Mutual Expectation: A Measure for Multiword Lexical Unit Extraction

Gaël Dias, José Gabriel Pereira Lopes
Universidade Nova de Lisboa
Faculdade de Ciencias e Tecnologia
2825-114, Monte da Caparica, Portugal
{ddg,gpl}@di.fct.unl.pt
Tel: (351) 1 295 44 64 ext. 0743
Fax: (351) 1 294 85 41

Sylvie Guilloré
Université d'Orléans
Laboratoire d'Informatique Fondamentale d'Orléans
BP 6102 - 45061,Orléans Cédex 2, France
sylvie.guillore@lifo.univ-orleans.fr
Tel: (33) 2 38 41 72 65
Fax: (33) 2 38 41 71 37

Abstract

Many applications in Information Extraction, Information Retrieval and Machine Translation would greatly benefit if reliable methods for extracting multiword lexical units (MWUs) were available. As a matter of fact, multiword lexical units such as *diritto di voto*, *Federazione russa*, *Amnesty International* and *in materia di* should be taken as indivisible units in the sense that their meaning or function does not necessarily follow from the compositionality of the meaning of their component words. We propose in this paper a language independent statistically-based system that identifies and extracts multiword lexical units.

1 Introduction

The majority of works in Natural Language Processing have traditionally been concerned with the recognition and extraction of explicit information from texts (knowledge about the world) and have generally neglected the extraction of implicit information (knowledge about the language used). Fortunately, the growing amount of available large-scale text corpora has initiated a new era for the retrieving of intrinsic information such as pp-attachment, sub-categorization and multiword lexical units. Multiword lexical units such as *diritto di voto*, *Federazione russa*, *Amnesty International* and *in materia di* represent implicit knowledge included in texts as they are indivisible lexical units in the sense that their meaning or function does not necessarily follow from the compositionality of the meaning of their component words. But, their study has for a long time been relegated to the margins of the lexicographic treatment. However, the extraction of multiword lexical units would enable more precise text processing and as a consequence would lead to an adequate normalization of texts for the extraction of more explicit information. Consequently, many applications in Information Extraction, Information Retrieval and Machine Translation would greatly benefit if reliable methods for extracting multiword lexical units were available.

In this paper, we propose a system based exclusively on a statistical methodology that retrieves, from naturally occurring text, contiguous multiword lexical units (i.e. uninterrupted sequences of words) and non-contiguous rigid multiword lexical units (i.e. sequences of words interrupted by one or several gaps filled in by interchangeable words). In order to extract MWUs, a new association measure based on the concept of normalized expectation, the Mutual Expectation (ME) [Dias1999-1], is conjugated with a new multiword lexical unit acquisition process based on an algorithm of local maxima, the LocalMax algorithm [Silva1999].

The proposed system comprises three stages. The first stage transforms the input text corpus into contingency tables, suitable for statistical analysis, by counting contiguous and non-contiguous *n*-grams. The second stage, presented in section 2, measures the cohesiveness of every *n*-gram by applying the ME measure to all of them. The final stage, exposed in section 3, elects the MWUs from the set of all cohesiveness-valued *n*-grams by using the LocalMax algorithm. In the fourth section, the quality of the extracted multiword units is tested by means of different comparisons with four other association measures over an Italian, Portuguese, English and French parallel corpus of political debates. Finally, the analysis of the results points at a partial solution to the problem of the election of hapaxes (i.e. multiword units with frequency equal to one) by evidencing patterns of multiword lexical units.

2 The Mutual Expectation

The transformation of the input text corpus into contingency tables allows to define mathematical

models that describe the degree of cohesiveness that stands between words. But, the mathematical models (or association measures) presented so far in the literature [Church1990] [Gale1991] [Smadja1993] [Dunning1993] and [Smadja1996] are unsatisfactory as they only evaluate the degree of cohesiveness between two discrete random variables and do not generalize for the case of *n* variables. Moreover, many of these association measures rely too much on the marginal probabilities misevaluating the attraction between words. In order to overcome both problems, we introduce the Mutual Expectation measure (ME) [Dias1999-1] based on a normalized expectation (NE).

2.1 Normalized Expectation

We define the normalized expectation measure existing between *n* words as the average expectation of one word occurring in a given position knowing the presence of the other *n*-1 words also constrained by their positions.

Taking the example of the *2*-gram [*Federazione* +1 *russa*], the normalized expectation measure will evaluate the degree of cohesiveness that stands between *Federazione* and *russa* by calculating the expectation of occurring *Federazione* before *russa* (i.e the expectation of occurring *Federazione* knowing the presence of *russa* constrained by the signed distance -1) and the expectation of appearing *russa* after *Federazione* (i.e the expectation of occurring *russa* knowing the presence of *Federazione* constrained by the signed distance +1). The underlying concept is based on the conditional probability defined in (1).

$$p(X = x \mid Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \qquad \textbf{(1)}$$

where $p(X = x, Y = y)$ 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*.

So, let's take the *n*-gram $[w_1\ p_{12}\ w_2\ p_{13}\ w_3 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]$ where p_{1i}, for i=2,...,n, denotes the signed distance[1] that separates word w_i from word w_1. This *n*-gram is equivalent to $[w_1\ p_{12}\ w_2\ p_{23}\ w_3 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]$ 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. This transformation is necessary, as we will be interested in considering an *n*-gram as the composition of *n sub*-(*n-1*)-gram, obtained from the *n*-gram by extracting one word at a time from it. And this can be thought as giving rise to the occurrence of any of the following *n* events where the underline denotes the missing word from the *n*-gram:

(n-1)-grams	**Missing word**
$[\ \underline{\quad}\ w_2\ p_{23}\ w_3 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n]$	w_1
$[w_1\ \underline{\quad}\ p_{13}\ w_3 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n]$	w_2
...	...
$[w_1 \ldots p_{1(i-1)}\ w_{(i-1)}\ \underline{\quad}\ p_{1(i+1)}\ w_{(i+1)} \ldots p_{1n}\ w_n]$	w_i
...	...
$[w_1 \ldots p_{1i}\ w_i \ldots p_{1(n-1)}\ w_{(n-1)}\ \underline{\quad}\ ]$	w_n

So, we are interested in the set of all the *n* conditional probabilities measuring the expectation of one word occurring knowing that the other ones occur in the *n*-gram in constrained positions. For our purpose, we need to capture in just one measure all the conditional probabilities. One way to solve this problem is to define one average event defining the conditional part of the probability (i.e. the *Y=y* event). The fair point of expectation (FPE) realizes this normalization. The FPE is theoretically defined as the average point of expectation embodying every particular points of expectation, thus reducing the *n* particular points of expectation to just one average point. Basically, the fair point of expectation is the arithmetic mean of all the joint probabilities[2] of the (*n*-1)-grams contained in the *n*-gram. The FPE for an *n*-gram is defined in (2).

$$FPE([w_1 p_{12} w_{2\ldots} p_{1i} w_{i\ldots} p_{1n} w_n]) = \frac{1}{n}\left(p([w_{2\ldots} p_{2i} w_{i\ldots} p_{2n} w_n]) + \sum_{i=2}^{n} p\left(\left[\begin{array}{c} \wedge \quad \wedge \\ w_1 p_{12\ldots} p1i\ wi_{\ldots} p_{1n} wn \end{array}\right]\right)\right) \qquad \textbf{(2)}$$

where $p([w_2 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n])$, for i=3,...,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 $p\left(\left[\begin{array}{c} \wedge \quad \wedge \\ w_1 p_{12\ldots} p1i\ wi_{\ldots} p_{1n} wn \end{array}\right]\right)$ 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*. So, the normalized expectation of a generic *n*-gram is defined as being the "fair" conditional probability using the fair point of expectation and is defined in (3)

[1] The sign "+" ("-") is used to represent words on the right (left) of w_1

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

$$NE([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n]) = \frac{p([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n])}{FPE([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n])} \quad (3)$$

where $p([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n])$ is the probability of occurrence of the *n*-gram $[w_1 p_{12} w_2 \ldots p_{1i} w_i \ldots p_{1n} w_n]$ and $FPE([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n])$ is as defined in (2). The reader will be aware of the fact that the Normalized Expectation measure is different than the Dice coefficient introduced by [Smadja1996] although they share the same expression for the case of *2*-grams. And it will become clearer when we will access the results by using the equivalently normalized Dice coefficient.

2.2 The Mutual Expectation Measure

[Daille1995] 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 (i.e. with the same value measuring the possible loss of one word in an *n*-gram) the more frequent *n*-gram is more likely to be a multiword lexical unit. So, the ME between *n* words is defined in (4) based on the NE and the simple frequency

$$ME([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n]) = f([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n]) \times NE([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n]) \quad (4)$$

where $f([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n])$ and $NE([w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n])$ are respectively the frequency of the particular *n*-gram $[w_1 p_{12} \ldots p_{1i} w_i \ldots p_{1n} w_n]$ and its normalized expectation. So, each *n*-gram is associated to its ME value in order to elect the potential MWUs.

3 The LocalMax Algorithm

Most of the approaches proposed for the extraction of multiword lexical units are based on association measure thresholds like in [Church1990], [Daille1995] and [Smadja1996]. This is defined by the underlying concept that there exits 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 association measures tend to favor certain properties of the MWUs, and as a consequence, the coarse grain threshold methodology may reject unjustifiably potential MWUs in the set of all valued *n*-grams. Finally, these thresholds may vary with the type, the size and the language of the document and vary obviously with the association measure. The LocalMax algorithm [Silva1999] proposes a more robust, flexible and fine tuned approach.

Electing MWUs among the sample space of all valued *n*-grams (i.e. each *n*-gram is associated to its cohesiveness value) may be defined as detecting combinations of features that are common to all instances of the concept of MWUs and/or features that must not be found for an object to be an instance of the concept. Taking into account that the only feature we have is the association measure value for each *n*-gram, the LocalMax algorithm elects the multiword units from the set of all the valued *n*-grams based on the two following assumptions. First, the association measures show that the more cohesive a group of words is the higher its score[3] will be. Second, MWUs are highly associated localized groups of words. From these two assumptions, we can deduce that an *n*-gram is a MWU if the degree of cohesiveness between its *n* words is higher or equal than the degree of cohesiveness of any sub-group of *n-1* words contained in the *n*-gram and if it is strictly higher than the degree of cohesiveness of any super-group of *n+1* words containing all the words of the *n*-gram. As a consequence, an *n*-gram is a MWU if its ME value is higher or equal than the ME value of any *(n-1)-gram* contained in the *n*-gram and if it is strictly higher than the ME value of any *(n+1)-gram* containing the *n*-gram. Though, the LocalMax algorithm avoids the ad hoc definition of any association measure threshold overcoming the problems of reliability and portability of the previous proposed methodologies.

4 Evaluation of the Results

We present the results obtained for Italian by applying the LocalMax algorithm and the mutual expectation to an Italian, Portuguese, English and French parallel corpus of political debates taken from the European Parliament debates collection with approximately 300000-words for each language. We then compare the results obtained with four other normalized association measures: the association ratio (N_AR)[4], the dice coefficient (N_DC)[5], the Φ^2 (N_PHI)[6] and the Log-likelihood ratio (N_LOG)[7].

---

[3] The conditional entropy measure is one exception.

[4] The N_AR is the application of the fair point of expectation methodology to the association ratio introduced by [Church1990].

[5] The N_DC is the application of the fair point of expectation methodology to the dice coefficient introduced by [Smadja1996].

[6] The N_PHI is the application of the fair point of expectation methodology to the Pearson's coefficient introduced by [Gale1991].

[7] The N_LOG is the application of the fair point of expectation methodology to the Log-likelihood ratio introduced by [Dunning1993].

We built all the non-contiguous *n*-grams (for *n*=1 to *n*=10) from the parallel corpus and applied to each one its mutual expectation value and finally ran the LocalMax algorithm on this data set. Contiguous multiword lexical units (CMWUs) and non-contiguous rigid multiword lexical units (NCMWUs) have been extracted (See Table 1). In the case of the extracted NCMWUs, 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 MWU the less this unit is meaningful and the more it is likely to be an incorrect multiword lexical unit.

Table 1: Sample extracted contiguous and non-contiguous multiword units sorted by frequency

Contiguous multiword units	**Frequency**
Stati membri	114
cooperazione politica	29
in materia di	21
Premio europeo di letteratura	4
esercitare la professione	3
codice di buena condotta	3
per motivi di	2

Non-contiguous multiword units	**Frequency**
la _____ di	65
di_____ e di	19
densità _____ popolazione	3
proposta di _____ del Consiglio	2
totale di _____ milioni	2
essere _____ in considerazione	2

We measured the precision of the results based on two assumptions. First, multiword lexical units are valid units if they are grammatically appropriate units (by grammatically appropriate units we refer to compound nouns/names, compound verbs and compound prepositions/adverbs/conjunctions). And second, multiword lexical units are valid units if they are relevant structures even though they are not grammatical[8] such as *al fine di _____ la* where the gap stands for any verb in the infinitive form. For the latter case, we used a concordancer to verify whether one elected *n*-gram was a relevant structure or not regarding to its immediate context. In these conditions, the system shows a precision of 88,56% (See Table 2). Although the evaluation of extraction systems is usually performed with precision and recall coefficients, we do not present the "classical" recall rate in this experiment due to the lack of a reference corpus where all MWUs are identified. Instead, we present the extraction rate, a measure of coverage, which is the percentage of well-extracted MWUs (i.e. correct MWUs) in relation with the size of the corpus which was evaluated at 1,81% for the ME (See Table 2).

Then, we applied to the same corpus the LocalMax algorithm with the Normalized association ratio (N_AR), the normalized Φ^2 (N_PHI), the normalized Dice coefficient (N_DC) and the normalized Log-likelihood ratio (N_LOG), and compared the results. The normalized measures are the result of the application of the fair point of expectation methodology, respectively, to the measures of [Church1990], [Gale1991], [Smadja1996] and [Dunning1993]. The experiment shows that the ME measure gives significantly better results than all the other normalized measures in terms of precision and consistency [Dias1999-2]. The N_AR makes rare word groups look more similar than they really are and as a consequence the average frequency of the extracted contiguous multiword units falls to 2.27 raising a weak extraction rate. The Figure 1 shows that almost 90% of the elected MWUs with the N_AR occur only twice in the text. Besides, almost no 2-grams are extracted over-evaluating the average length of the units to 3.25 words (See Figure 2). The N_DC, unlike the N_AR and the N_PHI, has a higher extraction rate. But, most of the extracted MWUs are only two words long being evidenced by an average length of only 2.18 (See Figure 2). The N_DC also shows one of the worst precision rate over-generalizing the concept of MWUs. Moreover, the N_DC tends to elect preferably frequent MWUs as shown in Figure 1. MWUs occurring three times in the corpus represent the highest proportion of the elected MWUs.

The N_PHI shows a more satisfying precision rate than the N_AR, the N_DC and the N_LOG measures but its extraction rate is weak comparing to the N_LOG, the N_DC and the ME. Like the N_DC and the N_LOG, the N_PHI also tends to elect short MWUs with a low average length rate of 2.78 words. The Figure 2 confirms the previous result showing that a great proportion of the extracted MWUs is two-word long, which is not satisfactory. Finally, the N_LOG evidences the worst precision rate of all measures in contrast with its extraction rate that evidences the best result. However, similarly to the N_DC, the N_LOG almost only elects *2*-grams (See

[8] This choice can easily be argued as a precision measure should be calculated in relation with a particular task. For instance, one may calculate the precision of the extracted multiword units for machine translation purposes, for information retrieval purposes or for lexicographic purposes.

Figure 2). Besides, the precision of the elected *n*-grams, for *n* higher than 2, is weak causing the low precision rate result.

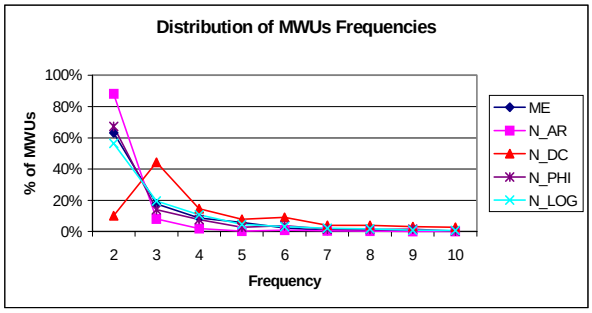
The ME is a much more satisfactory association measure as it shows the best precision and the second highest extraction rate of all the experimented measures. It also elects a more variegate set of multiword lexical units with an acceptable average length rate of 3.17 (See Figure 2).

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 four measures (N_AR, N_DC, N_PHI and N_LOG) elect the multiword lexical unit *selezione _____ personale docente universitario* although the probability that the preposition *del* fills in the gap is one. In fact, the following 5-gram [*selezione +1 del +2 personale +3 docente +4 universitario*] gets unjustifiably a lower value of cohesiveness than the *4*-gram [*selezione +2 personale +3 docente +4 universitario*]. Indeed, the high frequency of the word *del* underestimates the cohesiveness value of the *5*-gram. On the opposite, the ME elects the MWU *selezione del personale docente universitario* as it does not depend on marginal probabilities except for the case of the *2*-grams. So, all the non-contiguous multiword lexical units extracted with the mutual expectation measure define correct units as the gaps correspond to the occurrence of at least two different tokens. The problem shown by the other measures is illustrated by the high rate of extracted non-contiguous multiword lexical units (See Table 2). Identically, the N_AR, the N_DC, the N_PHI and the N_LOG elect the multiword lexical unit *in materia* although the probability that the preposition *di* occurs after *in materia* is very high. And one may expect to extract the well-formed MWU *in materia di*. In fact, at least one of the following *2*-grams [*in +1 materia*], [*in +2 di*] or [*materia +1 di*] gets unjustifiably a higher value of cohesiveness than the *3*-gram [*in +1 materia +2 di*]. The ME elects the MWU *in materia di* raising the precision of the elected MWUs comparing to the other three measures. Finally, the analysis of some particular non-contiguous rigid multiword units enables to partially solve the problem of the extraction of hapaxes (i.e. multiword units with frequency equal to one). Some particular non-contiguous rigid multiword lexical units are generalizations of one particular concept. As a consequence, all the possible instances of the generalized concept are also multiword units independently of their frequencies.

Table 2: Comparative results for Italian between 5 association measures

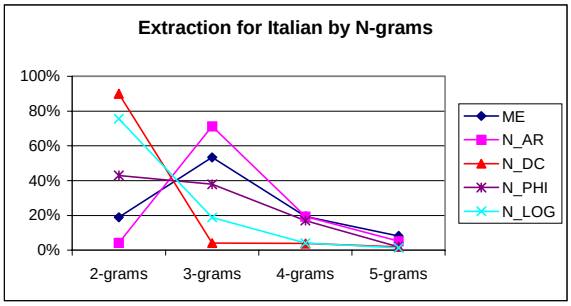
	ME	N_AR	N_DC	N_PHI	N_LOG
% of CMWU	75.49	48.70	61.91	58.50	64.28
% of NCMWU	24.51	51.30	38.09	41.50	35.72
Average frequency of CMWU	5.44	2.27	6.68	6.38	4.47
Average frequency of NCMWU	4.10	2.16	6.25	2.85	3.84
Average lenght of MWU	3.17	3.25	2.18	2.78	2.31
% Precision	88.56	64.23	56.24	70.50	50.80
% Extraction rate	1.81	0.91	1.78	0.94	3.46

Figure 1: The distribution of the extracted MWUs frequencies.



For example, the multiword unit *proposta di ____ del Consiglio* can be represented as the following model $\exists x$*(proposta di x del Consiglio)* where possible occurrences of *x* such as *direttiva* or *regolamento* specify the overall concept. As the gap defines a cluster of nouns, we may argue that each instance filling in the gap defines a new multiword lexical unit independently of its frequency. Therefore, *proposta di direttiva del Consiglio* and *proposta di regolamento del Consiglio* are MWUs even if they occur only once in the corpus. This technique can not obviously be applied to all non-contiguous rigid multiword lexical units but we believe that the problems raised can be easily overcome using linguistic information, such as part of speech tags, in order to extract the suitable non-contiguous rigid multiword units.

Figure 2: The Distribution of the extracted MWUs lengths



6 Conclusion

We proposed in this paper a language independent statistically-based system that automatically extracts contiguous and non-contiguous rigid multiword lexical units from unrestricted text corpora. The experiments realized on a corpus of the legal domain evaluate the precision of the system at 88,56%. We compared the mutual expectation with four other association measures and the comparative results show that the mutual expectation gives high precision and extraction rates, overcomes the problem of highly frequent words raised by the four other measures and tends to elect longer multiword units. Finally, the system ensures total portability as it is applicable to various languages as it uses plain text corpora and requires only the general information appearing in it. Finally, the analysis of the results pointed at a partial solution to the problem of the election of hapaxes by evidencing patterns of multiword units.

We experienced our system on Portuguese, French and English corpora and obtained similar results in terms of precision rate, extraction rate, length and frequency distributions than for Italian [Dias1999-3].

We are hardly convinced that the success of applications in the areas of Natural Language Processing, Information Retrieval and Information Extraction will rely on the preprocessing of corpora in order to benefit from their intrinsic information. The extraction of implicit knowledge (knowledge of the language) such as sub-categorization frames, pp-attachment and MWUs will enable more precise text processing and as a consequence will lead to an adequate normalization of texts in order to extract more explicit information (knowledge of the world).

References

Church K.et al. (1990), "Word Association Norms, Mutual Information, and Lexicography", *Computational Linguistics*, Vol. 16 (1), pp 23-29

Dagan I. (1994), "Termight : Identifying and Translating Technical Terminology", *4th Conference on Applied Natural Language Processing, ACL Proceedings*

Daille B. (1995), "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology", *The balancing act combining symbolic and statistical approaches to language* (MIT Press)

Dias G., Guilloré S. et Lopes J.G. (1999), "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora", In proceedings of TALN'99, 12-17 July, Cargèse, France

Dias G., Guilloré S. et Lopes J.G. (1999), "Multiword Lexical Unit Extraction", In proceedings of the International Conference on Machine Translation and Computer Language Information Processing, 26-28 June, Beijing, China

Dias G., Guilloré S. et Lopes J.G. (1999), "Multilingual Aspects of Multiword Lexical Units", In proceedings of the Workshop on Language Technologies – Multilingual Aspects, 8-11 July, Ljubljana, Slovenia

Dunning T. (1993), "Accurate Methods for the Statistics of Surprise and Coincidence", *ACL*, Vol. 19-1

Gale W. (1991), "Concordances for Parallel Texts", *Proceedings of Seventh Annual Conference of the UW Center for the New OED and Text Research*, Using Corpora, Oxford

Smadja F. (1993), "Retrieving Collocations From Text : XTRACT", *Computational Linguistics*, Vol. 19 (1), pp 143-177

Smadja F. (1996), "Translating Collocations for Bilingual Lexicons: A Statistical Approach", *Association for Computational Linguistics*, Vol 22-1

Silva J. et Lopes J. G. (1999), "A local Maxima Method and a Fair Dispersion Normalization for Extracting multiword units", In Proceedings of MOL'6, 23-25 July, Orlando, USA