Benefiting from Multidomain Corpora to Extract Terminologically Relevant Multiword Lexical Units

Gaël DIAS, Sylvie GUILLORÉ, José Gabriel PEREIRA LOPES,
Covilhã and Caparica, Portugal, Orléans, France

Abstract
The acquisition of terminology for particular domains has long been a significant problem in Natural Language Processing requiring a great deal of manual effort. In order to provide terminologists with powerful tools for the creation, the maintenance and the upgrade of terminological data collections, we present the SENTA software that retrieves, from naturally occurring text, terminologically relevant multiword lexical units. SENTA is a statistical system that conjugates a new association measure based on the concept of normalised expectation, the Mutual Expectation, with a new acquisition process based on an algorithm of local maxima, the LocalMaxs. The results obtained by applying SENTA to the IJS-ELAN Slovene-English parallel corpus stress the extraction of a great proportion of terms, with 74% precision on average. Moreover, by conducting further experiments, we show that the average precision rate can be drastically improved up to 82% benefiting from the multidomain structure of the IJS-ELAN Slovene-English parallel corpus.

1 Introduction

The acquisition of terminology for particular domains has long been a significant problem in Lexicography requiring a great deal of manual effort. In order to provide terminologists with powerful tools for the creation, the maintenance and the upgrade of terminological data collections, three main strategies have emerged from the research community. Linguistic [Dagan 1994] [Bourigault 1996], statistical [Church/Hanks 1990] [Dunning 1993] [Smadja 1993] [Shimohata 1997] and hybrid linguistic-statistical systems [Justeson/Katz 1993] [Daille 1995] [Heid 1999] have been proposed in order to extract terminologically relevant multiword lexical units from text corpora.

However, linguistic approaches, combined or not with statistical measures, present two major drawbacks. First, by reducing the searching space to groups of words that correspond uniquely to particular syntactic patterns (mainly regular noun phrases: Adj+Noun, Noun+Noun etc ...), such systems do not deal with a great proportion of terms. Consequently, terms like "*vitamin C*", "*supply and demand*", "*come into effect*", "*Getting Started*" or "*half a teaspoon*" are unlikely to be extracted. Second, the definition of syntactical constraints requires adjustments from one language to another and from one domain to another [Habert/Jacquemin 1998] thus preventing such systems to be widely used. On the other hand, purely statistical methods are more flexible being domain and language independent as they use plain text corpora and only require the information appearing in texts. However, the systems presented so far in the literature also emphasise two major drawbacks. First, by relying on the *ad hoc* establishment of global association measure thresholds such systems are prone to error. Second, most of them only allow the acquisition of binary associations thus requiring bootstrapping techniques[1] to acquire multiword lexical units with more than two words. Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable 2-grams (i.e. groups of two words) for the initiation of the iterative process.

In order to overcome the drawbacks highlighted by previous statistical methods, we present the SENTA software (Software for the Extraction of N-ary Textual Associations) that retrieves, from naturally occurring text, terminologically relevant contiguous and non-contiguous multiword lexical units. SENTA avoids the definition of global thresholds and does not require bootstrapping techniques as it conjugates a new association measure based on the concept of normalised expectation, the Mutual Expectation [Dias et al. 1999], with a new acquisition process based on an algorithm of local maxima, the LocalMaxs [Silva et al. 1999]. The results obtained by applying SENTA to the IJS-ELAN Slovene-English parallel corpus stress the extraction of a great proportion of terms, with 74% precision on average. By leading further experiments, we show that the average precision rate can be drastically improved up to 82% benefiting from the multidomain structure of the IJS-ELAN Slovene-English parallel corpus.

2 Architecture of SENTA

SENTA takes as input a text corpus that is neither lemmatised nor morpho-syntactically tagged nor pruned with lists of stop-words. This decision may be controversial but it is based on the idea that the general information appearing in texts should be enough to extract multiword lexical units (i.e. recurrent sequences of words that co-occur more often than expected by chance). Indeed, according to [Justeson/Katz 1993], the more a sequence of words is fixed (i.e. the less it accepts morphological and syntactical transformations), the more likely it is a multiword lexical unit. Based on this assumption, we believe that multiword lexical units are sufficiently fixed and recurrent sequences of words to propose that they should be retrieved without the introduction of further linguistic information. Furthermore, enhancing text corpora with linguistic information implies the introduction of constraints that are not previously contained in texts (i.e. definition of part of speech tag set, definition of the lemmatisation process, error rate due to deficient morpho-syntactic tagging etc...). As a consequence, we opted not to modify the input text and work on all the information contained inside the corpus. However, we are also aware that this claim may not stand so strongly for morphologically rich languages (like German or Slovene) for which the absolute minimum of linguistic pre-processing is lemmatisation.

The global architecture of SENTA is designed around four sequential steps. First, SENTA performs the transformation of the input text into a set of *n*-gram databases (an *n*-gram is a vector of *n* words indexed by their positions). Indeed, a great deal of applied works in lexicography evidence that most of the lexical relations associate words separated by at most five other words [Sinclair 1974]. So, being a multiword lexical unit a specific lexical relation, it can be defined in terms of structure as a specific contiguous or non-contiguous[2] *n*-gram in a six words wide window (i.e. three words to the left of a word under consideration and three on its right hand side). One non-contiguous 3-gram and one contiguous 3-gram are respectively shown in the first two rows of Table (1), taking as current input the following sentence and *General* (i.e. *w1*) the word under study: [3]

Linux Installation is covered by the GNU General Public License.

w_1	*Position*$_{12}$	w_2	*Position*$_{13}$	w_3
General	-2	*the*	+2	*License*
General	+1	*Public*	+2	*License*

Table 1: Two 3-grams containing *General*

As notation is concerned the non-contiguous 3-gram presented in the first row of Table(1) may be characterized by one of the following expressions where a gap (i.e.“_____”) embodies the set of all the occurrences in the corpus that fulfil the free space:

$$\text{the \_\_\_\_\_ General \_\_\_\_\_ License} \tag{1}$$
$$[\text{General -2 the +2 License}] \tag{2}$$

Similarly, the contiguous 3-gram of the second row may be characterised by one of the following equivalent expressions.

$$\text{General Public License} \tag{3}$$
$$[\text{General +1 Public +2 License}] \tag{4}$$

Generically, we will denote an *n*-gram as the following array [$w_1\ p_{12}\ w_2\ p_{13}\ w_3 \dots p_{1n}\ w_n$] where p_{1i} denotes the signed distance that separates word w_i from word w_1 for $i = 2, \dots, n$.

Then, SENTA respectively calculates the frequency and the Mutual Expectation of each unique *n*-gram. Finally, in the fourth and final step, SENTA applies the LocalMaxs algorithm in order to elect the multiword lexical unit candidates from the set of all valued *n*-grams. In section three and four, we rigorously define the Mutual Expectation and the LocalMaxs algorithm.

3 The Mutual Expectation Measure

In order to evaluate the degree of rigidity existing between words contained in an *n*-gram, various mathematical models have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness within 2-grams and do not generalise for the cause of *n* individual words [Church/Hanks 1990] [Gale 1991], [Dunning 1993] [Smadja 1993] [Smadja 1996] [Shimohata 1997]. As a consequence, these mathematical models only allow the acquisition of binary associations and bootstrapping techniques have to be applied to acquire associations with more than two words. 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. Moreover, their lack of generalisation has lead researchers to test association measures on plain word pairs (i.e. function words like determinants and prepositions are considered as stop-words) in order to evaluate the cohesiveness between words like in [Daille 1995]. So, we introduce a new association measure, the Mutual Expectation (ME) [Dias et al. 1999] , based on the concept of normalised expectation (NE) that evaluates the degree between words contained in an *n*-gram[4].

3.1 Normalised Expectation

By definition, multiword lexical units are recurrent groups of words that co-occur more often than expected by chance. Based on this assumption, we define the NE of an *n*-gram as the average expectation of occurring one word in a given position knowing the occurence 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 loss of one word in an *n*-gram. So the more cohesive a group of words is, that is the less it accepts the loss of one of its componennts, the higher its Normalised Expectation will be.

Expectation to occur the word	Knowing the gapped 3-grams
General	[_____ +1 *Public* +2 *License*]
Public	[*General* +1 _____ +2 *License*]
License	[*General* +1 *Public* +2 _____]

Table 2: Expectations and Normalised Expectation

For example, the average expectation of the 3-gram [*General* +1 *Public* +2 License] must take into account all the expectations presented in Table (2) that correspond to the loss of one word of the 3-gram at a time. Thus the average expectation of the 3-gram must take into account the expectation of occurring the word "*License*" after "*General Public*", but also the expectation of "*Public*" linking together "*General*" and "*License*" and finally, the expectation of the occurrence of an event $X = x$ knowing that an event $Y = y$ stands as illustrated in Equation (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.

$$p(X = x|Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)} \tag{1}$$

Equation 1: Conditional probability

As each word of the text corpus can be mapped to a discrete random variable in a given probability space[5], the definition of the conditional probability can be applied in order to measure the expectation of the occurence of one word in a given position knowing the occurence of the other n-1 words also constrained by their positions. However this definition does not accommodate the *n*-gram length factor. For example, Table(2) clearly points at three possible conditional probabilities for a 3-gram. Naturally, an *n*-gram is associated to n possible conditional probabilities. It is clear that the conditional probability definition needs to be normalised in order to take into account all the conditional probabilities involved by an *n*-gram. In order to explain this process, let's consider the following *n*-gram [w_1 p_{12} $w_2 \ldots p_{1i}$ $w_i \ldots p_{1n}$ w_n]. The extraction of one word at a time from the generic *n*-gram gives rise to the occurence of any of the n events shown in table (3) where the underline (i.e. "_____") denotes the missing word from the *n*-gram.

$(n-1)$-gram	missing word
$[\ \underline{\qquad}\ w_2\ p_{23}\ w_3 \ldots p_{2i}\ w_i \ldots p_{2n}\ w_n\ ]$	w_1
$[\ w_1\ \underline{\qquad}\ p_{13}\ w_3 \ldots p_{1i}\ w_i \ldots p_{1n}\ w_n\ ]$	w_2
$\ldots$	$\ldots$
$[\ w_1 \ldots p_{1(i-1)}\ w_{(i-1)}\ \underline{\qquad}\ p_{1(i+1)}\ w_{(i+1)} \ldots p_{1n}\ w_n\ ]$	w_i
$\ldots$	$\ldots$
$[\ w_1 \ldots p_{1i}\ w_i \ldots p_{1(n-1)}\ w_{(n-1)}\ \underline{\qquad}\ ]$	w_n

Table3: $(n-1)$-grams and missing words

So, each event may be associated with a respective conditional probability that evaluates the expectation of the missing word to occur knowing its corresponding $(n-1)$-gram. The n conditional probabilities are introduced in Equation (2) that evaluates the cost of the loss of all the other words of the n-gram:

$$p(w_1|[w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n]) = \frac{p([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n])}{p([w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n])} \quad i = 1, \ldots, n \tag{2}$$

Equation 2: Conditional Probability for the first word of the n-gram

$$p(w_1|[w_1 \ldots p_{1(i-1)}w_{(i-1)}p_{1(i+1)}w_{(i+1)} \cdots p_{1n}w_n]) = \frac{p([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n])}{p([w_1 \cdots p_{1(i-1)}w_{(i-1)}p_{1(i+1)}w_{(i+1)} \cdots p_{1n}w_n])} \tag{3}$$

Equation 3: Conditional Probability for the other words of the n-gram

The analysis of Equation (2) and Equation (3) highlights the fact that the numerators remain unchanged from one probability to another. Only the denominators change. So, in order to perform a sharp normalisation, it is convenient to evaluate the centre of gravity of the denominators thus defining an average event called the Fair Point of Expectation (FPE). Basically the FPE is the arithmetic mean of the denominators of all the conditional probabilities embodied by Equation (2) and Equation (3). Theoretically, the Fair Point of Expectation is the arithmetic mean of the n joint probabilities[6] of the $(n-1)$-grams contained in an n-gram and it is defined in Equation (4) where the "ˆ" corresponds to a convention commonly used in Algebra that consists in writing a "ˆ" on the top of the omitted terms of a given succession indexed from 2 to n.

$$FPE\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right) = \frac{p\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right)}{\frac{1}{n}\left(p\left([w_2 \ldots p_{2i}w_i \ldots p_{2n}w_n]\right) + \sum_{i-2}^{n} p\left([w_1 \ldots \hat{p_{1i}}\hat{w}_i \ldots p_{1n}w_n]\right)\right)} \tag{4}$$

Equation 4: Fair Point of Expectation

Hence the normalisation of the conditional probability is realised by the introduction of the Fair Point of Expectation into the general definition of the conditional probability. The resulting measure is called the Normalised Expectation and it is proposed as a “fair” conditional probability. The Normalised Expectation is defined in Equation (5).

$$NE\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right) = \frac{p\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right)}{FPE\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right)} \quad (5)$$

Equation 5: Normalised Expectation

3.2 Mutual Expectation

[Justeson/Katz 1993] and [Daille 1995] have shown in their studies that frequency is one of the most relevant statistics to identify multiword terms with specific syntactical patterns. The studies made by [Frantzi/Ananiadou 1996] in the context of extraction of interrupted collocations also assess that the relative frequency is an important clue for the retrieval process. From this assumption, we deduce that between two word *n*-grams with the same Normalised Expectation, the most frequent word *n*-gram is more likely to be a relevant multiword lexical unit. So, the Mutual Expectation between n words is defined in Equation (6) based on the Normalised Expectation and the relative frequency.

$$ME\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right) = p\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right) \times NE\left([w_1 \ldots p_{1i}w_i \ldots p_{1n}w_n]\right) \quad (6)$$

Equation 6: Mutual Expectation

4 The LocalMaxs Algorithm

Most of the approaches proposed in the literature base thier selection process on global association measure thresholds [Church/Hanks 1990] [Smadja 1993] [Daille 1995] [Shimohata 1997]. This is defined by the underlying concept that there exists a limit value of the association measure that allows to decide whether an *n*-gram is a multiword lexical unit or not. However, these thresholds are prone to error as they depend on experimentation. Moreover, they highlight evident flexibility constraints as they have to be re-tuned when the type, the size, the domain and the language of the document change. The LocalMaxs algorithm [Silva et al. 1999] proposes a more flexible and fine-tuned appoach for the election of multiword lexical units as it focuses on the identification of local maxima of the association measure values. Specifically, the LocalMaxs selects multiword lexical units from the set of all the valued word *n*-grams based on two assumptions. First the association measures show that the more cohesive a group of words is, the higher its score will be. Second, multiword lexical units are localized associated groups of words. So, we may deduce that a word *n*-gram is a multiword lexical unit if its association measure value is higher or equal than the association value of all its subgroups of $(n-1)$ words and if it is strictly higher than the association measure values of all its super-groups of $(n+1)$ words.[7] So, let *assoc* be an association measure, *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 words of an *n*-gram, the LocalMaxs is defined as follows:

$$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1}\ W \text{ is a Multiword Lexical Unit if}$$

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

5 Results of SENTA for the IJS-ELAN Corpus

We present an experiment made with SENTA over the IJS-ELAN Slovene-English parallel corpus[8] In this paper, we will focus on the results obtained with four sub-corpora that deal with different domains and vary in size: Vino (EC Council Regulation, 69 Kwords), Vade (Vademecum chemistry, 24 Kwords), Ligs (Linux Installation Guide, 173 Kwords), Usta (Constitution of the Republic of Slovenia, 20 Kwords). Contiguous and non-contiguous terminologically relevant multiword lexical units have been extracted with 74% precision on average. We present the measure of Precision (number of extracted terms/number of extracted multiword lexical units) and the number of terms that have been extracted for each sub-corpus in Table (4).

	Vino	Vade	Ligs	Usta
% Precision	71.55	81.12	74.64	70.76
Number of extracted terms	383	377	1832	310

Table 4: Results of the SENTA software

In order to measure the precision, we took advantage of the bilingual structure of the IJS-ELAN corpus and decided to define a term as a label that refers to a unique concept and which has a unique translation [9]. Thus we have based our evaluation on a broader sense than previous evaluations, considering that terms may not only embody specific Noun Phrase structures. For instance, we believe that expressions like "*come into force*" or "*Getting Started*" can be considered terms as they refer to unique concepts that are translated in a unique way, although some authors would prefer to classify these as Noun+Verb collocations and idioms. For the contiguous case, the results show that base-terms together with terms obtained from overcomposition and co-ordination have been extracted[10] . Correspondingly, most of the non-contiguous terminologically relevant multiword lexical units are terms obtained from modification [11] . We illustrate the results obtained for the sub-corpora under analysis in Appendix A.

However, Senta also extracts terminologically irrelevant multiword lexical units that are mainly functional associations: "*of course*", "*from the*", "*of the*", "*as well as*", "*in order to*", "*like the*", "*has been*", "*may be*", "*as a result of*". So, it si necessary to subdivide the results into two categories: terminologically relevant vs. non-relevant. [Heid 1999] proposes different criteria (depending on the linguistic structure of the units) in order to filter the candidate sets, relative frequency, pre-defined lists of general verbs or adjectives. However, using pre-defined lists of general-purpose units requires determining domain-dependent from domain-independent units, which is not straightforward. As a consequence, we propose a different criterion based on the multidomain structure of the IJS-ELAN corpus. In order to decide whether a multiword lexical unit is terminologically relevant or not, we simply constrain it to occur in only one domain or related domain. So, if the same multiword lexical unit occurs at least in two different domains, it is considered as terminologically irrelevant. Technically, for each one of the four corpora under study, we identified its multiword lexical units that were contained in the set of all the units extracted from all the sub-corpora that deal with a different domain. For example, in order to identify the terminologically relevant multiword lexical units contained in Ligs, we processed the intersection between the set of the multiword lexical units extracted from Ligs and the union of all the sets of multiword lexical units extracted from all the 14 remaining sub-corpora. So, all the units contained in the intersection were not considered for terminological purposes. This experiment shows that the average precision rate is drastically improved up to 82% as a great proportion of domain-independent units can be identified. In parallel the number of extracted terms remains almost unchanged. We present the results in Table (5). The slight loss in recall is mainly due to the Ecmr corpus (Slovenian Economic Mirror, 239 Kwords) that contains terms used in other domains. For example the following terms have been identifed as irrelevant by the experiment as they occur at the same time in a particular domain and in the Ecmr corpus: "*human being*", "*natural resources*", "*The Republic of Slovenia*", "*service provider*". However, the global results show that SENTA largely benefits from the multidomain structure of the IJS-ELAN corpus for the specific task of term extraction.

	Vino	Vade	Ligs	Usta
% Precision	79.29	89.76	80.14	79.64
Number of extracted terms	377	371	1829	308

Table 5: Results of the multidomain experiment

6 Conclusion

SENTA (Software for the Extraction of N-ary Textual Associations) retrieves, form naturally occurring text, terminologically relevant contiguous and non-contiguous multiword lexical units. As it conjugates a new association measure, the Mutual Expectation, with a new acquisition process, the LocalMaxs algorithm, SENTA avoids the definition of global thresholds based on experimentation and does not requiere bootstrapping techniques. The results obtained by applying SENTA to the IJS-ELAN Slovene-English parallel corpus stress the extraction of a great proportion of terms, with 74% precision on average. However, by benefiting from the multidomain structure of the IJS-ELAN corpus, we show that the average precision rate can be drastically improved up to 82% without significantly loosing in recall. However, in order to separate domain-dependent from domain-independent multiword lexical units, we are aware that better results can be obtained by using theoretically defined language models, such as the ones introduced by [Kromer 2000] and [Gotoh/Reginalds 2000].

Acknowledgement

The authors want to thank the reviewers for their valuable remarks and hope that the final version of the paper will have reached their expectations.

Notes

[1] First, relevant 2-grams are retrieved from the corpus. Then, *n*-ary associations may be identified by (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 and ending finally when no more 2-grams are identified.

[2] The results presented in Appendix (1) assess this assumption.

[3] $Position_{12}$ and $Position_{13}$ are respectively the signed distances between w_1 and w_2 and between w_1 and w_3. The sign "+"("–") is used for words on the right (left) of w_1.

[4] [Dias *et al.* 2000] shows that the Mutual Expectation leads to improved results comparing to well-known normalised mathematical models like the Dice coefficient [Smadja 1993], the association ratio [Church/Hanks 1990], the Φ^2 [Gale 1991] and the Log-likelihood ratio [Dunning 1993].

[5] More Details about the probability space can be found in [Dias *et al.* 2000].

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

[7] It is possible to compare association measure values of vectors of different length due to the normalisation process.

[8] The IJS-ELAN corpus can be downloaded at `http://nl.ijs.si/elan/`.

[9] This definition has first been introduced by [Daille 1995].

[10] The extraction of terms obtained by over-composition is a great issue for automatic construction of thesaurus, word sense disambiguation and pp-attachment problems.

[11] The notions of base-terms, overcomposition, modification and co-ordination are defined in [Daille 1995].

References

[Bourigault 1996] Bourigault, Didier (1996). "Lexter: a Natural Language Processing Tool for Terminology Extraction", in *Euralex Proceedings*

[Church/Hanks 1990] Church, Kenneth and Hanks, P. (1990). "Word Association Norms Mutual Information and Lexicography", in *Computational Linguistics*, Vol. 16(1), pp 23-29.

[Dagan 1994] Dagan, Ido and Church, Kenneth (1994) "Termight: Identifying and Translating Technical Terminology", in 4th *Conference on Applied Natural Language Processing*

[Daille 1995] Daille, Béatrice (1995). "Study and Implementation of Combined Techniques for Automatic Extraction of Terminology" in *The Balancing Act Combining Symbolic and Statistical Approaches to Language*, Cambridge, MIT Press.

[Dias et al. 1999] Dias, Gaël, Guilloré, Sylvie and Lopes, José G. P. (1999) "Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora", in *Traitement Automatique des Langues Naturelles (TALN)*, Cargèse, France. July 12-17.

[Dias *et al.* 2000] Dias, Gaël, Guilloré, Sylvie and Lopes, José 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* (ACIDCA '00), Monastir, Tunisia.

[Dunning 1993] Dunning, Ted (1993) "Accurate Methods for the Statistics of Surprise and Coincidence", in *Association for Computational Linguistics*, Vol. 19-1

[Frantzi/Ananiadou 1996] Frantzi, Katerina T. and Ananiadou, Sophia, “Extracting Nested Collocations” in 16th *International Conference on Computational Linguistics (COLING ’96)*, pp. 41-46.

[Gale 1991] 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.

[Gotoh/Reginalds 2000] Gotoh Yoshihiko and Renalds Steve (2000), “Variable Word Rate N-grams”, in 25th *International Conference on Acoustics, Speech and Signal Processing (ICASSP00).*

[Habert/Jacquemin 1998] Habert, Benoît and Jacquemin, Christian (1993) “Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques” in *Traitement Automatique des Langues*, Vol. 34-2

[Heid 1999] Heid, Ulrich (1999) “Extracting Terminologically Relevant Collocations from German Technical Texts” in *TKE’99*.

[Justeson/Katz 1993] Justeson, John and Katz, Slava (1993) “Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text”. *Research Report RC 18906 (82591) IBM*.

[Kromer 2000] Kromer, Victor (2000) “An Usage Measure Based on Psychophysical Relations” http://xxx.lanl.gov/abs/cs.CL/0002017

[Shimohata 1997] Shimohata, S (1997). “Retrieving Collocations by Co-occurrences and Word Order Constraints”, in *ACL-EACL’97*, pp.476-481

[Silva et al. 1999] Silva, Joaquim F., Dias, Gaël, Guilloré, Sylvie and Lopes, José G. P. (1999) “Using Local Maxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units”, in 9th *Portuguese Conference in Artificial Intelligence*, Springer-Verlag.

[Sinclair 1974] Sinclair, John (1974). *English Lexical Collocations: A study in computational linguistics*. Reprinted as chapter 2 of Foley, J.A. (ed). (1996). J.M. Sinclair on Lexis and Lexicography, Singapore: Uni Press.

[Smadja 1993] Smadja, Frank (1993). “Retrieving Collocations from Text: XTRACT” in *Computational Linguistics*, Vol.19-1.

[Smadja 1996] Smadja, Frank (1996). “Translating Collocations for Bilingual Lexicons: A Statistical Approach”, in *Association for Computational Linguistics*, Vol.22-1

A Samples of Extracted Terms

Vino	
Base-terms	*standard quality, flat-rate import value, exported with refunds, in terms of quantity, per tonne, the cif, at regular intervals*
Overcomposition	*agricultural products from third countries, common organisation of agricultural markets*
Modification	*derived from the* _____ *price* where the gap can be fulfilled with *intervention* or *buying-in*.
Co-ordination	*colza and rape seed, prices and availabilities of products*

Vade	
Base-terms	*intestinal spasms, lidocaine hydrochloride monohydrate, the skin, suitable for ederly people, recipe based on hundred years of experience*
Overcomposition	*therapeutic agents derived from natural sources, rhinolytic symptomatic treatment for common cold*
Modification	*children above* _____ *years* where the gap can be fulfilled with *five* or *twelve*
Co-ordination	*rheumatic and muscular pain relief, wounds and bones fractures*

Ligs	
Base-terms	*Boot Diskette, C compiler, core dump, World Wide Web, Install via NFS, access to a TCP/IP network, FTP e-mail servers, Digital Equipment Corporation, the kernel*
Overcomposition	*Press Tab at the boot prompt, Linux Documentation Project Home Page, Debian/GNU Linux Packages file*
Modification	*the* _____ *environment variable* where the gap can be fulfilled with *EDITOR*, *HOME* or *PATH*
Co-ordination	*input and output files, input and standard output*

Usta	
Base-terms	*Central Bank, Constitutional Court, come into effect, rights and obligations, for public purposes, President of the Republic, local government, the Constitution*
Overcomposition	*adherence to international agreements, principles of international laws, elections for the National Assembly*
Modification	*a* _____ *of no confidence* where the gap can be fulfilled with *vote* or *motion*
Co-ordination	*Italian and Hungarian ethnic communities, State and local government bodies*

In the case of the terms that embody a modification structure, the non-contiguous unit is the generalisation of a concept that can be realised with different words. A detailed analysis of these terms is necessary as they allow to determine hapaxes (i.e. multiword lexical units that occur only once in the text). Indeed, the instantiation of the concept *a* ____ *of no confidence* can be realised by the occurrence of *vote* that leads to the following hapax: *a vote of no confidence*.