology. In Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics, p. 15-22, Bergen, Norway June (1999)
7. Bourigault, D.: Lexter, a Natural Language Processing Tool for Terminology Extraction, 7th EURALEX International Congress, (1996)
8. Chengxiang, Z.: Exploiting Context to Identify Lexical Atoms: a Statistical View of Linguistic Context, cmp-lg/9701001, 2 Jan 1997, (1997)
9. Church, K. et al.: Word Association Norms Mutual Information and Lexicography, Computational Linguistics, Vol. 16 (1). (1990) 23-29
10. Church, K., Gale, W., Hanks, P., Hindle, D.: Using Statistical Linguistics in Lexical Analysis. In Lexical Acquisition: Using On-line Resources to Build a Lexicon, edited by Uri Zernik. Lawrence Erlbaum, Hilldale, New Jersey (1991) 115-165
11. Dagan, I.: Termight: Identifying and Translating Technical Terminology, 4th Conference on Applied Natural Language Processing, ACL Proceedings (1994)
12. Daille, B.: Study and Implementation of Combined Techniques for Automatic Extraction of Terminology. The Balancing Act Combining Symbolic and Statistical Approaches to Language, MIT Press (1995)
13. Dias, G., Gilloré, S., Lopes, G.: Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora. In Proceedings of the TALN'99 (1999).
14. Dias, G., Gilloré, S., Lopes, G.: Multilingual Aspects of Multiword Lexical Units. In Proceedings of the Workshop Language Technologies –Multilingual Aspects, Faculty of Arts, 8-11 July (1999), Ljubljana, Slovenia
15. Dunning, T.: Accurate Methods for the Statistics of Surprise and Coincidence, Association for Computational Linguistics, Vol. 19-1. (1993)
16. Enguehard, C.: Acquisition de Terminologie à partir de Gros Corpus, Informatique & Langue Naturelle, ILN'93 (1993) 373-384
17. Gale, W.: Concordances for Parallel Texts, Proceedings of Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora, Oxford (1991)
18. Habert, B. et al.: Les linguistiques du Corpus, Armand Colin, Paris (1997)
19. Herviou-Picard et al.: Informatiques, Statistiques et Langue Naturelle pour Automatiser la Constitution de Terminologies, In Proc. ILN'96 (1996)
20. Jacquemin, C., Royauté, J.: Retrieving Terms and their Variants in a Lexicalized Unification-Based Framework, in: SIGIR'94, Dublin, (1994) 132-141
21. Justeson, J.: Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text, IBM Research Report, RC 18906 (82591) 5/18/93 (1993)
22. Marques, N.: Metodologia para a Modelação Estatística da Subcategorização Verbal. Ph.D. Thesis. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Lisbon, Portugal, Previewed Presentation (1999) (In Portuguese)
23. Shimohata, S.: Retrieving Collocations by Co-occurrences and Word Order Constraints, Proceedings ACL-EACL'97 (1997) 476-481

24. Silva, J., Lopes, G.: A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In Proceedings of the 6th Meeting on the Mathematics of Language, Orlando, July 23-25 (1999)
25. Silva, J., Lopes, G.: Extracting Multiword Terms from Document Collections. In Proceedings of the VExTAL, Venezia per il Trattamento Automatico delle Lingu, Universiá Cá Foscari, Venezia November 22-24 (1999)
26. Silva, J., Lopes, G., Xavier, M., Vicente, G.: Relevant Expressions in Large Corpora. In Proceedings of the Atelier-TALN99, Corse, july 12-17 (1999)
27. Smadja, F. et al.: Translating Collocations for Bilingual Lexicons: A Statistical Approach, Association for Computational Linguistics, Vol. 22 (1) (1996)
28. Smadja, F.: From N-grams to Collocations: An Evaluation of Extract. In Proceedings, 29th Annual Meeting of the ACL (1991). Berkeley, Calif., 279-284
29. Smadja, F.: Retrieving Collocations From Text: XTRACT, Computational Linguistics, Vol. 19 (1). (1993) 143-177