

Normalising the IJS-ELAN Slovene-English Parallel Corpus for the Extraction of Multilingual Terminology

Gaël Dias
Universidade da Beira Interior

Spela Vintar
University of Ljubljana

José Gabriel Pereira Lopes
Universidade Nova de Lisboa

Sylvie Guilloré
Université d'Orléans

Abstract

Various efforts have been made for the development of tools and methods dedicated to the automatic processing of multilingual terminology databases. For that purpose, multilingual parallel corpora have been used as a basis resource. However, most of the neologisms in technical and scientific domains are realised by multiword terms that are rarely identified in parallel corpora. In this paper, we propose the normalisation of the IJS-ELAN Slovene-English parallel corpus by using the language-independent SENTA software.

1 Introduction

The need for multilingual terminology resources has become particularly acute owing to the globalization of scientific and technical exchanges and the concurrent development of international communication networks. As a consequence, various efforts have been made for the development of tools and methods dedicated to the automatic processing of multilingual terminology databases. For that purpose, multilingual parallel corpora have been used as a basis resource. However, most of the neologisms in technical and scientific domains are realised by multiword lexical units that are rarely identified in parallel corpora. For example, *World Wide Web*, *IP address* and *TCP/IP network* are multiword terms that clearly need to be identified in the corpora as concept units.

In order to identify multiword terms from text corpora, three main strategies have been

proposed in the literature. Purely linguistic systems (David, 1990; Dagan, 1993; Bourigault, 1996) propose to recognise relevant terms by using techniques that analyse specific syntactical structures in the texts. However, this methodology suffers from its monolingual basis, as the systems require highly specialised linguistic techniques to identify clues that isolate possible candidate terms. Hybrid linguistic-statistical methods (Enguehard, 1993; Justeson, 1993; Daille, 1995; Heid, 1999) define co-occurrences of interest in terms of syntactical patterns and statistical regularities. However, by reducing the searching space to groups of words that correspond to *a priori* defined syntactical patterns (Noun+Adj, Noun+Prep+Noun etc...), such systems do not deal with a great proportion of terms and introduce noise in the retrieval process. Finally, purely statistical systems (Church, 1990; Dunning, 1993; Smadja, 1993; Shimohata, 1997) detect discriminating multiword terms from text corpora by means of association measure regularities. As they use plain text corpora and only require the information appearing in texts, such systems are highly flexible and identify relevant units independently from the domain and the language of the input text. However, they emphasise two major drawbacks. On one hand, by relying on *ad hoc* establishment of global thresholds they are prone to error. On the other hand, as they only allow the acquisition of binary associations, these systems must apply enticement techniques to acquire multiword terms with more than two words. [1] Unfortunately, such techniques have

¹ First, relevant 2-grams are retrieved from the corpus. Then, n-ary

shown their limitations as their retrieval results mainly depend on the identification of suitable bigrams for the initiation of the iterative process.

In order to overcome the problems highlighted by the previous statistical systems, we propose a new architecture called SENTA (Software for the Extraction of N-ary Textual Associations). SENTA conjugates a new association measure called the Mutual Expectation (Dias, 1999) with a new acquisition process called the LocalMaxs (Silva, 1999). On one hand, the Mutual Expectation, based on the concept of Normalised Expectation, evaluates the degree of cohesiveness that links together all the words contained in an n -gram (i.e. $\forall n, n \geq 2$). On the other hand, the LocalMaxs retrieves the candidate terms 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 defined by experimentation.

In this paper, we show how SENTA can be used to normalise the IJS-ELAN Slovene-English parallel corpus (Erjavec, 1999; Vintar, 1999) for the specific task of automatic extraction of bilingual terminology. First, SENTA is run independently over the Slovene and the English sub-corpora in order to identify potential multiword terms. Then, the multiword term candidates are marked in the corpus with SGML markups following the Text Encoding Initiative (TEI) thus proposing a suitable normalised corpus for bilingual terminology extraction.

In the first section of this paper, we will present the IJS-ELAN Slovene-English parallel corpus. Then, we will respectively introduce the global architecture of the SENTA system around the three following topics: positional n -grams, Mutual Expectation and LocalMaxs algorithm. Finally, in the fifth section, we will access the results obtained after the normalisation.

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 IJS-ELAN Corpus

The IJS-ELAN Slovene-English parallel corpus has been compiled at the Jožef Stefan Institute in Ljubljana, Slovenia (Erjavec 1999; Vintar 1999) within the framework of the EU ELAN project (European Language Activity Network).[2] It consists of 15 original texts and their translations, together amounting to 1 million words. The texts were chosen according to several criteria, though in practice the main factors were terminological relevance and overall availability of the text, both in terms of copyright and digital format.

The texts can be divided into the following groups:

- Legislation: texts pertaining to the Slovenian accession to the EU (33%),
- Slovenian Economic Mirror: a periodical (23%),
- Linux user's guide: a glossary of computer terms (18%),
- "1984": novel by George Orwell (18%),
- Political speeches of the Slovenian president Mr Kučan (6%),
- Lek Vademecum: catalogue of medical products (2%).

Each of the parallel texts was first converted into plain text format and "cleaned up", which also meant removing graphical elements such as pictures and tables from text, and then aligned. For the alignment two different tools were used (1) the Unix-based Vanilla aligner and (2) the Windows-based aligner component of the Translation Memory System DéjàVu by Atril.[3] These were in turn tokenised and marked with SGML tags according to the TEI guidelines.

The texts were kept in 15 separate files containing a header and a body. The header contains all administrative data about the text, (e.g. title, source, length, provider tools used for processing) and the body is divided into translation units. Each translation unit contains two language segments, Slovene and English or

² The corpus is freely available and can be accessed at <http://nl.ijs.si/elan/>.

³ Both required hand-validation of the results.

vice versa, depending on which of the two was the source language as illustrated in Figure (1).

The organisation of language segments into translation units is a feature related to the notion of Translation Memory (TM) and was chosen deliberately so as to facilitate the conversion of the parallel text into a TM format.[4]

```
<tu lang="sl-en" id="parl.14">
<seg lang="sl"> <w>&Ccaron;e</w>
<w>je</w> <w>predsednik</w>
<w>odsoten</w> <c>,</c> <w>ga</w>
<w>nadome&scaron;&ccaron;a</w>
<w>tisti</w> <w>podpredsednik</w>
<c>,</c> <w>ki</w> <w>ga</w>
<w>dolo&ccaron;i</w> <w>predsednik</w>
<w>dr&zcaron;avnega</w> <w>zbor</w>
<c>.</c> </seg>
<seg lang="en"> <w>If</w> <w>the</w>
<w>President</w> <w>is</w> <w>absent</w>
<c>,</c> <w>he</w> <w>is</w>
<w>replaced</w> <w>by</w> <w>a</w>
<w>Deputy</w> <w>President</w>
<w>appointed</w> <w>by</w> <w>him</w>
<c>.</c> </seg>
</tu>
```

Figure 1: Sample Translation Unit

The goal of our experiment was to identify multiword term candidates in the translation units of the IJS-ELAN corpus and mark them with SGML markups. Indeed, the specific task of terminology extraction would deeply benefit from the normalisation of the corpus i.e. from the identification of multiword terms such as `<w>Deputy</w>` `<w>President</w>` that corresponds to a single concept. Indeed, its translation is realised in Slovene by the single word term `<w>podpredsednik</w>`. As a consequence, the normalisation of the corpus would result in the introduction of SGML markups that would highlight the multiword terms as illustrated in the following expression: `<mwu>` `<w>Deputy</w>` `<w>President</w>` `</mwu>` where `mwu` stands for multiword unit. [5] For that purpose, we used the SENTA architecture that is explained in the following section.

3 Architecture of SENTA

SENTA has been thought and developed around the idea of total flexibility. As it is exclusively based on a new probabilistic measure and a new acquisition process, SENTA detects multiword lexical units (MWUs) by processing only once a corpus of any language, any domain or any type. SENTA takes as input, a text corpus that is neither lemmatised 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 MWUs. Indeed, according to Justeson (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 MWU. Based on this assumption, we believe that multiword lexical units are sufficiently fixed and recurrent sequences of words to propose that they should be identified without the introduction of any extra-linguistic information. As a consequence, we opted not to modify the input text and work on all the information contained inside the corpus. The global architecture of SENTA is designed around four sequential steps.

First, SENTA performs the transformation of the input text into a set of databases of n-grams. 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 n-gram in a six words wide window (i.e. three words to the left of the considered word and three on its right hand side). One non-contiguous bigram and one contiguous bigram are respectively shown in the first two rows of Table (1), taking as current input the English sentence of Figure (1) and `<w>Deputy</w>` (i.e. w_i) the word under study.

As notation is concerned, the non-contiguous bigram presented in the first row of Table (1) may be characterised by one of the following equivalent expressions where a gap (i.e. “___”)

⁴ Some of the text providers, especially the Government Office of European Affairs, were very eager to benefit from this kind of exchange and we indeed returned all the texts we obtained from them in the TRADOS translation memory format `.tmw`.

⁵ We will access later in the article the SGML notation.

embodies the set of all the occurrences in the corpus that fulfil the free space:

$$\langle w \rangle \text{by} \langle /w \rangle \text{ ___ } \langle w \rangle \text{Deputy} \langle /w \rangle \quad (\text{a})$$

$$[\langle w \rangle \text{Deputy} \langle /w \rangle \text{ } -2 \text{ } \langle w \rangle \text{by} \langle /w \rangle] \quad (\text{b})$$

Similarly, the contiguous bigram of the second row may be characterised by one of the following equivalent expressions:

$$\langle w \rangle \text{Deputy} \langle /w \rangle \langle w \rangle \text{President} \langle /w \rangle \quad (\text{c})$$

$$[\langle w \rangle \text{Deputy} \langle /w \rangle \text{ } 1 \text{ } \langle w \rangle \text{President} \langle /w \rangle] \quad (\text{d})$$

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

w_1	p_{12}	w_2
$\langle w \rangle \text{Deputy} \langle /w \rangle$	-2	$\langle w \rangle \text{by} \langle /w \rangle$
$\langle w \rangle \text{Deputy} \langle /w \rangle$	1	$\langle w \rangle \text{President} \langle /w \rangle$

Table 1: Two 2-grams containing $\langle w \rangle \text{Deputy} \langle /w \rangle$ [6]

Following this first step, 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 sections 3 and 4, we rigorously define the Mutual Expectation and the LocalMaxs.

4 The Mutual Expectation Measure

In order to evaluate the degree of cohesiveness existing between words, various mathematical models have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness between two words and do not generalise for the case of n individual words (Church, 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 enticement 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 bigrams for the initiation of the iterative process. On the other hand, the proposed mathematical models tend to be over-sensitive to frequent words. In particular, this has lead researchers to consider function words like determinants or prepositions meaningless to the sake of the statistical evaluation process and to test association measures on plain word pairs (Daille, 1995).

In order to overcome both problems, we introduce a new association measure called the Mutual Expectation that evaluates the degree of rigidity that links together all the words contained in an n-gram ($\forall n, n \geq 2$) based on the concept of Normalised Expectation.

4.1 Normalised Expectation

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 Normalised 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 textual units is, that is the less it accepts the loss of one of its components, the higher its Normalised Expectation will be.

For example, the Normalised Expectation for $[\langle w \rangle \text{Linux} \langle /w \rangle \text{ } 1 \text{ } \langle w \rangle \text{Operating} \langle /w \rangle \text{ } 2 \text{ } \langle w \rangle \text{System} \langle /w \rangle]$ must take into account the cost of the loss of one the three individual words $\langle w \rangle \text{Linux} \langle /w \rangle$, $\langle w \rangle \text{Operating} \langle /w \rangle$ and $\langle w \rangle \text{System} \langle /w \rangle$ one at a time of the n-gram. Thus, the average expectation of the 3-gram must take into account the expectation of occurring the word $\langle w \rangle \text{System} \langle /w \rangle$ after $\langle w \rangle \text{Linux} \langle /w \rangle \langle w \rangle \text{Operating} \langle /w \rangle$, but also the expectation of $\langle w \rangle \text{Operating} \langle /w \rangle$ linking together $\langle w \rangle \text{Linux} \langle /w \rangle$ and $\langle w \rangle \text{System} \langle /w \rangle$ and finally, the expectation of occurring $\langle w \rangle \text{Linux} \langle /w \rangle$ before $\langle w \rangle \text{Operating} \langle /w \rangle \langle w \rangle \text{System} \langle /w \rangle$. This situation is illustrated in Table (2) where one possible expectation corresponds to one respective row.

⁶ In Table 1, p_{12} is the signed distance between w_1 and w_2 . The sign "+" ("-") is used for words on the right (left) of w_1 .

The underlying concept of the Normalized Expectation is based on the conditional probability.

Expectation to occur <w>Linux</w> knowing the gapped 3-gram [_ 1 <w>Operating</w> 2 <w>System</w>]
Expectation to occur <w>Operating</w> knowing the gapped 3-gram [<w>Linux</w> 1 _ 2 <w>System</w>]
Expectation to occur <w>System</w> knowing the gapped 3-gram [<w>Linux</w> 1 <w>Operating</w> 2 _]

Table 2: Example of expectations

Indeed, the conditional probability measures the expectation of the occurrence of an event $X=x$ knowing that an event $Y=y$ stands as illustrated in Equation (1)

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

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 .

As each word of the text corpus can be mapped to a discrete random variable in a given probability space, the definition of the conditional probability can be applied in order to measure the 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.[7]

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.

⁷ More Details about the probability space can be found in (Dias,2000a).

In order to explain this process, let's consider the following n -gram $[w_1 p_{12} w_2 \dots p_{1i} w_i \dots p_{1n} w_n]$. The extraction of one word at a time from the generic n -gram gives rise to the occurrence 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	word
[<u> </u> $w_2 p_{23} w_3 \dots p_{2i} w_i \dots p_{2n} w_n$]	w_1
[w_1 <u> </u> $p_{13} w_3 \dots p_{1i} w_i \dots p_{1n} w_n$]	w_2
...	...
[$w_1 \dots p_{1(i-1)} w_{(i-1)}$ <u> </u> $p_{1(i+1)} w_{(i+1)} \dots p_{1n} w_n$]	w_i
...	...
[$w_1 \dots p_{1i} w_i \dots p_{1(n-1)} w_{(n-1)}$ <u> </u>]	w_n

Table 3: (n-1)-grams and missing words

So, each event may be associated to a respective conditional probability that evaluates the expectation to occur the missing word knowing its corresponding (n-1)-gram. The n conditional probabilities are introduced in Equation (2) and Equation (3). Equation (2) evaluates the cost of the loss of the first word of the n -gram (i.e. the pivot word).

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

Equation (3) evaluates the cost of the loss of all the other words of the n -gram.

$$\forall i, i = 2..n, \\ p(w_i | [w_1 \dots p_{1(i-1)} w_{(i-1)} p_{1(i+1)} w_{(i+1)} \dots p_{1n} w_n]) = \frac{p([w_1 \dots p_{1i} w_i \dots p_{1n} w_n])}{p([w_1 \dots p_{1(i-1)} w_{(i-1)} p_{1(i+1)} w_{(i+1)} \dots p_{1n} w_n])} \quad (3)$$

The analysis of the 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 gravity centre 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 of the $(n-1)$ -grams contained in an n -gram and it is defined in Equation (4). [8]

$$FPE([w_1 p_{12} w_2 \dots p_{1i} w_i \dots p_{1n} w_n]) = \frac{1}{n} \left(p([w_2 \dots p_{2i} w_i \dots p_{2n} w_n]) + \sum_{i=2}^n p\left(\left[\overset{\wedge}{w_1} \dots p_{1i} \overset{\wedge}{w_i} \dots p_{1n} w_n \right] \right) \right) \quad (4)$$

where $p([w_2 \dots p_{2i} w_i \dots p_{2n} w_n])$, for $i=3, \dots, n$, is the probability of the occurrence of the $(n-1)$ -gram $[w_2 \dots p_{2i} w_i \dots p_{2n} w_n]$ which is the result of the extraction of w_1 from the whole n -gram and $p\left(\left[\overset{\wedge}{w_1} \dots p_{1i} \overset{\wedge}{w_i} \dots p_{1n} w_n \right] \right)$ is the probability of the occurrence of one $(n-1)$ -gram containing necessarily the first word w_1 . The " \wedge " corresponds to a convention frequently used in Algebra that consists in writing a " \wedge " on the top of the omitted term of a given succession indexed from 2 to n .

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 symmetric resulting measure is called the Normalised Expectation and it is proposed as a "fair" conditional probability. The Normalised Expectation is defined in Equation (5) where $p([w_1 \dots p_{1i} w_i \dots p_{1n} w_n])$ is the relative frequency of $[w_1 \dots p_{1i} w_i \dots p_{1n} w_n]$ and $FPE([w_1 \dots p_{1i} w_i \dots p_{1n} w_n])$ is the Fair Point of Expectation defined in Equation (4).

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

4.2 Mutual Expectation

Daille (1995) and Justeson (1993) evidence that one effective criterion for multiword lexical unit identification is frequency. From this assumption, we pose that between two n -grams with the same Normalised 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 (6) based on the Normalised Expectation and the relative frequency.

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

$p([w_1 \dots p_{1i} w_i \dots p_{1n} w_n])$ and $NE([w_1 \dots p_{1i} w_i \dots p_{1n} w_n])$ are respectively the relative frequency of the particular n -gram $[w_1 \dots p_{1i} w_i \dots p_{1n} w_n]$ and its Normalised Expectation.

5 The LocalMaxs Algorithm

Electing multiword terms among the sample space of all the valued word n -grams may be defined as detecting combinations of features that are common to all the instances of the concept of multiword term. In the case of purely statistical methods, frequencies and association measure values are the only features available to the system. Consequently, most of the approaches have based their selection process on the definition of global frequency thresholds and/or on the evaluation of global association measure thresholds (Church, 1990; Smadja, 1993; Daille, 1995; Shimohata, 1997; Feldman, 1998). This is defined by the underlying concept that there exists a limit value of the association measure that allows to decide whether a word n -gram is a pertinent word association or not.

However, these thresholds are prone to error as they 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 document change (Habert, 1997). [9]

The LocalMaxs (Silva, 1999) proposes a more flexible and fine-tuned approach for the selection process as it concentrates on the identification of local maxima of association measure values. Specifically, the LocalMaxs elects multiword terms 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 textual units is, the

⁸ In the case of $n=2$, the FPE is the arithmetic mean of the marginal probabilities.

⁹ They obviously vary with the association measure.

higher its score will be. [10] Second, multiword terms are localised associated groups of words. So, we may deduce that a word n-gram is a multiword term if its association measure value is higher or equal than the association measure values of all its sub-groups of (n-1) words and if it is strictly higher than the association measure values of all its super-groups of (n+1) words.

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 a word n-gram. The LocalMaxs is defined as follows:

$$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1}$$

W is a multiword lexical unit if

$$\begin{aligned} & (sizeof(W)=2 \wedge assoc(W) > assoc(y)) \\ & \vee \\ & (sizeof(W) \neq 2 \wedge assoc(W) \geq assoc(x) \wedge \\ & assoc(W) > assoc(y)) \end{aligned}$$

The LocalMaxs evidences two interesting properties. On one hand, it allows the testing of various association measures that respect the first assumption described above (i.e. the more cohesive a sequence of words is, the higher its association measure value will be). Using this property, we performed many experiments with different association measures. In particular, we tested the following normalised mathematical models: the Association Ratio (Church, 1990), the Dice coefficient (Smadja, 1996), the ϕ^2 (Gale, 1990) and the Log-Likelihood Ratio (Dunning, 1993).[11]

On the other hand, the LocalMaxs allows extracting multiword terms obtained by composition. Indeed, as the algorithm retrieves pertinent units by analysing their immediate context, it may identify multiword terms that are composed by one or more other terms. For example, the LocalMaxs conjugated with the Mutual Expectation, elects the multiword term

`<w>Operating</w>` `<w>System</w>`
`<w>Windows</w>` `<w>NT</w>` built from the composition of the extracted multiword terms `<w>Operating</w>` `<w>System</w>` and `<w>Windows</w>` `<w>NT</w>`. This situation is illustrated in Figure (2).

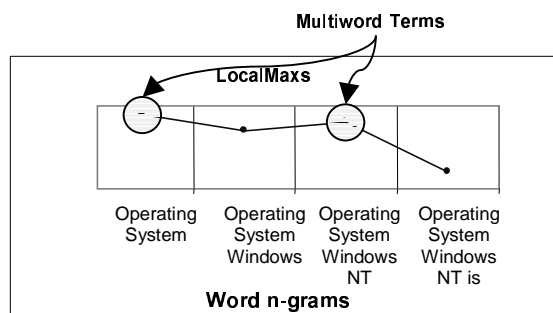


Figure 2: Illustration of the LocalMaxs Algorithm

Indeed, roughly exemplifying, one can expect that there are many operating systems. Therefore, the association measure value of `<w>Operating</w>` `<w>System</w>` `<w>Windows</w>` should be lower than the one for `<w>Operating</w>` `<w>System</w>` as there are many possible words, other than `<w>Windows</w>`, that may occur after `<w>Operating</w>` `<w>System</w>`. Thus, the association measure of any super-group containing the unit `<w>Operating</w>` `<w>System</w>` should theoretically be lower than the association measure for `<w>Operating</w>` `<w>System</w>`. But, if the first name of the operating system is `<w>Windows</w>`, the expectation to appear `<w>NT</w>` is very high and the association measure value of `<w>Operating</w>` `<w>System</w>` `<w>Windows</w>` `<w>NT</w>` should then be higher than the association measure values of all its sub-groups and super-groups, as in the latter case, no word can be expected to strengthen the overall unit `<w>Operating</w>` `<w>System</w>` `<w>Windows</w>` `<w>NT</w>`.

6 Marking Multiword Lexical Units

The Text Encoding Initiative (TEI) has focused its research efforts on defining a set of generic Guidelines for the representation of textual

¹⁰ The conditional entropy measure is one of the exceptions.

¹¹ Cramer and Pearson coefficients (Bhattacharyya, 1977) have also been tested. In all cases, the Mutual Expectation has overperformed the other measures.

materials in electronic form. In particular, the TEI Guidelines provide a means of making explicit certain features of a text in such a way as to aid its processing by computer programs running on different machines. This process of making explicit is called *markup* or *encoding*. For that purpose, the TEI Guidelines use the Standard Generalised Markup Language (SGML). So, in order to normalise the IJS-ELAN parallel corpus by identifying its multiword terms, we decided to follow the TEI Guidelines thus defining new elements and attributes.

In the IJS-ELAN DTD the element `w` is defined as follows which means that a token may be formed by combinations of strings, segments, words, morphemes or characters:

```
<!ELEMENT w --( #PCDATA | seg | w | m | c ) * >
```

A multiword lexical unit can be defined by a sequence of words or multiword lexical units and it must contain at least one word or one multiword lexical unit. So, in order to define the new element for multiword lexical units we propose the following definition:

```
<!ELEMENT mwu --( w | mwu ) + >
```

Possible encodings are evidenced in the examples illustrated in Figure 3.

```
<mwu><mwu>
<w>Operating</w><w>System</w></mwu>
<mwu><w>Windows</w><w>NT</w></mwu>
</mwu>
<mwu>
<w>Linux</w>
<mwu><w>Operating</w><w>System</w>
</mwu></mwu>
```

Figure 3: Example of encoding

Moreover, as SENTA extracts contiguous and non-contiguous multiword lexical units we had to define a list of attributes for the `mwu` element as defined as follows:

```
<!ATTLIST mwu
  sequence (cont | non-cont) cont
  id ID #IMPLIED
  next ID #IMPLIED
  prev ID #IMPLIED >
```

where `sequence` stands for contiguous or non-contiguous and is contiguous by default, `prev` links the current unit to the previous unit, `next` links the current unit to the next unit and `id` is a unique code for each `mwu`. [12]

```
<mwu sequence="non-cont" id=mwu15
next=mwu16>
<w>to</w> <w>allow</w></mwu>
. . . .
<mwu sequence="non-cont" id=mwu16
prev=mwu15>
<w>to</w></mwu>
```

Figure 4: Encoding non-contiguous sequences

For example, the non-contiguous multiword lexical unit `<w>to</w> <w>allow</w> ... <w>to</w>` may be encoded by the following SGML markups illustrated in Figure (4).

7 Normalising the IJS-ELAN Corpus

SENTA has been applied to each one of the fifteen parallel texts in the IJS-ELAN Slovene-English parallel corpus and produced a list of multiword lexical units for each language. Then, the multiword lexical units have been marked in the overall corpus to produce the normalised version. In the first subsection we will focus on the result of the experiment and in the second, we will discuss about the work that is being carried on.

7.1 Evaluation

SENTA has produced lists of multiword lexical units containing word groups of length ranging from 2 to 6 and also including non-contiguous units for each one of the languages. Not surprisingly, we have been faced to different results between Slovene and English namely in the number of extracted MWUs. We propose some figures in Table (4).[13] [14]

¹² The `id`, `next` and `prev` attributes are necessary for marking non-contiguous multiword lexical units.

¹³ The non-contiguous units evidence problems that need to be dealt apart from the contiguous. In this early stage of our work we will focus exclusively on the contiguous multiword lexical units.

¹⁴ We remind the reader that in order to be identified, a MWU must occur at least twice, which means that the inflected form of the MWU must occur at least twice.

	Slovene	English
Tokens	112,669	125,796
MWUs	1,934	4,697
MWUs in words	6,699	20,792

Table 4: Sample text ECMR

The ratio between Slovene and English recall is approximately 1:3. This is particularly due to the fact that Slovene is part of the Slavic languages that handle compounds using morphological transformations. For example, Deputy President is translated into the single word *podpredsednik* in Slovene. These results strengthen our original idea that normalisation of text corpora is a fundamental issue for the extraction of terminology databases.

The results of the normalisation are encouraging although they differ in terms of quality, as we will evidence it with some examples. Correct normalisation are obtained i.e. all the multiword terms are identified in the Slovene and English translation units as illustrated in Figure (5) where *operacijski sistem* is the translation of operating system and *Motherboard* is translated into *Mati•na ploš•a*.

```
<tu lang="en-sl" id="lig5.17">
<seg lang="en"> <w>It</w> <w>is</w>
<w>a</w> <w>true</w> <w type=dig>32-
bit</w> <mwu> <w>operating</w>
<w>system</w> </mwu> <w>solution</w>
<c>.</c> </seg>
<seg lang="sl"> <w>Linux</w> <w>je</w>
<w>pravi</w> <w type=dig>32-bitni</w>
<mwu> <w>operacijski</w> <w>sistem</w>
</mwu> <c>.</c> </seg>
</tu>
```

```
<tu lang="en-sl" id="lig5.730">
<seg lang="en"> <w>Motherboard</w>
<w>and</w> <w>CPU</w>
<w>requirements</w> <c>.</c> </seg>
<seg lang="sl"> <mwu>
<w>Mati&ccaron;na</w>
<w>plo&scaron;&ccaron;a</w> </mwu>
<w>in</w> <w>procesorske</w>
<w>zahteve</w> <c>.</c> </seg>
</tu>
```

Figure 5: Normalised Translation Units

Other normalisations are incomplete as SENTA failed to identify some multiword lexical units. For example, although *source code* has been well identified in two different inflected forms i.e. *izvorna koda* and *izvorne kode* and

device driver i.e. *gonilnik naprave* and part of i.e. *kot del* have also been detected, *parallel port* and the *kernel* have not been recognised in Slovene as illustrated in Figure (6).

```
<seg lang="en"> <w>However</w>
<c>,</c> <w>the</w> <mwu> <w>source</w>
<w>code</w> </mwu> <w>for</w>
<w>the</w> <w>Zip</w> <mwu>
<w>parallel</w> <w>port</w> </mwu>
<mwu> <w>device</w> <w>driver</w>
</mwu> <w>is</w> <w>included</w>
<w>as</w> <mwu> <w>part</w> <w>of</w>
</mwu> <mwu> <w>the</w> <w>kernel</w>
</mwu> <mwu> <w>source</w> <w>code</w>
</mwu> <w>distribution</w> <c>.</c>
</seg>
<seg lang="sl"> <w>Vendar</w>
<w>pa</w> <w>je</w> <mwu>
<w>izvorna</w> <w>koda</w> </mwu>
<w>za</w> <mwu> <w>gonilnik</w>
<w>naprave</w> </mwu> <w>Zip</w>
<w>na</w> <w>vzporednih</w>
<w>vratih</w> <w>vklju&ccaron;ena</w>
<mwu> <w>kot</w> <w>del</w>
<w>distribucije</w> </mwu> <mwu>
<w>izvorne</w> <w>kode</w> </mwu>
<w>jedra</w> <c>.</c> </seg>
</tu>
```

Figure 6: Normalised Translation Unit

Finally, some results clearly show the limitations of pure statistical methodologies. Indeed, in order to be identified, a MWU must occur at least twice in the corpus. However, a great deal of multiword terms occur just once and can not be detected.

For example, in Figure (7), *customs* *duties* and *sugar* *products* and *carinskih* *dajatev* and *sladkorne* *izdelke* were not detected as multiword units.

Noisy multiword lexical units are also marked. For example, *na* *trgu* *Skupnosti* is the equivalent of *the* *Community* *market* (correctly recognised in English) but just *na* *trgu* is identified as a relevant unit.

```

<tu lang="en-sl" id="vino.17">
<seg lang="en"> <w>whereas</w>
<c>,</c> <w>moreover</w> <c>,</c>
<w>in</w> <w>order</w> <w>to</w>
<w>avert</w> <w>problems</w> <w>of</w>
<w>supply</w> <w>to</w> <mwu>
<w>the</w> <w>Community</w>
<w>market</w> </mwu> <c>,</c>
<w>the</w> <w>suspension</w> <w>of</w>
<w>customs</w> <w>duties</w> <w>on</w>
<w>certain</w> <w>sugar</w>
<w>products</w> <w>should</w> <w>be</w>
<w>permitted</w> <c>;</c> </seg>
<seg lang="sl"> <w>glede</w>
<w>na</w> <w>to</w> <c>,</c>
<w>da</w> <w>je</w> <w>zaradi</w>
<w>prepre&ccaron;evanja</w>
<w>problemov</w> <w>ponudbe</w> <mwu>
<w>na</w> <w>trgu</w> </mwu>
<w>Skupnosti</w> <w>treba</w>
<w>dovoliti</w> <w>za&ccaron;asno</w>
<w>ukinitev</w>
<w>pla&ccaron;evanja</w>
<w>carinskih</w> <w>dajatev</w>
<w>za</w> <w>dolo&ccaron;ene</w>
<w>sladkorne</w> <w>izdelke</w>
<c>;</c> </seg>
</tu>

```

Figure 7: Normalised Translation Unit

Moreover, statistical works based on the study of text corpora identify textual associations in the context of their usage. As a consequence, many multiword units can not be considered as terms (although their identification is useful). For example, in Figure (8), <w>be</w> <w>notified</w> and <w>valid</w> <w>for</w> are not terms.

```

<tu lang="en-sl" id="vino.830">
<seg lang="en"> <w>the</w>
<w>Council</w> <w>and</w> <w>the</w>
<w>Member</w> <w>States</w>
<w>shall</w> <mwu> <w>be</w>
<w>notified</w> </mwu> <w>of</w>
<w>such</w> <w>measures</w> <c>,</c>
<w>which</w> <w>shall</w> <w>be</w>
<mwu> <w>valid</w> <w>for</w> </mwu>
<w>no</w> <w>more</w> <w>than</w> <mwu>
<w>six</w> <w>months</w> <w>and</w>
</mwu> <w>shall</w> <w>be</w>
<w>immediately</w> <w>applicable</w>
<c>.</c> </seg>
<seg lang="sl"> <w>Svet</w> <w>in</w>
<w>dr&zcaron;ave</w>
<w>&ccaron;lanice</w> <w>se</w>
<w>obvesti</w> <w>o</w> <w>teh</w>
<w>ukrepih</w> <c>,</c> <w>ki</w>
<w>pa</w> <w>ne</w> <w>smejo</w>
<w>veljati</w> <w>ve&ccaron;</w>
<w>kot</w> <w>&scaron;est</w>

```

```

<w>mesecev</w> <w>in</w> <w>se</w>
<w>za&ccaron;nejo</w> <w>takoj</w>
<w>uporabljati</w> <c>.</c> </seg>
</tu>

```

Figure 8: Normalised Translation Unit

7.2 Future Work

The real usefulness of these results naturally depends on the purpose we intend to use them for. In our case, the primary aim was a research of the terminological inventory of the texts, especially in a bilingual context for translation oriented terminography. However, some words definitely cannot be terms (e.g. locutions) and it is highly unlikely that a multiword term would begin or end in a preposition or an article (though it can by all means occur in the middle, as in <w>Ministry</w> <w>of</w> <w>Foreign</w> <w>Affairs</w>). A filtering stage seems to be inevitable for the sake of our purpose. In order to filter raw MWU lists we first obtained a stoplist for both languages and a list of single-word terms for each of the 15 texts in the corpus.[15] The former was done simply on the basis of a corpus word frequency list, as function words tend to be the most frequent items in any corpus.[16]

The list of single-word terms was produced on the basis of lists of keywords for each text, which were obtained by matching a word frequency list of a single text against a reference frequency list of a whole corpus. Extracted were only words whose relative frequency in the selected text was higher than the overall frequency. These words represent the core vocabulary of the text and its domain, and in technical/legal texts such as the ones in our corpus a great majority of these words are in fact terms, either independent single-word terms or parts of (yet unknown) multiword terms.

The stopword filter therefore removed stopwords where they occurred either at the beginning or at the end of a MWU, and if only a single word was left, this was excluded too. This reduced the initial size of the raw MWUs

¹⁵ Multiword terms are likely to contain words that could be regarded as single-word terms, therefore the more single-word terms a MWU contains, the more likely it is a term.

¹⁶ Certain less frequent (inflected) forms of function words were added manually.

by approximately 20%. The second step (i.e. the term filter) selected terminologically relevant MWUs according to the following rules: a bigram must contain at least one single-word term in order to be selected, a 3-gram at least two and a 4-gram also at least two terms. 5- and 6-grams were treated separately because they conform less well to such generalizations. This procedure reduced the initial MWU list to only about 10% of its original size, but the units obtained in this way are indeed terminologically relevant. A sample of the final results can be seen in the Table (5).

The results from the filtering stage have not been integrated yet in the overall corpus as we first need to take into account the non-contiguous multiword lexical units that allow under certain conditions the extraction of hapaxes (i.e. terms that occur only once in the overall corpus). This topic is out of the purpose of this paper but the reader will be able to find some details in (Dias, 2000b).

capital formation	CEFTA countries
capital intensity	EU countries
capital market	GDP growth
capital markets	labour productivity
cash benefits	all sectors
commercial banks	annual growth
coomercial firms	applicant countries
After rising	automotive gasoline
average wage	available data
base wage	budget deficit

Table 5: List of filtered terms

8 Conclusion

We hardly believe that the extraction of implicit knowledge (knowledge of the language) such as pp-attachment and multiword lexical units will enable more precise text processing and as a consequence will lead to an adequate normalisation of texts in order to extract more explicit information (knowledge of the world). In this paper, we have presented a new statistically based system called SENTA (Software for the Extraction of N-ary Textual Associations) that retrieves, from naturally occurring texts, relevant 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 require enticement techniques. SENTA has finally been applied to normalise the IJS-ELAN Slovene-English parallel corpus leading to encouraging results for the specific purpose of multilingual terminology extraction.

References

- Bhattacharyya, G. & Johnson, R. (1977). *Statistical Concepts and Methods*. New York, John Wiley & Sons.
- Bourigault D. (1996) *Lexter, a Natural Language Processing Tool for Terminology Extraction*, In "Proceedings of the 7th EURALEX International Congress".
- Chengxiang Z. (1997) *Exploiting Context to Identify Lexical Atoms: a Statistical View of Linguistic Context*, cmp-lg/9701001, 2 Jan 1997.
- Church K.W. & Hanks P. (1990) *Word Association Norms Mutual Information and Lexicography*, *Computational Linguistics*, 16/1, pp. 23--29.
- Dagan I. (1994) *Termight: Identifying and Translating Technical Terminology*, In "Proceedings of the 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.
- David, S. & Plante, P. (1990) *Termino Version 1.0. Research Report of Centre d'Analyse de Textes par Ordinateur*. Université du Québec. Montréal.
- Dias G., Guilloire S. & Lopes J.G.P. (1999a). *Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text corpora*. In "Proceedings of Traitement Automatique des Langues Naturelles", Institut d'Etudes Scientifiques, Cargèse, France.
- Dias G., Guilloire S. & Lopes J.G.P. (2000a) *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., Guilloire S. & Lopes J.G.P. (2000b) *Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?*, In "Proceedings of Recherche

- d'Informations Assistée par Ordinateur (RIO2000)", Collège de France, Paris, France.
- Dunning T. (1993) *Accurate Methods for the Statistics of Surprise and Coincidence*, Association for Computational Linguistics, 19/1.
- Enguehard C. (1993) *Acquisition de Terminologie à partir de Gros Corpus*, In "Proceedings of Informatique & Langue Naturelle ILN'93", pp. 373--384
- Erjavec T. (1999) *A TEI Encoding of Aligned Corpora as Translation Memories*, In "Proceedings of the EACL-99 Workshop on Linguistically Interpreted Corpora (LINC-99)", Bergen. ACL.
- Feldman R. (1998) *Text Mining at the Term Level*, In "Proceedings of PKDD'98", Lecture Notes in AI 1510, Springer Verlag.
- Gale, W. (1991) *Concordances for Parallel Texts*, In "Proceedings of the Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora" Oxford, England.
- Habert B. (1997) *Les linguistiques du Corpus*, Armand Colin, Paris, France.
- Heid U. (1999) *Extracting Terminologically Relevant Collocations from German Technical Texts*, In "Proceedings of TKE'99".
- Justeson J. (1993) *Technical Terminology: Some Linguistic Properties and an Algorithm for Identification in Text*, IBM Research Report, RC 18906 (82591) 5/18/93.
- Silva J., Dias G., Guilloré S. & Lopes J.G.P. (1999) Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units, In "Proceedings of 9th Portuguese Conference in Artificial Intelligence". Springer-Verlag.
- Sinclair J. (1974). *English Lexical Collocations: A study in computational linguistics*. Singapore, reprinted as chapter 2 of Foley, J. A. (ed). (1996), J. M. Sinclair on *Lexis and Lexicography*, Uni Press.
- Shimohata S. (1997) *Retrieving Collocations by Co-occurrences and Word Order Constraints*, In "Proceedings of ACL-EACL'97", pp. 476—481.
- Smadja F. (1993) *Retrieving Collocations From Text: XTRACT*, Computational Linguistics, 19/1, pp. 143—177.
- Smadja F. (1996) *Translating Collocations for Bilingual Lexicons: A Statistical Approach*, Association for Computational Linguistics, 22/1.
- Vintar S. (1999) *A lexical Analysis of the ELAN Slovene-English Corpus*, In "Proceedings of the Workshop on Language Technologies - Multilingual Aspects", Ljubljana, Slovenia, University of Ljubljana.