Normalization of Association Measures for Multiword Lexical Unit Extraction

Gaël Dias
Universidade da Beira Interior
Departamento de Matemática e Informática
6200-001 Covilhã, Portugal
ddg@noe.ubi.pt

Sylvie Guilloré
Université d'Orléans
LIFO
45061, Orléans, France
sylvie.guillore@lifo.univ-orleans.fr

José Gabriel Pereira Lopes
Universidade Nova de Lisboa
FCT/DI
2825-114, Caparica, Portugal
gpl@di.fct.unl.pt

Abstract

The acquisition of Multiword Lexical Units (MWUs) has long been a significant problem in Natural Language Processing. The access to large-scale text corpora has recently originated a new interest in phraseology allowing testing assumptions made about word flexibility constraints. For that purpose, many statistical measures have been proposed in the literature. However, most of them do not accomodate the MWU length factor and so can only evaluate binary word associations. In order to overcome the lack of generalization for n individual words, we propose the normalization of four well-known mathematical models and combine each one with a new acquisition process based on local maxima. In order to face unsatisfactory results obtained with the previous normalized measures, we introduce a new association measure based on the normalized expectation, the Mutual Expectation.

1. Introduction

The acquisition of Multiword Lexical Units (MWUs) has long been a significant problem in Natural Language Processing, being relegated to the borders of lexicographic treatment. Most of the work in knowledge acquisition has aimed at extracting explicit information from texts (i.e. knowledge about the world) and has generally neglected the extraction of implicit information (i.e. knowledge about the language). For the past ten years, there has been a renewal in phraseology mostly stimulated by full access to large-scale text corpora in machine-readable format. The evolution from formalisms towards lexicalization[1] has lead to propose the hypothesis that the more a sequence of words is fixed, that is the less it accepts lexical and syntactical transformations, the more likely it will be a MWU. Compound nouns (*Prime minister*), compound names (*Republic of Yugoslavia*), compound determinants (*a number of*), verbal locutions (*to give rise*), adverbial locutions (*as soon as possible*), prepositional locutions (*such as*) and conjunctive locutions (*on the other hand*) share the properties of MWUs[2][3]. In order to test the assumptions made about word flexibility constraints inherent to MWUs, a great deal of statistical measures have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness that exists within 2-grams (i.e. groups of two words) and do not deal with the general case of n-grams (i.e. groups of n words, with $n \geq 2$). As a consequence, these mathematical models only allow the acquisition of binary associations and enticement techniques[4] have to be applied to acquire associations with more than two words [13] [17] [18] [14] [2]. 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. In order to overcome the lack of generalization for the case of n individual words, we propose the normalization

[1] i.e. The evolution from "general" rules towards rules specifying the usage of words on a case-by-case basis.

[2] This classification is proposed by [12].

[3] For explanatory purposes, we'll access the non-contiguous MWUs further in the paper.

[4] 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).

of four well-known mathematical models (the Dice coefficient [18], the specific mutual information [1], the ϕ^2 [11] and the Log-likelihood ratio [8]) and combine each one with a new acquisition process based on local maxima of association measure values, the LocalMaxs algorithm [15]. However, in order to face unsatisfactory results obtained with the previous normalized measures, we have to introduce a new association measure based on the normalized expectation, the Mutual Expectation [3]. As a consequence, its combination with the LocalMaxs algortithm provides a new solution for the acquisition of n-ary word associations that avoids the definition of global thresholds and does not require entice-ment techniques.

2. Data Preparation

A great deal of applied works in lexicography evidence that most of the lexical relations associate words separated by at most five other words [16]. But a MWU is a specific lexical relation and so can be defined in terms of structure as a specific word n-gram calculated in an immediate span of five words to the left hand side and five words to the right hand side of a focus word.

```
All the European countries forced the
Republic of Yugoslavia to accept
drastic economical sanctions.
```

Figure 1. Sample sentence

As an example, if Figure 1 is the current input and *Republic* (w_1) is the focus word, the set of all the word n-grams can be calculated in a span that starts at the first determinant *the* and ends at the word *drastic*, as illustrated in the following Figure.

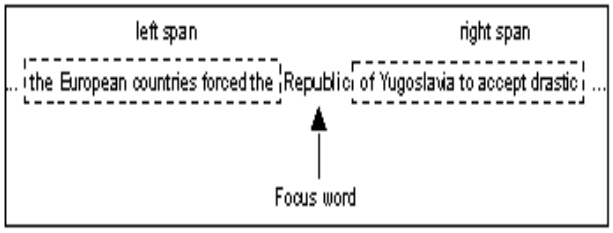

Figure 2. Context around the focus word

Two possible 3-grams are shown in Table 1, being the second one a typical MWU.

w_1	Pos_{12}	w_2	Pos_{13}	w_3
$Republic$	-2	$forced$	$+5$	$drastic$
$Republic$	$+1$	of	$+2$	$Yugoslavia$

Table 1. Sample 3-grams from Figure 1

By definition, an n-gram is a vector of n words where each word is indexed by the signed distance that separates it from its associated focus word. Consequently, an n-gram can be contiguous or non-contiguous whether the words involved in the n-gram represent or not a continuous sequence in the corpus. By convention, the focus word is always the first element of the vector and its signed distance is equivalent to zero. We represent an n-gram by the vector $[w_1 p_{12} w_2 ... p_{1i} w_i ... p_{1n} w_n]$ where p_{1i} (for i=2 to n) denotes the signed distance that separates the word, w_i, from the focus word, w_1.
As notation is concerned, we may characterize an n-gram either 1) by its generic notation or 2) by the sequence of its constituents as they appear in the corpus. In the latter case, each interruption of a non-contiguous n-gram is identified by a gap ("–") that represents the set of all the occurrences that fulfill the free space in the text corpus. Table 2 illustrates the alternative notation for the sample 3-grams presented in Table 1.

Alternative Notation
$forced\ -\ Republic\ -\ -\ -\ -\ drastic$
$Republic\ of\ Yugoslavia$

Table 2. Alternative Notation

As computation is concerned, each word is successively a focus word 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 mathematical models that will evaluate its degree of cohesiveness.

3. Normalized Association Measures

In order to evaluate the degree of cohesiveness existing between words contained in an n-gram, various mathematical models adopted from the Theory of Information or based on the statistical analysis of contingency tables have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness within 2-grams and do not generalize for the case of n individual words. In this section, we propose the normalization of four well-known mathematical models in order to accommodate the n-gram length

factor: the Dice coefficient [18], the specific mutual information [1], the ϕ^2 [11] and the Log-likelihood ratio [8]. In order to face unsatisfactory results[5] obtained with the previous normalized measures, we introduce a new association measure called Mutual Expectation [3] that is the based on the normalized expectation.

3.1. Mutual Expectation

By definition, MWUs are groups of words that occur together more often than expected by chance. From this assumption, we define a new mathematical model, the Mutual Expectation (*ME*), based on the concept of Normalized Expectation (*NE*).

Normalized Expectation We define the *NE* of an n-gram as the average expectation of occurring one word in a given position knowing the occurrence of the other $n-1$ words also constrained by their positions. For example, the average expectation of the 3-gram $[Republic +1\ of +2\ Yugoslavia]$ must take into account all the expectations presented in Table 3.

Expectation of	Knowing the gapped 3-gram
$Republic$	$[\ —\ +1\ of\ +2\ Yugoslavia\ ]$
of	$[\ Republic\ +1\ —\ +2\ Yugoslavia\ ]$
$Yugoslavia$	$[\ Republic\ +1\ of\ +2\ —\ ]$

Table 3. Expectations and *NE*

The underlying concept of the *NE* is based on the conditional probability defined as follows in Equation 1.

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

However, in order to capture in one measure the n conditional probabilities associated to the n events obtained by extracting one word at a time from the n-gram, we need to introduce the concept of the fair point of expectation (*FPE*).

$$FPE\left([w_1 p_{12} w_2 \cdots p_{1i} w_i \cdots p_{1n} w_n]\right) = \frac{1}{n}\left(\sum_{i1=1}^{2}\sum_{i2=i1+1}^{3}\cdots\sum_{\substack{i(n-1)=\\ i(n-2)+1}}^{n} p\left(\left[\begin{array}{c} w_{i1} p_{i1i2} w_{i2} \cdots \\ p_{i1i(n-1)} w_{i(n-1)} \end{array}\right]\right)\right) \tag{2}$$

We know that only the n denominators of the n conditional probabilities vary while the n numerators remain unchanged from one probability to another. So, in order to perform the normalization process, we evaluate the gravity center of the denominators thus defining an average event, the *FPE*. Basically, the *FPE* is the arithmetic mean of the n joint probabilities of the sub-$(n-1)$-grams contained in an n-gram and is defined for each n-gram as in Equation 2.

[5] We will access the comparative results later in the article.

Hence, the normalization of the conditional probability, is realized 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 as defined in Equation 3.

$$NE\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \frac{p\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)}{FPE\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)} \tag{3}$$

Mutual Expectation Daille in [2] shows that one effective criterion for multiword lexical unit identification is frequency. From this assumption, we deduce that between two n-grams with the same *NE*, the most frequent n-gram is more likely to be a MWU.

$$ME\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \begin{array}{l} p\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) \times \\ NE\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) \end{array} \tag{4}$$

So, the Mutual Expectation between n words is defined in Equation 4 based on the *NE* and the relative frequency.

3.2. Association measures from Information Theory

Dice Coefficient The Dice coefficient has been formulated by Dice [7] and introduced by Smadja [18] for the extraction of translation equivalents.

$$Dice\left([w_1 p_{12} w_2]\right) \quad = \quad \frac{2 \times f\left([w_1 p_{12} w_2]\right)}{f\left([w_1]\right) + f\left([w_2]\right)} \tag{5}$$

It measures the cohesiveness that stands between two words of a 2-gram, $[w_1\ p_{12}\ w_2]$, as defined in Equation 5 where $f\left([w_1 p_{12} w_2]\right)$, $f\left([w_1]\right)$ and $f\left([w_2]\right)$ are the respective frequencies of the 2-gram $[w_1\ p_{12}\ w_2]$ and the 1-grams $[w_1]$ and $[w_2]$.

Specific Mutual Information The specific mutual information, based on the mutual information [10], has been introduced by Church and Hanks [1] for the extraction of collocations.

$$MI\left([w_1 p_{12} w_2]\right) \quad = \quad \log_2 \frac{N \times f\left([w_1 p_{12} w_2]\right)}{f\left([w_1]\right) \times f\left([w_2]\right)} \tag{6}$$

It measures the cohesiveness within two words of a 2-gram as defined in Equation 6 where N is the total number of words in the corpus.

Normalization process The normalization of both measures needs to take into account all the possible combinations of dividing an n-gram, $[w_1 ... p_{1i} w_i ... p_{1n} w_n]$, into two complementary sub-groups of words. Indeed, the denominators of both measures point at the division of a 2-gram, $[w_1\ p_{12}\ w_2]$, into two complementary sub-(1-grams), $[w_1]$ and $[w_2]$. Correspondingly to the *ME*, the normalization will be realized by the introduction of an average event called the Fair Point of Dispersion (*FPD*) into the general definitions of the association measures. Let's first introduce

the dispersion frontier point (*DFP*) that represents a symbolic border that divides an n-gram into two complementary sub-groups of words. The *DFP* may take values from 1 to *E(n/2)*[6] as an n-gram can be divided into *E(n/2)* pairs of complementary sub-groups as shown in Table 4.

DFP value	# Words 1^{st} sub-group	# Words 2^{nd} sub-group
1	1	$n-1$
2	2	$n-2$
...	...	...
$E(n/2)$	$E(n/2)$	$n-E(n/2)$

Table 4. Division of an n-gram into 2 complementary sub-groups

Moreover, for each value of the *DFP* there exists a combination of complementary sub-groups of words. For example, when *DFP*=1, there are n possible complementary sub-groups as there are n possible sub-groups of one word in an n-gram and correspondingly n possible sub-groups containing $n-1$ words (See Table 5).

	1^{st} sub-group	2^{nd} sub-group
1	$[w_1]$	$[w_2...p_{2i}w_i...p_{2n}w_n]$
2	$[w_2]$	$[w_1p_{13}w_3...p_{1i}w_i...p_{1n}w_n]$
...	...	...
n	$[w_n]$	$[w_1p_{12}w_2...p_{1i}w_i...p_{1(n-1)}w_{(n-1)}]$

Table 5. All complementary sub-groups of an n-gram for *DFP*=1

#	Event
1	$f([w_1]) \times f([w_2...p_{2i}w_i...p_{2n}w_n])$
2	$f([w_2]) \times f([w_1p_{13}w_3...p_{1i}w_i...p_{1n}w_n])$
...	...
n	$f([w_n]) \times f\left([w_1p_{12}w_2...p_{1i}w_i...p_{1(n-1)}w_{(n-1)}]\right)$

Table 6. All events of an n-gram for DFP=1

So, for both measures, if we define an event as a particular denominator[7], the *FPD* for a generic n-gram $[w_1...p_{1i}w_i...p_{1n}w_n]$ will be the arithmetic mean of all the possible events involved in an n-gram. For example, the respective events for the Dice coefficient associated to Table 5 are illustrated in Table 6. All the necessary equations to calculate the *FPDs* are presented in Appendix A.

Normalized measures The normalized measures are obtained by the introduction of the respective fair points of dispersion into the general definitions.

$$Normalized\_Dice([w_1...p_{1i}w_i...p_{1n}w_n]) = \frac{2 \times f([w_1...p_{1i}w_i...p_{1n}w_n])}{FPD_{Dice}([w_1...p_{1i}w_i...p_{1n}w_n])} \quad (7)$$

$$Normalized\_MI([w_1...p_{1i}w_i...p_{1n}w_n]) = \log_2 \frac{N \times f([w_1...p_{1i}w_i...p_{1n}w_n])}{FPD_{MI}([w_1...p_{1i}w_i...p_{1n}w_n])} \quad (8)$$

The normalized Dice coefficient and the normalized specific mutual information are respectively defined in Equation 7 and Equation 8 where $f([w_1...p_{1i}w_i...p_{1n}w_n])$ is the frequency of the generic n-gram $[w_1...p_{1i}w_i...p_{1n}w_n]$, N the number of words in the corpus and, $FPD_{Dice}([w_1...p_{1i}w_i...p_{1n}w_n])$ the fair point of dispersion for the Dice coefficient and $FPD_{MI}([w_1...p_{1i}w_i...p_{1n}w_n])$ the fair point of dispersion for the specific mutual information.

3.3. Association measures based on the statistical analysis of contingency tables

Contingency tables In order to investigate the relationships between words, an n-dimension contingency table is built for each n-gram providing a convenient display of the data for analysis. For comprehension purposes, we only detail the case where $n = 2$ that involves the definition of a two-dimension contingency table for each 2-gram. A contingency table is defined as in Table 7 for each 2-gram, $[w_1p_{12}w_2]$, where N is the number of words in the input text, f the frequency function and $\overline{arg}$ mentions the absence of the argument.

	w_2	$\overline{w_2}$	Total
w_1	$f([w_1p_{12}w_2])$	$f([w_1p_{12}\overline{w_2}])$	$f([w_1])$
$\overline{w_1}$	$f([\overline{w_1}p_{12}w_2])$	$f([\overline{w_1}p_{12}\overline{w_2}])$	$f([\overline{w_1}])$
Total	$f([w_2])$	$f([\overline{w_2}])$	N

Table 7. Contingency Table

ϕ^2 The ϕ^2 is based on the Pearson's χ^2 test for 2×2 contingency tables and is concerned with testing the null hypothesis that two random variables are independent. It has been introduced by Gale [11] in the context of concordances in parallel texts. The null hypothesis of statistical independence is commonly stated by H_0: $p(w_ip_{ij}w_j) =$

[6] *E(n/2)* returns the integer part of the quotient $(\frac{n}{2})$.

[7] For the case of 2-grams, $f([w_1]) + f([w_2])$ is an event for the Dice coefficient and $f([w_1]) \times f([w_2])$ is an event for the specific Mutual Information.

$p(w_i) \times p(w_j)$. So, if ϕ^2 is minimum then the null hypothesis H_0 is true and the discrete random variables (or words) under study are independent otherwise it may be stated that the two discrete random variables are highly related, with a certain degree of freedom. The ϕ^2 is defined in Equation 9.

$$\phi^2\left([w_1 p_{12} w_2]\right) = \frac{(N \times f([w_1 p_{12} w_2]) - f([w_1]) \times f([w_2]))^2}{f([w_1]) \times (N - f([w_1])) \times f([w_2]) \times (N - f([w_2]))} \quad (9)$$

Log-Likelihood Ratio The Log-likelihood ratio has been introduced by Dunning [8] for the extraction of collocations and is concerned with testing the null hypothesis that two random variables are independent. The null hypothesis of statistical independence is stated by H_0: $p(w_i p_{ij}|w_j) = p(w_i p_{ij}|\overline{w_j})$ thus setting the independence paradigm between two rows of the contingency table defined in Table 7. The Log-likelihood ratio is defined in Equation 10.

$$Loglike\left([w_1 p_{12} w_2]\right) = -2\log\lambda = \quad (10)$$
$$2 \times (\ \log\theta_1^{s1}(1-\theta_1)^{n1-s1} + \log\theta_2^{s2}(1-\theta_2)^{n2-s2} - \log\theta^{s1}(1-\theta)^{n1-s1} - \log\theta^{s2}(1-\theta)^{n2-s2})$$

where

$s1 = f\left([w_1 p_{12} w_2]\right)$ $\quad s2 = f\left([w_2]\right) - f\left([w_1 p_{12} w_2]\right)$
$n1 = f\left([w_1]\right)$ $\quad n2 = N - f\left([w_1]\right)$
$\theta_1 = \frac{s1}{n1}$ $\quad \theta_2 = \frac{s2}{n2}$
$\theta = \frac{f([w_2])}{N}$

Normalization process We propose the normalization of the ϕ^2 and the Log-likelihood ratio based on two particular fair points of dispersion, the *FPD-left* and the *FPD-right*. Indeed, the structure of the contingency table suggests that an n-gram should be divided into a left hand-side (for the w_1 part of the contingency table) and a right hand-side (for the w_2 part of the contingency table). We define the left hand-side of an n-gram as the set all the sub-groups of words that contain from 1 to *E(n/2)* words of the n-gram and the right hand-side of an n-gram as the set all the sub-groups of words that contain from *n-1* to *n-E(n/2)* words of the n-gram as illustrated in Table 8.

Left Hand Side	Right Hand Side
1	$n-1$
2	$n-2$
...	...
$E(n/2)$	$n-E(n/2)$

Table 8. # of words in each sub-group of an n-gram

So, for both measures, if we define an event as the frequency of one particular sub-group of an n-gram, the *FPD-left* for a n-gram $[w_1 ... p_{1i} w_i ... p_{1n} w_n]$ will be the arithmetic mean of all the possible events of its left part and the *FPD-right* for the same n-gram will be the arithmetic mean of all the possible events of its right part. All the expressions to calculate the particular fair points of dispersion are presented in Appendix B.

Normalized measures The normalized ϕ^2 and the normalized Log-likelihood ratio are respectively defined in Equation 11 and Equation 12.

$$Normalized\_\phi^2\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \frac{(N \times f([w_1 ... p_{1i} w_i ... p_{1n} w_n]) - x \times y)^2}{x \times (N-x) \times y \times (N-y)} \quad (11)$$

where

$x \equiv FPD\_left\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)$
$y \equiv FPD\_right\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)$

$$Normalized\_Loglike\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \quad (12)$$
$$2 \times (\ \log\theta_1^{s1}(1-\theta_1)^{n1-s1} + \log\theta_2^{s2}(1-\theta_2)^{n2-s2} - \log\theta^{s1}(1-\theta)^{n1-s1} - \log\theta^{s2}(1-\theta)^{n2-s2})$$

where

$s1 \equiv f\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)$
$s2 \equiv FPD\_right\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) - f\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)$
$n1 \equiv FPD\_left\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)$
$n2 \equiv N - FPD\_left\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right)$
$\theta_1 \equiv \frac{s1}{n1}$
$\theta_2 \equiv \frac{s2}{n2}$
$\theta \equiv \frac{FPD\_right([w_1 ... p_{1i} w_i ... p_{1n} w_n])}{N}$

At this stage, each n-gram can be associated to one association measure value. The following step consists in extracting MWU candidates from the set of all the n-grams associated to their cohesiveness value.

4. The Acquisition Process

Most of the approaches proposed in the literature base their selection process on global association measure thresholds [1] [2] [17] [18] [13] [14] [9]. This is defined by the underlying concept that there exits a limit value of the association measure that allows to decide whether an n-gram is a MWU 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 retuned when the type, the size, the domain and the language of the document change. The LocalMaxs algorithm [15]

proposes a more robust, flexible and fine-tuned approach for the election of MWUs as it focuses on the identification of local maxima of the association measure values.
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:

$\forall x \in \Omega_{n-1}, \forall y \in \Omega_{n+1} \quad W$ is a MWU if
$(sizeof(W) = 2 \wedge assoc(W) > assoc(y)) \vee$
$(sizeof(W) \neq 2 \wedge assoc(W) \geq assoc(x) \wedge$
$assoc(W) > assoc(y))$

Table 9. LocalMaxs Algorithm

The LocalMaxs algorithm proposes a theoretically sound acquisition process that does not depend on experimentation and avoids the definition of global thresholds. As a consequence, it overcomes the problems of portability of the existing approaches. Indeed, no tuning is needed in order to run the system and any association measure can be tested. For the purpose of our study, we applied the LocalMaxs algorithm to the Mutual Expectation, the normalized Dice coefficient, the normalized specific mutual information, the normalized ϕ^2 and the normalized Log-likelihood ratio.

5. Evaluation of the Results

In this section, we compare the results obtained by applying the LocalMaxs algorithm with the five association measures mentioned so far in the paper, over an English corpus of political debates with approximately 300000 words [8]. The results illustrate that the Mutual Expectation leads to very much improved results for the specific task of multiword lexical unit extraction and show that it is possible to extract precisely compound nouns, names and determinants as well as verbal, adverbial, adjectival, conjunctive and prepositional locutions (See Table 10 and Table 11 [9]).
There is no consensus among the research community about how to evaluate the output of multiword lexical unit extraction systems. Indeed, the quality of the output strongly depends on the task being tackled, as a lexicographer or a translator may not evaluate the same results in the same manner. A precision measure (*P*) should surely be calculated in relation with a particular task. However, in order to define some "general" rule to measure the precision of the system, we propose the following two assumptions. Multiword units are valid units if they are grammatically appropriate units (by grammatically appropriate units we refer to compound determinants/nouns/names and verbal/prepositional/adverbial/conjunctive locutions) or if they are meaningful units even though they are not grammatical.

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

[9]The numbers are expressed within a 10^{-3} scale and "F" stands for Frequency.

ME	F	MWUs
1.04	3	Peace Accord
1.04	3	Court of Justice
1.31	4	to fall within the competence
1.74	5	as soon as possible
1.01	2	bearing in mind
1.00	2	nuclear warhead design calculations

Table 10. Elected contiguous MWUs

ME	F	MWUs
1.04	2	to allow — — to {foreign observers, the IRC}
1.00	2	Article — of the Council Directive {12, 15}
1.00	2	the — on Climate Change {Conventions, Decisions}

Table 11. Elected non-contiguous MWUs

$$P = \frac{\#\ correct\ MWUs}{\#\ extracted\ MWUs} \tag{13}$$

$$ER = \frac{\#\ correct\ MWUs}{Size\ of\ the\ corpus} \tag{14}$$

Besides, the evaluation of extraction systems is usually performed with the well-known recall rate. However, we do not present the "classical" recall rate in this experiment due to the lack of a reference corpus where all MWUs would be identified. Instead, we present the extraction rate (*ER*), a measure of coverage, defined as the percentage of well-extracted MWUs in relation with the size of the corpus (by well-extracted we mean that the extracted MWUs are precise according to the definition of precision). The global results reveal that the Mutual Expectation exhibits significant progresses in terms of Precision comparing to all the other measures as illustrated in Table 12.

	ME	$Dice$	MI	ϕ^2	$Loglike$
P (%)	90.35	49.33	59.31	73.22	48.31
ER (%)	1.71	1.50	0.91	0.93	3.05

Table 12. Comparative results of *P* and *ER*

One of the most important points that we can express

against the four other normalized models is that they raise the typical problem of high frequency words as they highly depend on marginal probabilities. Indeed, they underestimate the degree of cohesiveness when the marginal probability of one word is high. For instance, the Dice coefficient, the Log-likelihood ratio and the ϕ^2 elect the following 2-gram $Turkish\ -\ Kurdish$ as the most significant expression in the overall corpus. However, the probability that the conjunction *and* fills in the gap is one. In fact, the 3-gram [$Turkish\ +1\ and\ +2\ Kurdish$] gets unjustifiably a lower value of cohesiveness than the 2-gram [$Turkish\ +2\ Kurdish$]. Indeed, the high frequency of the conjunction *and* underestimates the cohesiveness value of the 3-gram. On the opposite, applied with the Mutual Expectation, the LocalMaxs algorithm elects the longest and most frequent MWU that contains both words *Turkish* and *Kurdish* that is *Turkish and Kurdish political refugees*. In order to assess the results obtained with the ME, we present in Table 13 the output of the concordancer when *Turkish* and *Kurdish* are searched in the overall text corpus separated by just one word.

v e n	Turkish and Kurdish political refugees	b e i
e b y	Turkish and Kurdish political refugees	i m p
v e n	Turkish and Kurdish political refugees	i n G

Table 13. Concordances for *Turkish* $+2$ *Kurdish*

The same statement as for the Dice coefficient, the Log-likelihood ratio and the ϕ^2 ratio can be extended to the specific mutual information that elects the 2-gram *Code - Practice* instead of the well-formed expression *Code of practice* extracted with the Mutual Expectation.

The results presented in Table 12 allow the classification of the association measures on a general basis. However, in order to characterize precisely each mathematical model, we propose more detailed figures in terms of length and frequency of the extracted MWUs. The results presented in Figure 3 clearly reveal that most of the MWUs contain between 2 to 4 words although there are differences of distribution between each model.

Another important result illustrated in Figure 4 is the fact that most of the extracted MWUs occur only twice in the corpus. Indeed, all the models, with the exception of the Dice coefficient, elect in a great proportion MWUs that occur only two times in the overall corpus [10].

[10] More figures can be found in [4] [5] [6].

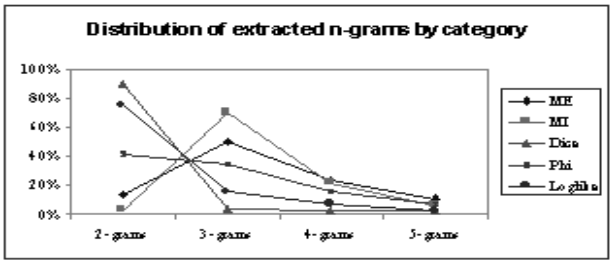

Figure 3. Extracted n-grams by Category

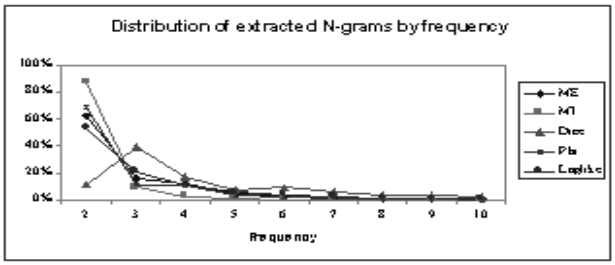

Figure 4. Extracted n-grams by Frequency

6. Conclusion

In order to avoid the use of unsatisfactory enticement techniques for the extraction of n-ary textual associations, we proposed the normalization of five association measures. The results obtained with four well-known mathematical models (Dice coefficient, specific Mutual Information, ϕ^2 and Log-likelihood ratio) lead us to introduce a new association measure, the Mutual Expectation, based on the concept of Normalized Expectation. We also proposed a new acquisition process, the LocalMaxs algorithm, that automatically extracts contiguous and non-contiguous multiword lexical units without relying on empirically defined global thresholds. The system evidences itself by its flexibility, allowing any user to retrieve contiguous and non-contiguous textual associations from texts of all domains and languages without any pre-treatment of the corpus or pre-definition of thresholds. We hardly believe that the success of applications in the areas of Information Extraction, Information Retrieval and Machine Translation will rely on the pre-processing of text corpora in order to benefit from their intrinsic information. The extraction of implicit knowledge such as sub-categorization frames, pp-attachment and multiword lexical units will enable more precise text processing and as a consequence will lead to an adequate normalization of texts in order to extract more explicit information.

7. Appendix A

We present in this appendix the details of the normalization of the Dice coefficient and the specific mutual information. We will treat both cases as a single case as the normalization processes only differ in the operators i.e. $+$ for the Dice coefficient and $\times$ for the specific mutual expectation. As a consequence, we will use the generic notation $\oplus$ to represent both the $+$ and the $\times$ operators.
For each association measure, the respective fair point of dispersion is the quotient between the sum of all the events involved by the n-gram and the number of events, as illustrated in Equation 15.

$$FPD_{(Dice \vee MI)}\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \frac{sum\_events([w_1 ... p_{1i} w_i ... p_{1n} w_n])}{\#\_events([w_1 ... p_{1i} w_i ... p_{1n} w_n])} \quad (15)$$

In order to calculate the number of events of a generic n-gram, two cases have to be distinguished. On one hand, if the n-gram contains an even number of words, the total number of events is the sum of all the combinations of i words among n words for $i = 1, ..., (E(n/2) - 1)$ plus half of the combinations of $E(n/2)$ words among n words. Indeed, for this specific case, the last dispersion frontier point (i.e. $DFP = E(n/2)$) divides the n-gram into two sub-groups with the same size in terms of words. Therefore, the number of events must be reduced to half. On the other hand, if the n-gram contains an odd number of words, the total number of events is the sum of all the combinations of i words among n words for $i = 1, ..., E(n/2)$. So, the $\#\_events$ function is defined in Equation 16.

$$\#\_events\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \begin{cases} even(n), & \begin{cases} if\ n = 2, 1 \\ if\ n > 2, \begin{cases} \sum_{i=1}^{E(\frac{n}{2})-1} \begin{pmatrix} n \\ i \end{pmatrix} + \\ \frac{1}{2} \begin{pmatrix} n \\ E(\frac{n}{2}) \end{pmatrix} \end{cases} \end{cases} \\ odd(n), & \sum_{i=1}^{E(\frac{n}{2})} \begin{pmatrix} n \\ i \end{pmatrix} \end{cases} \quad (16)$$

For the case of the sum_events function, the sum of all the events resulting from the division of an n-gram into two sub-groups is the sum of all the particular sums of events for all the possible DFP values and is defined in Equation 17.

$$sum\_events\left([w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]\right) = \begin{cases} n = 2, spec\_sum\_1(E\left(\frac{n}{2}\right), [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ even(n) \wedge (n > 2), \\ \sum_{DFP=1}^{E\left(\frac{n}{2}\right)-1} spec\_sum\_1(DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) + \\ \quad spec\_sum\_1(n - DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ + spec\_sum\_1(E\left(\frac{n}{2}\right), [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ odd(n), \\ \sum_{DFP=1}^{E\left(\frac{n}{2}\right)} spec\_sum\_1(DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) + \\ \quad spec\_sum\_1(n - DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \end{cases} \quad (17)$$

Finally, the $spec\_sum\_1$ function is introduced and defined in Equation 18 where $\hat{w_{ij}}$ corresponds to an omitted word of a given succession indexed from 1 to n.

$$spec\_sum\_1\left(k, [w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \sum_{i1=2}^{j} \sum_{i2=i1+1}^{j+1} \cdots \sum_{ik=i(k-1)+1}^{n} f\left([w_{i1} p_{i1i2} w_{i2} ... p_{i1ik} w_{ik}]\right) \oplus f\left([w_1 ... \hat{w_{i1}} ... \hat{w_{ik}} ... p_{1n} w_n]\right) \quad (18)$$

where $j = n - k + 1$

8. Appendix B

We present in this appendix the details of the normalization of the ϕ^2 and the Log-likelihood ratio. In both case, two fair points of dispersion are introduced, i.e. one for the left hand side and one for the right hand side of an n-gram. Both FPDs are respectively defined in Equation 19 and Equation 20.

$$FPD\_left\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \frac{sum\_events\_left([w_1 ... p_{1i} w_i ... p_{1n} w_n])}{\#\_events([w_1 ... p_{1i} w_i ... p_{1n} w_n])} \quad (19)$$

$$FPD\_right\left([w_1 ... p_{1i} w_i ... p_{1n} w_n]\right) = \frac{sum\_events\_right([w_1 ... p_{1i} w_i ... p_{1n} w_n])}{\#\_events([w_1 ... p_{1i} w_i ... p_{1n} w_n])} \quad (20)$$

We want to notice the reader that in both cases, the $\#\_events$ function is defined in as in Equation 16.
In order to calculate the fair point of dispersion of the left hand-side and the right hand-side of an n-gram, the sum_events_left and sum_events_right functions are defined in Equation 21 and Equation 22. For the specific case of an "even" n-gram, we rule that the left hand-side of the n-gram will contain only the sub-groups of $(\frac{n}{2})$ words that contain w_1 and the right hand-side all the other sub-groups. Indeed, an n-gram that contains an even number of words can be divided into two equal parts in terms of word number.

$$sum\_events\_left\left([w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]\right) = \begin{cases} n = 2, spec\_sum\_left(E\left(\frac{n}{2}\right), [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ even(n) \wedge (n > 2), \\ \sum_{DFP=1}^{E\left(\frac{n}{2}\right)-1} spec\_sum\_2(DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ + spec\_sum\_left(E\left(\frac{n}{2}\right), [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ odd(n), \\ \sum_{DFP=1}^{E\left(\frac{n}{2}\right)} spec\_sum\_2(DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \end{cases} \quad (21)$$

$$sum\_events\_right\left([w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]\right) = \begin{cases} n = 2, spec\_sum\_right(n - E(\frac{n}{2}), [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ even(n) \wedge (n > 2), \\ \sum_{\substack{DFP = \\ n - E(\frac{n}{2}) + 1}}^{n-1} spec\_sum\_2(DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ + spec\_sum\_right(n - E(\frac{n}{2}), [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \\ odd(n), \\ \sum_{DFP=n-E(\frac{n}{2})}^{n-1} spec\_sum\_2(DFP, [w_1 \cdots p_{1i} w_i \cdots p_{1n} w_n]) \end{cases} \quad (22)$$

The $spec\_sum\_2$, $spec\_sum\_left$ and $spec\_sum\_right$ functions are finally respectively defined in Equation 23,

Equation 24 and Equation 25.

$$spec\_sum\_2\,(k,[w_1...p_{1i}w_i...p_{1n}w_n]) = \sum_{i1=1}^{j}\sum_{i2=i1+1}^{j+1}\cdots\sum_{ik=i(k-1)+1}^{n} f\,([w_{i1}p_{i1i2}w_{i2}...p_{i1ik}w_{ik}]) \quad (23)$$

where $j = n - k + 1$

$$spec\_sum\_left\,(k,[w_1...p_{1i}w_i...p_{1n}w_n]) = \sum_{i2=2}^{j+1}\sum_{i3=i2+1}^{j+2}\cdots\sum_{ik=i(k-1)+1}^{n} f\,([w_1p_{1i2}w_{i2}...p_{1ik}w_{ik}]) \quad (24)$$

where $j = n - k + 1$

$$spec\_sum\_right\,(k,[w_1...p_{1i}w_i...p_{1n}w_n]) = \sum_{i1=2}^{j}\sum_{i3=i2+1}^{j+1}\cdots\sum_{ik=i(k-1)+1}^{n} f\,([w_{i1}p_{i1i2}w_{i2}...p_{i1ik}w_{ik}]) \quad (25)$$

where $j = n - k + 1$

9. Acknowledgements

We want to thank the reviewers for their valuable comments and we hope that our efforts to clarify the paper have achieved the results the reviewers and the Program Committee expected. This work has been funded by the PhD grant PRAXIS 4/4.1/BD/3895 and the project "DIXIT - Multilingual Intentional Dialog Systems", Ref. PRAXIS XXI 2/2.1/TIT/1670/95.

References

[1] K. Church and P.Hanks. Word association norms mutual information and lexicography. *Computational Linguistics*, 16(1):23–29, 1990.

[2] B. Daille. Study and implementation of combined techniques for automatic extraction of terminology. *The balancing act combining symbolic and statistical approaches to language*, 1995.

[3] G. Dias, S. Guilloré, and G. Lopes. Language independent automatic acquisition of rigid multiword units from unrestricted text corpora. *Traitement Automatique des Langues Naturelles*, pages 12–17, July 1999.

[4] G. Dias, S. Guilloré, and G. Lopes. Multilingual aspects of multiword lexical units. *Workshop Language Technologies-Multilingual Aspects*, pages 8–11, July 1999.

[5] G. Dias, S. Guilloré, and G. Lopes. Multiword lexical units extraction. *International Symposium on Machine Translation and Computer Language Information Processing*, pages 26–28, June 1999.

[6] G. Dias, S. Guilloré, and G. Lopes. Mutual expectation: a measure for multiword lexical unit extraction. *Venezia per il Trattamento Automatico delle Lingue*, pages 22–24, November 1999.

[7] L. Dice. Measures of the amount of ecologic association between species. *Journal of Ecology*, 1945.

[8] T. Dunning. Accurate methods for the statistics of surprise and coincidence. *Association for Computational Linguistics*, 19(1), 1993.

[9] C. Enguehard. Acquisition de terminologie à partir de gros corpus. *Informatique et Langue Naturelle*, pages 373–384, 1993.

[10] R. Fano. *Transmission of Information: A statistical theory of communications*. MIT Press, MA, 1961.

[11] W. Gale. Concordances for parallel texts. *Proceedings of Seventh Annual Conference of the UW Center for the New OED and Text Research, Using Corpora*, 1991.

[12] G. Gross. *Les expressions figées en franais*. Ophrys, Paris, 1996.

[13] A. Salem. *La Pratique des segments répétés*. Klincksieck, Paris, 1987.

[14] S. Shimohata. Retrieving collocations by co-occurrences and word order constraints. *ACL-EACL'97*, 1997.

[15] J. Silva, G. Dias, S. Guilloré, and G. Lopes. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *9th Portuguese Conference in Artificial Intelligence*, pages 21–24, September 1999.

[16] J. Sinclair. *English Lexical Collocations: A study in computational linguistics. J. M. Sinclair on Lexis and Lexicography*. Uni Press, Singapore, 1974.

[17] F. Smadja. Retrieving collocations from text: Xtract. *Computational Linguistics*, 19(1):143–117, 1993.

[18] F. Smadja. Translating collocations for bilingual lexicons: A statistical approach. *Association for Computational Linguistics*, 22(1), 1996.