Extracting Textual Associations in Part-of-Speech Tagged Corpora

Gaël Dias[123], Sylvie Guilloré[2] &
José Gabriel Pereira Lopes[3]

[1] Universidade da Beira Interior
Departamento de Matemática-Informática
ddg@noe.ubi.pt
[2] Université d'Orléans
Laboratoire d'Informatique Fondamentale d'Orléans
sylvie.guillore@lifo.univ-orleans.fr
[3] Universidade Nova de Lisboa
Departamento de Informática
gpl@di.fct.unl.pt

Abstract

Identifying textual associations from text corpora is a useful pre-processing step for many applications in Natural Language Processing. In this paper, we will present an innovative system that extracts relevant sequences of characters, words and part-of-speech tags from corpora. We will show that the combination of a new association measure (Mutual Expectation) with a new acquisition process (LocalMaxs) proposes an integrated solution to the problems of enticement techniques and global thresholds highlighted by previous researches.

Introduction

Identifying textual associations from corpora is a useful pre-processing step for many applications in Natural Language Processing. In the context of word associations, Mel'čuk (1988) argues that multiword lexical units (sequences of words that co-occur more often than expected by chance) are often opaque in the comprehension phase and cause hesitations in the production process. For instance, "Bill of Rights", "swimming pool", "as well as", "in order to", "to comply with" and "to put forward" are multiword lexical units. Consequently, their identification within the process of text normalisation is a crucial issue for the specific tasks of machine translation, information extraction and information retrieval. Textual associations are not restricted to word associations. Indeed, Argamon-Engelson *et al.* (1999) assess that the identification of local patterns of syntactical sequences is essential for various application areas including word sense disambiguation, bilingual alignment and text summarisation. The principle approach for the detection of syntactical patterns is the task of shallow parsing which consists, according to Abney's definition (1991), in a chunker that offers potential part-of-speech tag sequences to an attacher that solves attachment ambiguities and selects the final chunks. It is clear that systems that may identify meaningful part-of-speech tags associations such as "AT JJ NN"[1], "JJ NP CC JJ NP", "NP $ JJ NN", "NP CO NP CO NP CC NP" and "HV RB BEN" would greatly benefit the shallow parsing task. Finally, in the context of character associations, the decomposition of words into morphemes has proved to lead to improved results in different areas including text indexing, bilingual alignment and information extraction. Indeed, Grabar and Zweigenbaum (1999) confirm that a great deal of words in European Languages share common Greek and Latin morphemes that allow the generalisation of concepts. As a consequence, it is convenient to define morphological segments as meaningful sequences of characters that should be automatically extracted from corpora. For instance, the following words belong to the same morphological family as they share the same stem: **ocean***ography*, **ocean***arium*, **ocean***ic* and **ocean**. In order to identify and extract meaningful sequences of words, part-of-speech tags and characters, we present a statistically-based architecture called SENTA (Software for the Extraction of N-ary Textual Associations) that retrieves, from naturally occurring text, relevant contiguous and non-contiguous textual associations. For that purpose, we combine a new association measure called the Mutual

[1] We use the following part-of-speech tag set: AT = determinant, JJ = adjective, RB = adverb, NN = singular noun, NP = personal noun, $=possessive markup ('s), CO = comma, CC = coordination conjunction, HV=auxiliary have, BEN = past participle of verb be.

Expectation with a new acquisition process called the LocalMaxs algorithm. On one hand, the Mutual Expectation, based on the concept of Normalised Expectation, evaluates the degree of cohesiveness that links together all the textual units contained in an n-gram ($\forall n, n \geq 2$). On the other hand, the LocalMaxs algorithm retrieves the potential associations from the set of all the valued n-grams by evidencing local maxima of association measure values. This combination proposes an innovative integrated solution to the problems of enticement techniques and global thresholds highlighted by previous researches. As an illustration, we access the results obtained by running SENTA on three different data sets built from the tagged Brown corpus i.e. a corpus of words[2], a corpus of part-of-speech tags and a corpus of characters (Figure (2))[3]. In the context of word associations, the results point at the extraction of compound nouns and verbs, and various types of locutions. Analogously, the system retrieves contiguous and non-contiguous noun-phrase (NP), verb-phrase (VP), subject-verb (SV) and verb-object (VO) chunks from the corpus of part-of-speech tags. Finally, in the context of character associations, bound and free morphemes are identified. In the following section, we will present the first stage of the system that consists in the transformation of the input text corpus into contiguous and non-contiguous n-grams of textual units. In section 2 and 3, we will respectively introduce the Mutual Expectation measure and the LocalMaxs algorithm. Finally, in section 4, we will detail and discuss the experiments realised over the Brown corpus.

1 Data Preparation

Van den Bosch (1998) assesses that most relations between textual units (TUs) occur within a local context (span) of at most six other textual units. As a consequence, a textual association can be defined in terms of structure as a specific n-gram calculated in an immediate span of three TUs to the left hand side and three TUs to the right hand side of a focus TU, as illustrated in Figure (1). By definition, an n-gram is a vector of n TUs where each TU is indexed by the signed distance that separates it from its associated focus TU. Consequently, an n-gram can be contiguous or non-contiguous whether the TUs in the n-gram represent or not a continuous sequence in the corpus.

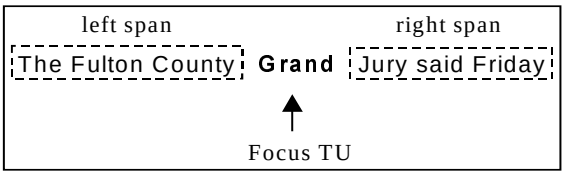

Figure 1: Local context of a focus TU

By convention, the focus TU is always the first element of the vector and its signed distance is equivalent to zero. We represent an n-gram by a vector $[p_{11}\, u_1\, p_{12}\, u_2 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n]$ where p_{11} is equal to zero and p_{1i} (for i=2 to n) denotes the signed distance that separates the TU, u_i, from the focus TU, u_1. For instance, let's consider a focus TU for each one of the three corpora built from the original Brown corpus as illustrated in Figure (2). We'll respectively take as focus TUs, the word "Fulton", the part-of-speech tag "/NP"[4] and the character "F"[5]. Three possible contiguous and non-contiguous 3-grams are illustrated in the first three rows of Table (1).

p_{11}	u_1	p_{12}	u_2	p_{13}	u_3
0	Fulton	-1	the	+1	County
0	/AT	+1	/NP	+2	/NP
0	F	+1	u	+3	t

Table 1: Sample 3-grams.

As notation is concerned, we may characterise an n-gram 1) by the sequence of its constituents as they appear in the corpus or 2) by explicity mentionning the signed distances associated to each TU.

Notation 1
the Fulton County
/AT /NP /NP
F u _____ t

Table 2: First notation.

[2] We refer to a word as a sequence of characters surrounded by empty spaces but containing no internal space.

[3] For presentation purposes, the space character in the character corpus is identified by the "*" character.

[4] This is the first /NP tag in the tag corpus.

[5] This is the first "F" in the character corpus.

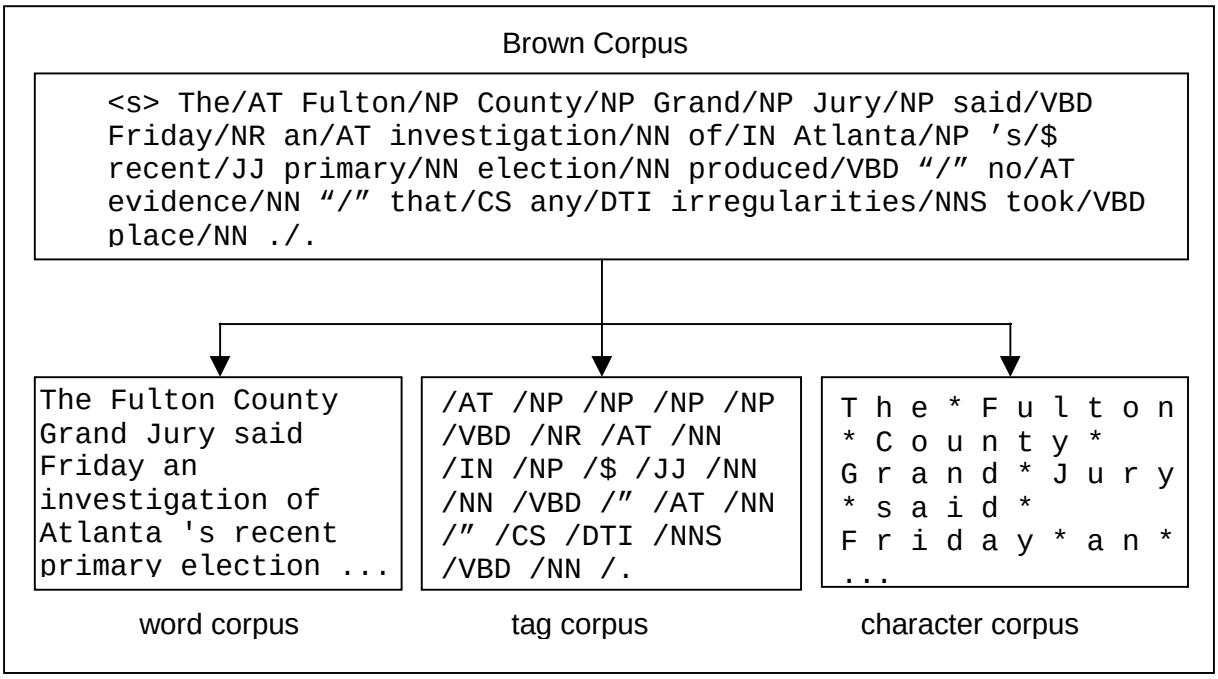

Figure 2: Dividing the Brown Corpus into three different data sets.

In the first case, each interruption of a non-contiguous n-gram is identified by a gap ("_____") that represents the set of all the occurrences that fulfil the free space in the text corpus. In the second case, the distance of the focus TU is ommited. Table (2) and Table (3) respectively illustrate both notations for the sample 3-grams presented in Table (1).

Notation 2
[Fulton -1 the +1 County]
[/AT +1 /NP +2 /NP]
[F +1 u +3 t]

Table 3: Second notation.

As computation is concerned, we developed an algorithm that sequentially processes each TU of the input text corpus[6]. Each TU is successively a focus TU and all its associated contiguous and non-contiguous n-grams are calculated avoiding duplicates. Finally, each n-gram is associated to its frequency in order to apply the Mutual Expectation measure that evaluates its degree of cohesiveness.

[6] The corpus is not pruned with stop-lists, thus all the information in the input text is taken into account.

2 Mutual Expectation

In order to evaluate the degree of cohesiveness existing between TUs, various mathematical models have been proposed in the literature. Church (1990), Gale (1991), Dunning (1993), Smadja (1993; 1996) and Shimohata (1997) are some references. However, most of these models only evaluate the degree of cohesiveness between two TUs and do not generalise for the case of n individual TUs ($\forall n,\ n \geq 2$). As a consequence, these association measures only allow the acquisition of binary associations and enticement techniques[7] have to be applied to acquire associations with more than two TUs. Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable 2-grams for the initiation of the iterative process.

On the other hand, these models have shown to be over-sensitive to frequent TUs. In particular,

[7] As a first step, relevant 2-grams are retrieved from the input corpus. Then, n-ary associations may be identified by either 1) gathering overlapping 2-grams or 2) by marking the extracted 2-grams as single words in the text and re-running the system to search for new 2-grams (the process ends when no more 2-grams are identified).

in the context of word associations, this has lead researchers to regard function words like determinants or prepositions as meaningless to the sake of the statistical evaluation process. For instance, Daille (1995) tested various association measures on plain word pairs only. In order to overcome both problems, we present a new association measure called the Mutual Expectation, introduced by Dias *et al.* (1999a), that evaluates the degree of cohesiveness that links together all the TUs contained in an n-gram ($\forall n,\ n \geq 2$) based on the concept of Normalised Expectation.

2.1 Normalised Expectation

The basic idea of the Normalised Expectation (NE) is to evaluate the cost of loosing one TU in a given n-gram. Thus, the less an n-gram would accept the loss of one of its components, the higher its NE value should be. Consequently, the NE for a given n-gram can be defined as the **average expectation** of occurring one of its constituents in a given position knowing the occurrence of the other (n-1) ones[8]. Indeed, the more the (n-1) TUs in an n-gram expect for the occurrence of a specific TU, the more the degree of cohesiveness between the n constituents will be high. So, for instance, the NE of the 3-gram [the +1 Fulton +2 County] would be the average expectation embodying all the expectations presented in Table (4).

Expectation to occur	Knowing the gapped 3-gram
the	[_____ +1 Fulton +2 County]
Fulton	[the +1 _____ +2 County]
County	[the +1 Fulton +2 _____]

Table 4: Sample expectations.

But, each raw of Table (4) corresponds exactly to one conditional probability that evaluates the specific expectation of occurring one TU in a given position knowing the (n-1) other ones.
It is clear that the NE is based on the conditional probability (Equation (1)) that measures the expectation of occurring the event X=x knowing the conditional event Y=y.

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

Equation 1: Conditional Probability.

However, this definition does not accommodate the n-gram length factor. Indeed, an n-gram is naturally associated to n possible conditional probabilities. Thus, a normalisation process is necessary.
At this stage, we introduce the concept of the Fair Point of Expectation (FPE) that proposes an elegant solution for the process of normalisation. As the numerators remain unchanged from one specific probability to another, the FPE defines one **average conditional event** that embodies all the specific conditional events specified by each conditional probability. Theoretically, the FPE for a given n-gram is defined as the arithmetic mean of the n joint probabilities[9] of the (n-1)-grams contained in the n-gram. It is defined in Equation (2).

$$FPE(\lfloor p_{11}u_1 p_{12}u_2 \ldots p_{1i}u_i \ldots p_{1n}u_n \rfloor) = \frac{1}{n}\left(\sum_{i1=1}^{2} \sum_{i2=i1+1}^{3} \cdots \sum_{i(n-1)=i(n-2)+1}^{n} p\left(\begin{bmatrix} p_{(i1)(i1)}u_{(i1)}p_{(i1)(i2)}u_{(i2)}\ldots \\ \ldots p_{(i1)(i(n-1))}u_{(i(n-1))} \end{bmatrix} \right) \right)$$

Equation 2: Fair Point of Expectation.

So, the normalisation of the conditional probability is realised by the introduction of the FPE into the general definition of the conditional probability. The resulting measure is called the NE and it is proposed as a "fair" conditional probability. It is defined in Equation (3).

$$NE(\lfloor p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n \rfloor) = \frac{p(\lfloor p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n \rfloor)}{FPE(\lfloor p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n \rfloor)}$$

Equation 3: Normalised Expectation.

[8] The (n-1) other TUs are also constrained by their positions.

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

2.2 Mutual Expectation

Many applied works in Natural Language Processing have shown that frequency is one of the most relevant statistics to identify relevant textual associations. For instance, in the context of word associations, Gross (1996) corroborates Daille (1995) and Justeson (1993)'s opinions that the comprehension of a multiword lexical unit is an iterative process being necessary that a unit be pronounced more than one time to make its comprehension possible. We hardly believe that this phenomenon can be enlarged to part-of-speech tag and character associations. From this assumption, we deduce that between two n-grams sharing the same NE, the most frequent n-gram is more likely to be a relevant textual association. So, the Mutual Expectation of an n-gram is the product between its NE and its relative frequency as defined in Equation (4).

$$ME([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n]) = p([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n]) \times NE([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n])$$

Equation 4: Mutual Expectation.

Comparing to previously proposed mathematical models, the ME allows evaluating the degree of cohesiveness that links together all the textual units contained in an n-gram (i.e. $\forall n, n \geq 2$) as it accommodates the n-gram length factor. So, it is possible to classify each n-gram by its degree of pertinence. In the following section, we present the LocalMaxs algorithm that retrieves the potential textual associations from the set of all the valued n-grams by evidencing local maxima of association measure values.

3 LocalMaxs Algorithm

Electing textual associations among the sample space of all the valued n-grams may be defined as detecting combinations of features that are common to all the instances of the concept of textual association. In the case of statistical methodologies, frequency and association measure are the only two features available to the system. As a consequence, most of the approaches have based their selection process on the definition of global thresholds of frequency and/or association measure as in Church (1990), Smadja (1993), Daille (1995), Shimohata (1997) and Feldman (1998). This is defined by the underlying concept that there exit limit values of frequency and/or association measure that allow to decide whether an n-gram is a pertinent textual association or not. However, these thresholds are prone to error as they mainly depend on experimentation. Furthermore, they highlight evident constraints of flexibility, as they need to be re-tuned when the type, the size, the domain and the language of the input corpus change[10]. In order to deal with both problems, Silva *et al.* (1999b) has introduced the LocalMaxs algorithm that concentrates the acquisition process on the identification of local maxima of association measure values. So, an n-gram is a textual association if its association measure value is higher or equal than the association measure values of all its sub-groups of (n-1) TUs and if it is strictly higher than the association measure values of all its super-groups of (n+1) TUs. The LocalMaxs is defined in Figure (3) being *assoc* any association measure[11], W an n-gram, Ω_{n-1} the set of all the (n-1)-grams contained in W, Ω_{n+1} the set of all the (n+1)-grams containing W and *sizeof* a function that returns the number of TUs in an n-gram.

$\forall x \in \Omega_{n-1}$, $\forall y \in \Omega_{n+1}$ W is a textual association

IF

$(sizeof(W)=2 \wedge assoc(W) > assoc(y)) \vee (sizeof(W) \neq 2 \wedge assoc(W) \geq assoc(x) \wedge assoc(W) > assoc(y))$

Figure 3: The LocalMaxs. algorithm.

The LocalMaxs highlights two interesting properties. On one hand, it allows the testing of various association measures. In particular, Dias *et al.* (2000) shows that the ME evidences improved results comparing to the Association Ratio introduced by Church (1990), the Dice coefficient proposed by Smadja (1996), Gale (1990)'s ϕ^2 coefficient and Dunning (1993)'s

[10] They obviously vary with the association measure.

[11] The association measure must give higher scores to more cohesive n-grams. For instance, the conditional entropy could not be used with the LocalMaxs.

Log-Likelihood Ratio[12]. On the other hand, the algorithm allows the extraction of textual associations obtained by composition. As it retrieves pertinent textual units by analysing their immediate context, the LocalMaxs may identify textual associations that are composed of one or more other textual associations. This will be discussed in the following section by illustrating the results obtained by combining the LocalMaxs with the ME over the three data sets obtained from the Brown corpus.

4 Results and Discussion

SENTA has been applied over three data sets built from the part-of-speech tagged version of the Brown corpus[13]. In the context of word associations, the results point at the extraction of compound nouns and verbs, and various types of multiword locutions, as illustrated in Table (5).

Multiword Lexical Units	
United States	of course
atom of calcium	later on
Terrier Club of America	in conjunction with
to be able to	can ____ be made with
to compete with	to allow ____ ____ to

Table 5: Multiword Lexical Units.

Analogously, the system retrieves NP, VP, SV and VO chunks from the corpus of part-of-speech tags as evidenced in Table (6).

Chunks	
AT NN	TO BE JJ CC JJ
NP $ NN	JJ CC JJ JJ NP
NP $ JJ NN	HV RB VBN AT JJ NN
AT JJ CC AT JJ	AT JJ NN HV BE VBG
NP $ NN CC NN	VBG CO VBG CO CC VBG

Table 6: Chunks.

The results also highlight the extraction of number-coherent part-of-speech tag associations. Thus, the tag association, "EX BED ____ NNS" has successfully been extracted. Indeed, being EX the tag for the word "*there*", BED for the verb "*were*" and NNS for any plural noun, the number noun-verb correspondence is recognised.

[12] Cramer and Pearson's coefficients have also been tested (Bhattacharyya and Johnson (1977)).

[13] http://morph.ldc.upenn.edu/ldc/online/

In the context of character associations, prefixes, suffixes and stems have been identified, as illustrated in Table (7).

Morphemes	
*atom	room*
*inter	human
*judg	view
ful*	join
ism*	cogni

Table 7: Morphemes.

But, the results also evidence the extraction of allomorphs. Basically, an allomorph can be defined as an alternative manifestation of a morpheme. As a consequence, the following character association, "*** b e g ____ n**", is an allomorph as it corresponds to the change of the stem vowel of the verb to begin, e.g. **begin**, **began**, **begun**. The same occurs with many other verbs e.g. to write, to swim. Finally, we provide some interesting quantitative results about of the extracted n-grams. We compare the frequency distributions per n-gram for each task being tackled i.e. word, part-of-speech tag and character associations.

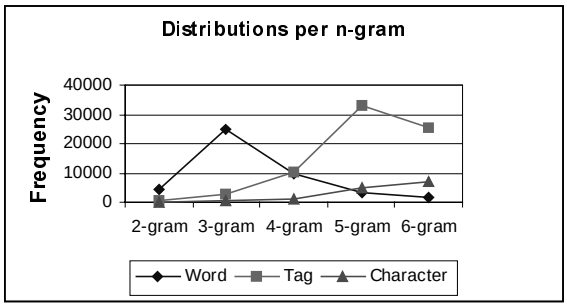

Figure 4: Distribution per n-gram

The different shapes of the lines show that for each case, different patterns are being identified. While word associations are maximum for the case of 3-grams, tag associations are at their maximum for the case of the 5-grams and for character associations, the maximum frequency is shown for the 6-grams. In the context of word associations, the results obtained are similar in terms of distribution to previous works reported in Daille (1995) and Justeson (1993) that confirm that the greatest part of multiword lexical units contain between two and four words. In the context of part-of-speech tag associations, the results show that a very limited

number of phrases are "simple" in the sense that they embody less than four part-of-speech tags. Indeed, complex phrases embodying co-ordinations and relative clauses are recursively used. Finally, in the context of character associations, the results are not surprising. Indeed, in contrast to the number of words, the number of morphemes is finite. But, as we carry on adding characters to one another, words are being formed. Thus, for the case of 6-grams, many words are evidenced e.g. "*** e a c h ***", "*** u s e d ***" and "*** m a k e ***".

Conclusion

In this article, we proposed an innovative methodology for the extraction of textual associations from unrestricted texts. We introduced the Mutual Expectation measure and the LocalMaxs algorithm that allow identifying relevant contiguous and non-contiguous textual associations without defining global thresholds or using enticement techniques. Nevertheless, efforts must be made in order to propose organised sets of data instead of unrelated textual associations. We are actually working in that sense.

References

Abney S.P. (1991). *Parsing by chunks*. In R.C. Berwick, S.P. Abney and C. Tenny, ed., Principle-Based Parsing: Computation and Psycholinguistics, Kluwer, Dordrecht, pp. 257-278.

Argamon-Engelson S., Dagan I. and Krymolowski (1999). A Memory-Based Approach to Learning Shallow Natural Language Patterns. In cmp-lg/980611 v3, April.

Bhattacharyya G. and Johnson R. (1977). *Statistical Concepts and Methods*, New York, Wiley & Sons.

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

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

Dias, G., Guilloré, S. and Lopes, J.G.P. (2000). *Normalisation of Association Measures for Multiword Lexical Unit Extraction*. In "International Conference on Artificial and Computational Intelligence for Decision, Control and Automation in Engineering and Industrial Applications", Monastir, Tunisia.

Dias, G., Guilloré, S. and Lopes, J.G.P. (1999a). *Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora.* In "Traitement Automatique des Langues Naturelles", Institut d'Etudes Scientifiques, Cargèse, France.

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

Feldman R. (1998). *Text Mining at the Term Level*. In "Principles and Practice of Knowledge Discovery in Databases", Lecture Notes AI 1050, Springer Verlag.

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

Grabar N. and Zweigenbaum P. (1999). *Acquisition Automatique de connaissances morphologiques sur le vocabulaire médical.* In "Traitement Automatique des Langues Naturelles", Institut d'Etudes Scientifiques, Cargèse, France.

Gross, G. (1996). Les expressions figées en français. Ophrys, Paris, France.

Justeson J. (1993). *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text*. IBM Research Report, RC 18906 (82591) 5/18/93.

Mel'čuk I. (1988). *Paraphrase et lexique dans la théorie linguistique sens-texte* In "Lexique", vol 6.

Shimohata S. (1997). *Retrieving Collocations by Co-occurrences and Word Order Constraints*. In "ACL-EACL", pp. 476-481.

Silva, J., Dias, G., Guilloré, S. and Lopes J.G.P. (1999b). *Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units*. In "9th Portuguese Conference in Artificial Intelligence", Lecture Notes, Springer-Verlag, Évora, Portugal.

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

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

Van den Bosch A. (1998). *Instance Families in Memory-Based Language Learning*. In Van Eynde F., Schuurman I. and Schelkens N, ed., Computational linguistics in the Netherlands 1998, Rodopi, Amsterdam, pp 3-17.