Combining Linguistics with Statistics for Multiword Term Extraction: A Fruitful Association?

Gaël Dias[123] & Sylvie Guilloré[2] & Jean-Claude Bassano[2] & José Gabriel Pereira Lopes[3]
[1] Universidade da Beira Interior
Departamento de Matemática e Informática
Rua Marquês de Ávila e Bolama
6200-001 Covilhã, Portugal
ddg@noe.ubi.pt
[2] LIFO-Université d'Orléans
BP 6102 - 45061, Orléans Cédex 2, France
{dias, guillore, bassano}@lifo.univ-orleans.fr
[3] Universidade Nova de Lisboa
Departamento de Informática
Quinta da Torre, 2825-114, Caparica, Portugal
{ddg, gpl}@di.fct.unl.pt

Abstract

The acquisition of multiword terms from large text collections is a fundamental issue in the context of Information Retrieval. Indeed, their identification leads to improvements in the indexing process and allows guiding the user in his search for information. In this paper, we present an original methodology that allows extracting multiword terms by either (1) exclusively considering statistical word regularities or by (2) combining word statistics with endogenously acquired linguistic information. For that purpose, we conjugate a new association measure called the Mutual Expectation with a new acquisition process called the LocalMaxs. On one hand, the Mutual Expectation, based on the concept of Normalised Expectation, evaluates the degree of cohesiveness that links together all the textual units contained in an n-gram (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. Finally, we compare the results obtained by applying the methodology over a raw Portuguese text with the results reached by combining word statistics with linguistic information endogenously acquired from the same corpus previously tagged.

1. Motivations

The acquisition of terminologically relevant multiword lexical units from large text collections is a fundamental issue in the context of Information Retrieval. Indeed, their identification leads to improvements in the indexing process and allows guiding the user in his search for information.

On one hand, selecting discriminating terms in order to represent the contents of texts is a critical problem. Ideally, the indexing terms should directly describe the concepts present in the documents. However, most of the information retrieval systems index the documents of a text collection based on individual words that are not specific enough to evidence the contents of texts. In order to improve the quality of the indexing process, some systems take advantage of pre-existing thesauri. In that case, the discriminating terms are selected from the thesaurus (Betts, 1991). Unfortunately, most of the domains do not contain pre-defined thesauri and very few projects include automatic construction of specialised thesauri (Grefenstette, 1994). In order to overcome the lack of domain specific thesauri, evolutionary retrieval systems use multiword terms previously extracted from text collections to represent the contents of texts (Evans, 1993). Indeed, multiword terms embody meaningful sequences of words that are less ambiguous than single words and allow approximating more accurately the contents of texts.

On the other hand, an information retrieval system should also be able to guide the user in his search for information allowing him to refine his query from a list of relevant terms that semantically match

his initial search. Thus, the user should be able to "visit" the document space from one concept to another keeping trace of his route. The discriminating power of the multiword terms plays a fundamental role in this operation by proposing a more refined view of the contents that are present in the text collection (Quaresma *et al*, 1998).

However, most of the multiword terms are not listed in lexical databases. Indeed, the creation, the maintenance and the upgrade of terminological data banks often require a great deal of manual efforts that can not cope with the ever growing number of text corpora to analyse. Moreover, due to the constant dynamism of specialised languages, the set of multiword terms is opened and to be completed. Indeed, most of the neologisms in technical and scientific domains are realised by multiword terms. For example, *World Wide Web*, *IP address* and *TCP/IP network* are terminologically relevant multiword lexical units that are particularly new in the domain of Computer Science. As a consequence, there has been a growing interest in developing techniques for automatic term extraction. In order to extract multiword terms from text corpora, three main strategies have been proposed in the literature.
First, purely linguistic systems (David, 1990; Dagan, 1993; Bourigault, 1996) propose to extract relevant terms by using techniques that analyse specific syntactical structures in the texts. However, this methodology suffers form its monolingual basis, as the systems require highly specialised linguistic techniques to identify clues that isolate possible candidate terms.
Second, hybrid methodologies (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 as we'll explain further in this paper.
Finally, purely statistical systems (Church & Hanks, 1990; Dunning, 1993; Smadja, 1993; Shimohata, 1997) extract 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 extract 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[1] to acquire multiword terms 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.

In order to overcome the problems previously highlighted by the statistical systems, we conjugate a new association measure called the Mutual Expectation (Dias *et al*, 1999*a*) with a new acquisition process called the LocalMaxs (Silva *et al*, 1999). On one hand, the Mutual Expectation, based on the concept of Normalised Expectation, evaluates the degree of cohesiveness that links together all the textual units contained in an n-gram (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. The combination of the new association measure with the new acquisition process proposes an innovative integrated solution to the problems of enticement techniques and global thresholds defined by experimentation.

As an illustration of our system, we present the results obtained by performing two distinct experiments. First, multiword terms are extracted from a Portuguese raw text by exclusively considering statistical word regularities. In that case, each word n-gram is associated to its Mutual Expectation value and the LocalMaxs extracts all the possible terms from the set of all valued word n-grams. In the second experiment, the system extracts terminologically relevant multiword lexical units

[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.

by combining word statistics with linguistic information that is acquired endogenously from the same Portuguese corpus previously tagged. In that case, each word n-gram is linked to its corresponding tag n-gram and the final association measure value of the word n-gram is the product between its Mutual Expectation value and the Normalised Expectation value of its associated tag n-gram. Finally, the LocalMaxs elects the candidate n-grams by evidencing local maxima of association measure values. The comparative results show that the combination of linguistics with statistics does not necessarily lead to improvements although interesting results are obtained.

2. Data Preparation

We base our methodology on two main principles. First, 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 multiword term. Based on this assumption, we propose that the general information appearing in morpho-syntactically tagged corpora (i.e. words or/and part-of-speech tags) should be sufficient to extract meaningful textual units without applying domain-dependent or language-dependent heuristics. As a consequence, we opted to design a generic methodology that can be applied to any text input, independently of its domain, type or language. We will refer to this first principle as the **rigidity principle**. Second, we propose that the input text corpus should not be modified at all. Indeed, on one hand, it has not been proved, as far as we know, that lemmatisation improves the extraction process although many works focus on its necessity. On the contrary, we believe that lemmatisation is worth introducing noise in some cases. For example, it is obvious that lemmatising the term *United Nations* into *united nation* would not bring benefit to the extraction process. On the other hand, we also believe that pruning texts with lists of stop words should be avoided as this process introduces constraints that are not contained in the original input text. Indeed, it is still not clear for the research community which set of stop words should be used for a given task. To that respect, we may formulate an even more radical belief saying that pruning has been the only solution encountered so far to handle frequent and "meaningless" units in texts. So, our objective has been to design a methodology that does not modify the input text at all (i.e., the text is neither lemmatised, nor pruned with lists of stop words) and work on all the information contained inside a morpho-syntactically tagged corpus. Thus, less external constraints are introduced in the extraction process. We may refer to that principle as the **corpus integrity principle**.

The first step of our methodology performs the transformation of the input text into a set of n-grams. 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[2] and assess that multiword terms are specific lexical relations that share this property (Sinclair, 1974). As a consequence, a multiword term can be defined in terms of structure as a specific word n-gram calculated in the immediate context of three words to the left hand side and three words to the right hand side of a pivot word[3]. This situation is illustrated in Figure (1) for the pivot word *Lei* (-*Law*-) being given the input sentence (1). Indeed, *Lei de Imprensa* (-*Press Law*-) is a specific multiword term.

(1) *O artigo 35 da Lei de Imprensa prevê esse procedimento em caso de burla agravada.*

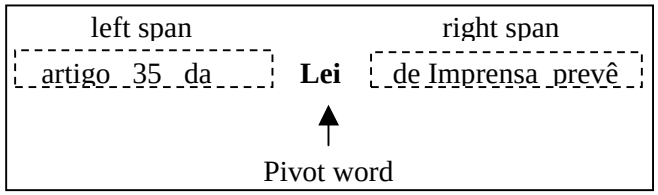

Figure 1: Context Span for the pivot word *Lei*.

[2] O. Mason (1997) suggests that lexical relations involving one word vary in terms of word length. So, ideally there should exist a different span for every word under study.

[3] Different sizes for the context span have been tested in (Dias *et al*, 1999*f*). The best results have been obtained for a context span of three words on each side of the pivot word.

By definition, a word n-gram is a vector of n words where each word is indexed by the signed distance that separates it from its associated pivot 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 of words in the corpus. For instance, if we consider the sentence (1) as the current input text and "*Lei*" the pivot word, a contiguous and a non-contiguous word 3-gram are respectively illustrated in the first two rows of Table (1).

w_1	$position_{12}$[4]	w_2	$position_{13}$	w_3
Lei	+1	*de*	+2	*Imprensa*
Lei	-3	*artigo*	+3	*Prevê*

Table 1: Sample word 3-grams calculated from the pivot word *Lei.*

Generically, an n-gram is a vector of n textual units where each textual unit is indexed by the signed distance that separates it from its associated pivot textual unit. By convention, the pivot textual unit is always the first element of the vector and its signed distance is equivalent to zero. We represent an n-gram by a vector $[p_{11}\ u_1\ p_{12}\ u_2\ p_{13}\ u_3...p_{1i}\ u_i...p_{1n}\ u_n]$ where p_{11} is equal to zero and p_{1i} (for i=2 to n) denotes the signed distance that separates the textual unit u_i from the pivot textual unit u_1.

On the other hand, each word n-gram may be linked to its corresponding tag n-gram[5]. In that case, a tag n-gram is a vector of part-of-speech tags where each tag is indexed by the signed distance that separates it from the part-of-speech tag of its associated pivot word. So, if we consider the tagged sentence (2)[6] as the current input text and "*[N]*" the part-of-speech tag associated to the pivot word "*Lei*", the respective tag 3-grams corresponding to the word 3-grams listed in the Table (1) are respectively illustrated in the first two rows of Table (2).

(2) *O[ART] artigo[N] 35[NUM] da[PREP] Lei[N] de[PREP] Imprensa[N] prevê[V] esse[ART] procedimento[N] em[PREP] caso[N] de[PREP] burla[N] agravada[ADJ]*[7].

t_1	$position_{12}$	t_2	$position_{13}$	t_3
[N]	+1	*[PREP]*	+2	*[N]*
[N]	-3	*[N]*	+3	*[V]*

Table 2: Corresponding tag 3-grams to the word 3-grams of Table (1).

As notation is concerned, we may characterise a contiguous n-gram by the sequence of its constituents as it appears in the corpus (Exp. (1)) or by explicity mentionning the signed distances associated to each textual unit of the n-gram where the distance of the pivot is ommited (Exp. (2)). Similarly, each interruption of a non-contiguous n-gram may be identified in the sequence of its constituents by a gap (i.e. "_____") that represents the set of all the occurrences that fulfil the free space in the text corpus (Exp. (3)). A non-contiguous n-gram can also be represented by the explicit mention of the signed distances associated to its constituents (Exp. (4)).

Lei de Imprensa (1)

[*Lei* +1 *de* +2 *Imprensa*] (2)

artigo _____ _____ *Lei* _____ _____ *prevê* (3)

[*Lei* –3 *artigo* +3 *prevê*] (4)

[4] *Position*$_{12}$ and *position*$_{13}$ are respectively the signed distances between w_1 (= the pivot word) and w_2 and between w_1 and w_3. The sign "+" ("-") is used for words on the right (left) of w_1.

[5] In particular, we will need to consider the tag n-gram associated to each word n-gram of the corpus for the purpose of our second experiment.

[6] Sentence (2) is the result of the tagging process over Sentence (1).

[7] For comprehension purposes, the set of tags is voluntarily simplified.

So, the data preparation stage ends up with the definition of seven tables that are used to store the set of all the n-grams[8] obtained by sequentially processing the input text corpus, each word being successively pivot. Then, we respectively calculate the frequency and the association measure of each unique n-gram. Finally, in the fourth and final step, we apply the LocalMaxs algorithm in order to elect the multiword term candidates from the set of all valued n-grams. In the next section, we will present the Mutual Expectation measure that will be applied to each word n-gram in order to evaluate its degree of cohesiveness. For that purpose, we will introduce the concept of the Normalised Expectation.

3. Normalised Expectation and Mutual Expectation

In order to evaluate the degree of cohesiveness existing between textual units, various mathematical models have been proposed in the literature. However, most of them only evaluate the degree of cohesiveness between two textual units and do not generalise for the case of n individual textual units (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 enticement techniques have to be applied to acquire associations with more than two textual units. Unfortunately, such techniques have shown their limitations as their retrieval results mainly depend on the identification of suitable 2-grams for the initiation of the iterative process. On the other hand, for the specific case of word associations, 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 (ME) that evaluates the degree of rigidity that links together all the textual units contained in an n-gram ($\forall n, n \geq 2$) based on the concept of Normalised Expectation (NE) (Dias *et al*, 1999*a*).

3.1. Statistical Background

By definition, a textual association is characterized by some kind of attraction between its components. In order to investigate this particular relationship between words or part-of-speech tags, an n-dimensions contingency table is built for each n-gram providing a convenient display of the data for analysis. Defining a Probability Space is the preliminary step towards building the contingency tables. For our purpose, the Probability Space ($\Omega, A, P[.]$) where Ω is the Domain space, A the Event space and $P[.]$ the Probability function is introduced in Table (3).

A	The event space A maps to each textual unit u_i a binary discrete random variable X_{ip} that takes the value "1" if the textual unit u_i appears in an n-gram at position[9] p and "0" if not.
Ω	The Domain space Ω is the collection of all possible outcomes of a conceptual experiment over the instance space and is therefore defined as $\Omega=\{0, 1\}$.
$P[.]$	A good approximation for the Probability function $P[.]$ is defined as the number of successes for a particular outcome divided by the number of instances.

Table 3: The Probability Space ($\Omega, A, P[.]$).

The instance space over which the Probability Space can be applied is the set of all n-grams built from the input text. Indeed, each n-gram provides a new independent Bernoulli trial for every variable X_{ip}. For example, if we take the discrete random variable X_{ip} which maps the following word u_i=*Imprensa*

[8] As the n-grams are built in a window of six textual units, excluding the pivot, it is possible to calculate combinations of one to seven words.

[9] The position p is the position of the textual unit u_i in relation with the first textual unit of the n-gram.

and the position p=2, the outcome of the trial for the first *3*-gram of Table (1) is "1" and for the second *3*-gram is "0".

The contingency tables may now be built supported by the Probability Space mentioned above. For comprehension purposes, we only detail the case of the 2-grams involving the definition of a two-dimensions contingency table for each *2*-gram[10].
We can define a 2-gram as being a quadruplet $[p_{11}\, u_1\, p_{12}\, u_2]$ where u_1 and u_2 are two textual units and p_{12} denotes the signed distance that separates both words and p_{11} equals to zero. As defined in Table (3), u_1 and u_2 are respectively mapped to two discrete random variables X_{1p} and X_{2k}[11] whose cohesiveness has to be tested in order to measure their attraction. A contingency table is defined as in Table (4) for each quadruplet $[p_{11}\, u_1\, p_{12}\, u_2]$ of the instance space.

	$\mathbf{X_{2k}}$	$\neg\, \mathbf{X_{2k}}$	**Row Total**
$\mathbf{X_{1p}}$	$f(p_{11}, u_1, p_{12}, u_2)$	$f(p_{11}, u_1, p_{12}, \neg u_2)$	$f(u_1)$
$\neg\, \mathbf{X_{1p}}$	$f(p_{11}, \neg u_1, p_{12}, u_2)$	$f(p_{11}, \neg u_1, p_{12}, \neg u_2)$	$f(\neg u_1)$
Column Total	$f(u_2)$	$f(\neg u_2)$	N

Table 4: A contingency table for 2-grams.

where N is the number of words present in the input text, $f(p_{11}, u_1, p_{12}, u_2)$ is the frequency of u_1, u_2 occurring together at the signed distance p_{12}, $f(p_{11}, u_1, p_{12}, \neg u_2)$ is the frequency of u_1 occurring with words other than u_2 at the signed distance p_{12}, $f(p_{11}, \neg u_1, p_{12}, u_2)$ is the frequency of u_2 occurring with words other than w_1 at the signed distance p_{21}[12], $f(p_{11}, \neg u_1, p_{12}, \neg u_2)$ is the frequency of u_1, u_2 never occurring at the signed distance p_{12}, $f(u_1)$ and $f(u_2)$ are the respective marginal frequencies of u_1 and u_2, $f(\neg u_1)$ and $f(\neg u_2)$ are respectively equal to N- $f(u_1)$ and N- $f(u_2)$.

3.2. Normalised Expectation

We define the Normalised Expectation existing between n textual units as the average expectation of one textual unit occurring in a given position knowing the occurrence of the other n-1 textual units 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 textual unit 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. The underlying concept of the Normalised Expectation is based on the conditional probability defined in Equation (1).

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

Equation 1: Conditional Probability.

Each textual unit of the text corpus is mapped to a discrete random variable in the Probability Space $(\Omega, A, P[.])$. Consequently, the definition of the conditional probability can be applied in order to measure the expectation of the occurrence of one textual unit in a given position knowing the occurrence of the other n-1 textual units also constrained by their positions. However, this definition does not accommodate the n-gram length factor. Naturally, an n-gram is associated to n possible conditional probabilities. The Normalised Expectation, based on a normalisation of the conditional probability, proposes an elegant solution to represent in a unique formula all the n conditional

[10] The representation of a contingency table with more than two dimensions is obviously not suitable for comprehension purposes.
[11] Positions p and k must satisfy the constraint imposed by p_{12} that the two textual units occur together at the signed distance p_{12}.
[12] p_{21} corresponds to the signed distance between u_2 and u_1.

probabilities involved by an n-gram. For that purpose we introduce the concept of the Fair Point of Expectation (FPE).
Let's consider a generic n-gram $[p_{11}\, u_1\, p_{12}\, u_2\, p_{13}\, u_3 \ldots p_{1i}\, u_i \ldots p_{1n}\, u_n]$ where p_{11} is equivalent to zero and p_{1i} (for i=2 to n) denotes the signed distance that separates the textual unit u_i from its pivot u_1 [13]. The extraction of one textual unit at a time from the generic n-gram gives rise to the occurrence of any of the n events shown in Table (5) where the underline (i.e. "_____") denotes the missing textual unit from the n-gram.

(n-1)-gram	**Missing textual unit**
$[\ \_\_\_\_\_\ p_{12}\, u_2\ \ p_{23}\ u_3 \ldots p_{2i}\ u_i \ldots p_{2n}\ u_n]$	$p_{11}\, u_1$
$[p_{11}\, u_1\ \_\_\_\_\_\ \ p_{13}\ u_3 \ldots p_{1i}\ u_i \ldots p_{1n}\ u_n]$	$p_{12}\, u_2$
...	...
$[p_{11}\, u_1\, p_{12}\, u_2\, p_{13}\, u_3 \ldots p_{1(i-1)}\, u_{(i-1)}\ \_\_\_\_\_\ p_{1(i+1)}\, u_{(i+1)} \ldots p_{1n}\, u_n]$	$p_{1i}\, u_i$
...	...
$[p_{11}\, u_1\, p_{12}\, u_2\, p_{13}\, u_3 \ldots p_{1i}\, u_i \ldots p_{1(n-1)}\, u_{(n-1)}\ \_\_\_\_\_\ ]$	$p_{1n}\, u_n$

Table 5: (n-1)-grams and missing textual units.

So, each event may be associated to a respective conditional probability that evaluates the expectation to occur the missing textual unit knowing its corresponding (n-1)-gram. The n conditional probabilities are introduced in Equation (2) and Equation (3).

$$p(p_{11}u_1 \mid [p_{12}u_2 \ldots p_{2i}\,u_i \ldots p_{2n}\,u_n]) = \frac{p([p_{11}u_1\,p_{12}u_2 \ldots p_{2i}\,u_i \ldots p_{2n}\,u_n])}{p([p_{12}u_2 \ldots p_{2i}\,u_i \ldots p_{2n}\,u_n])}.$$

Equation 2: Extraction of the pivot of the n-gram.

$$\forall i, i = 2..n, \quad p(p_{1i}u_i \mid [p_{11}u_1 \ldots p_{1(i-1)}u_{(i-1)}\,p_{1(i+1)}u_{(i+1)} \ldots p_{1n}u_n]) = \frac{p([p_{11}u_1\,p_{12}u_2 \ldots p_{1i}\,u_i \ldots p_{1n}\,u_n])}{p([p_{11}u_1 \ldots p_{1(i-1)}\,u_{(i-1)}\,p_{1(i+1)}\,u_{(i+1)} \ldots p_{1n}\,u_n])}.$$

Equation 3: Extraction of all the textual units except the pivot.

The analysis of the equations 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. Basically, the Fair Point of Expectation 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[14] of the (n-1)-grams contained in an n-gram and it is defined in Equation (4).

$$FPE([p_{11}u_1\,p_{12}u_2 \ldots p_{1i}\,u_i \ldots p_{1n}\,u_n]) = \frac{1}{n}\left(p([p_{12}u_2 \ldots p_{2i}\,u_i \ldots p_{2n}\,u_n]) + \sum_{i=2}^{n} p\left(\left[p_{11}u_1 \ldots \hat{p}_{1i}\ \hat{u}_i \ldots p_{1n}\,u_n\right]\right)\right)$$

Equation 4: Fair Point of Expectation.

In particular, the "^" corresponds to a convention frequently used in Algebra that consists in writing a "^" on the top of the omitted term of a given succession indexed from 2 to n.

[13] This n-gram is equivalent to the vector $[p_{11}\, u_1\, p_{12}\, u_2\, p_{23}\, u_3 \ldots p_{2i}\, u_i \ldots p_{2n}\, u_n]$ where p_{2i} denotes the signed distance that separates the textual unit u_i from u_2 and $p_{2i} = p_{1i} - p_{12}$ (for i=3 to n).
[14] In the case of n=2, the FPE is the arithmetic mean of the marginal probabilities.

So, 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 is proposed as a "fair" conditional probability. It is defined in Equation (5).

$$NE([p_{11}u_1 \ldots p_{1i}\,u_i \ldots p_{1n}\,u_n]) = \frac{p([p_{11}u_1 \ldots p_{1i}\,u_i \ldots p_{1n}\,u_n])}{FPE([p_{11}u_1 \ldots p_{1i}\,u_i \ldots p_{1n}\,u_n])}.$$

Equation 5: Normalised Expectation.

3.3. Mutual Expectation

For the specific purpose of word association, we introduce the Mutual Expectation measure. Justeson (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 and Ananiadou (1996) in the context of the 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 unit. So, the Mutual Expectation between n words is defined in Equation (6) based on the Normalised Expectation and the relative frequency.

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

Equation 6: Mutual Expectation.

Comparing to the previously proposed mathematical models, the Normalised Expectation allows evaluating the degree of cohesiveness that links together all the textual units contained in an n-gram (i.e. $\forall n,\ n \geq 2$) as it accommodates the n-gram length factor. For the specific case of word associations, the Mutual Expectation, based on the Normalised Expectation, enables to classify each word n-gram of the corpus by its degree of pertinence. In the following section, we present the LocalMaxs algorithm that retrieves the candidate terms from the set of all the valued n-grams by evidencing local maxima of association measure values.

4. Acquisition Process

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 & Hanks, 1990; Smadja, 1993; Daille, 1995; Shimohata, 1997, Feldman, 1998). This is defined by the underlying concept that there exits 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[15] (Habert *et al*, 1997). The LocalMaxs (Silva *et al*, 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 higher its score will be[16]. Second, multiword terms are localized 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

[15] They obviously vary with the association measure.
[16] The conditional entropy measure is one of the exceptions.

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:

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

$$W \text{ is a multiword term if } \begin{array}{l} (sizeof(W)=2 \ \wedge \ assoc(W) > y\ ) \\ \vee \\ (sizeof(W)\neq 2 \ \wedge \ assoc(W) \geq x \ \wedge \ assoc(W) > y) \end{array}$$

Among others, the LocalMaxs shows 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 experiences with different association measures. In particular, we tested the following normalised mathematical models in (Dias *et al*, 1999*d*): the Association Ratio (Church & Hanks, 1990), the Dice coefficient (Smadja, 1996), the ϕ^2 (Gale, 1990) and the Log-Likelihood Ratio (Dunning, 1993)[17]. In all cases the Mutual Expectation has proved to lead to better results than the other models.

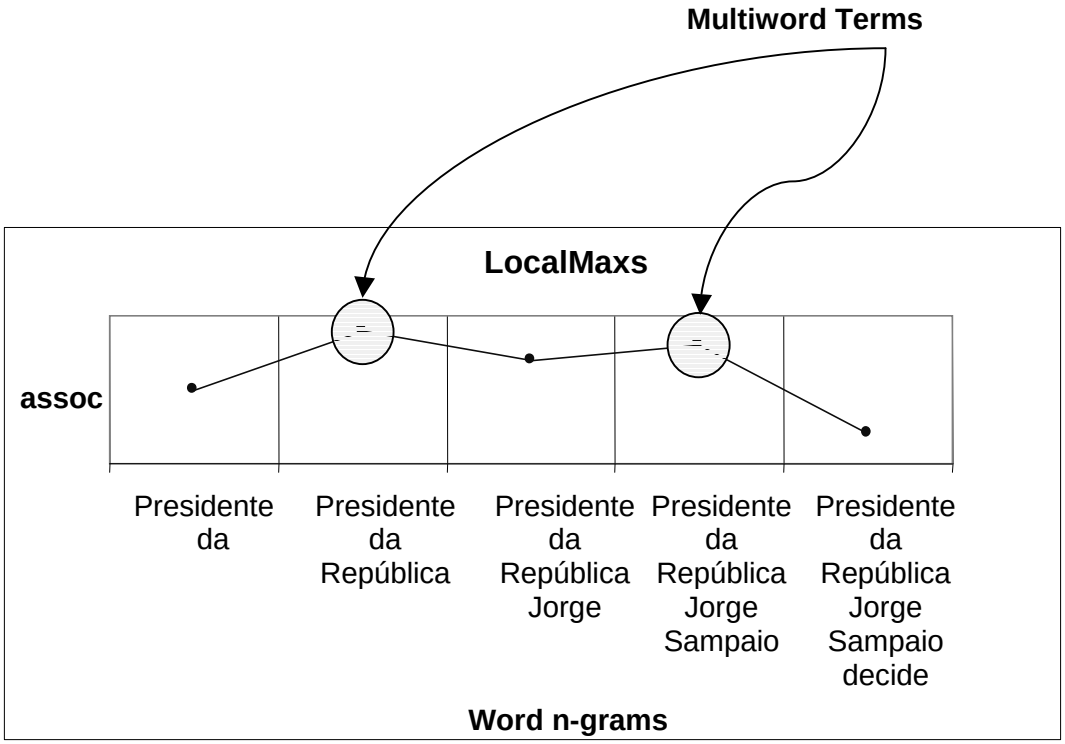

Figure 2: Election by Composition.

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 *Presidente da República Jorge Sampaio* (-*State President Jorge Sampaio*-) built from the composition of the extracted terms *Presidente da República* (-*State President*-) and *Jorge Sampaio* (-*Jorge Sampaio*-). This situation is

[17] Cramer and Pearson Coefficients have also been tested (Bhattacharyya & Johnson, 1977).

illustrated in Figure (2). Indeed, roughly exemplifying, one can expect that there are many State Presidents. Therefore, the association measure value of *Presidente da República Jorge* (-*State President Jorge*-) should be lower than the one for *Presidente da República* (-*State President*-) as there are many possible words, other than *Jorge*, that may occur after *Presidente da República* (-*State President*-). Thus, the association measure of any super-group containing the unit *Presidente da República* (-*State President*-) should theoretically be lower than the association measure for *Presidente da República* (-*State President*-). But, if the first name of the President is *Jorge*, the expectation to appear *Sampaio* is very high and the association measure value of *Presidente da República Jorge Sampaio* (-*State President Jorge Sampaio*-) 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 *Presidente da República Jorge Sampaio* (-*State President Jorge Sampaio*-).

So, the LocalMaxs algorithm proposes a flexible and robust solution for the extraction of multiword term candidates as it avoids the definition of global frequency and/or association measure thresholds based on experimentation. In the next section, we show the results obtained by applying the LocalMaxs and the Mutual Expectation over a raw Portuguese text and we access the results reached by combining word statistics with linguistic information endogenously acquired from the same corpus previously tagged.

5. Two Experiments

Due to its flexibility, our methodology has been tested on text corpora of different sizes, domains (Dias *et al*, 1999*c*) and languages (Dias *et al*, 1999*b*; Dias *et al*, 1999*e*). We have also performed experiments with a variegate number of normalised mathematical measures (Dias *et al*, 1999*d*). In this paper, we present the results obtained by performing two distinct experiments.

First, multiword terms are extracted from a Portuguese raw text by exclusively considering statistical word regularities. In that case, each word n-gram is associated to its Mutual Expectation value and the LocalMaxs extracts all the possible terms from the set of all valued word n-grams. In the second experiment, the system extracts terminologically relevant multiword lexical units by combining word statistics with linguistic information that is acquired endogenously from the same Portuguese corpus previously tagged. In that case, each word n-gram is linked to its corresponding tag n-gram and the final association measure value of the word n-gram is the product between its Mutual Expectation value and the Normalised Expectation value of its associated tag n-gram. This situation is illustrated in Figure (3).

The input text corpus has been extracted from a collection of advises formulated by the *Procuradoria Geral da República* in the context of the project "*PGR – Acesso Selectivo aos Pareceres da Procuradoria Geral da República*" funded by the *Fundação para a Ciência e Tecnologia*[18]. The overall corpus contains up to 3 million words but only 1.5 million words have been tagged using the neural network tagger developed by Marques (2000). Both the experiments presented in this paper have been tested on a sub-corpus containing 500000 words previously tagged. We will discuss the technical conditions of the experiment in the last sub-section.

5.1. Purely Statistical Approach: 1º Experiment

The first experiment consists in extracting multiword terms by using exclusively statistical word regularities. For that purpose, part-of-speech tags are considered meaningless for the acquisition process and all the tags are excluded from the original input tagged corpus. The set of all the word n-grams is then calculated from the "text of words" and each word n-gram is associated to its Mutual Expectation value. Finally, the LocalMaxs extracts the multiword term candidates from the set of all

[18] The reader may find more information about this project by acessing the following web site http://kholosso.di.fct.unl.pt/~di/people.phtml?it=CENTRIA&ch=gpl.

the valued word n-grams. The results show that three categories of multiword terms were extracted: base terms, terms obtained by composition and terms obtained by modification.

5.1.1. Base Terms A base term corresponds to a contiguous word n-gram that does not contain any other extracted word n-gram. So, a base term is neither constrained by its length nor by its syntactical pattern. We exemplify some of the extracted base terms in Table (6).

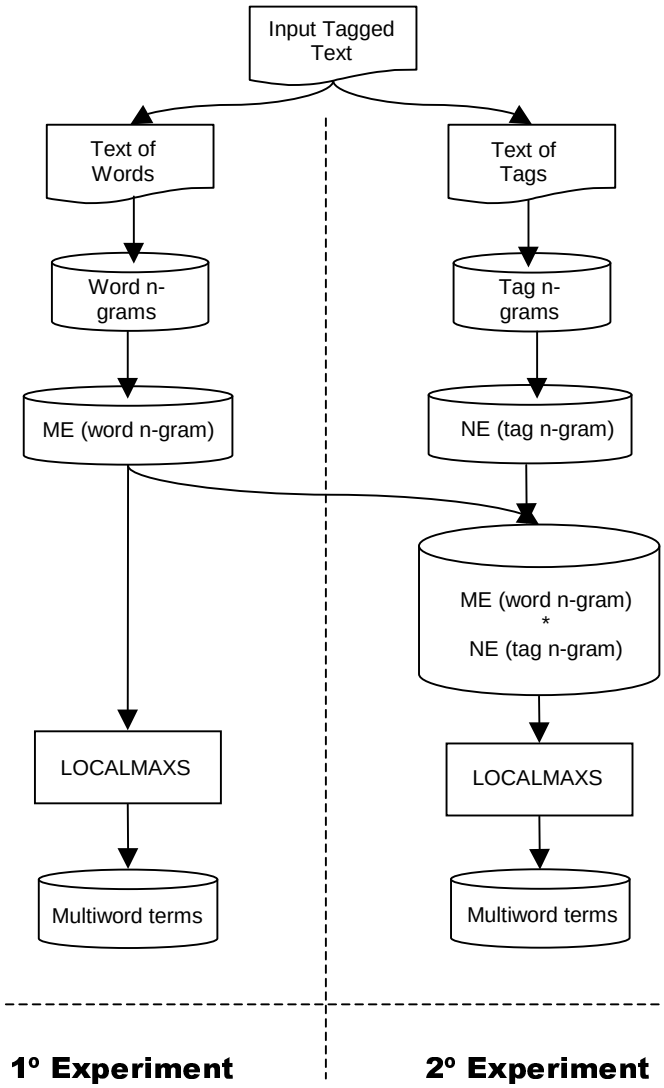

Figure 3: Description of two experiments.

5.1.2. Terms obtained by composition A term obtained by composition corresponds to a multiword term that is built from one or more base terms. This category embodies the specific constructions of juxtaposition, substitution, postposition and co-ordination proposed by B. Daille (1995)[19]. We exemplify some of the extracted terms obtained by composition in Table (7) [20].

5.1.3. Terms obtained by modification A multiword term obtained by modification is a non-contiguous word n-gram containing exactly one interruption[21]. Indeed, the insertion of modifiers

[19] We do not follow Daille´s classification as the results show a more variegate set of linguistic phenomena than she evidences by using pre-defined syntactical patterns for each category.
[20] The base terms are identified in brackets.
[21] The terms characterised by several interruptions exhibit lexicographically relevant units that are most of the time terminologically irrelevant.

inside a term implies the introduction of a flexibility factor that corresponds to an interruption. In fact, it is likely that several word occurrences may modify a complex sequence of words.

For example, both words *Europeu* (-*European*-) and *Mundial* (-*World*-) may modify the complex sequence *Conselho das Telecomunições* (-*Telecommunication Council*-) by introducing some specification factor and resulting in the following pair of modified word sequences: *Conselho Europeu das Telecomunições* (-*European Telecommunication Council*-) and *Conselho Mundial das Telecomunições* (-*World Telecommunication Council*-). The analysis of the non-contiguous word n-grams allows representing the terms obtained by modification by identifying in the word sequence each set of modifiers as an interruption. Thus, the LocalMaxs associated to the Mutual Expectation would elect the multiword term *Conselho __ das Telecomunições* (-*__ Telecommunication Council*-) that embodies both the modified complex sequences. In that case, the interruption identifies the flexible component of the complex term. The results show that the modifiers do not correspond compulsorily to adjectives or adverbs as Daille (1995) claims. In fact, the modifiers may embody a great deal of morpho-syntactical categories as illustrated in Table (8).

The terms obtained by modification are an important asset in the context of Information Retrieval. Indeed, on one hand, a non-contiguous term represents a concept that may be specialised by matching its interruption to one of its modifiers. Thus, it is possible to generalise a query by proposing as indexing term[22] the general concept of the multiword term instead of its specialised occurrences. For example, if a user searches information about *transporte de matérias perigosas* (-*transport of dangerous materials*-), the system should reasonably retrieve related documents that deal with *transporte de substâncias perigosas* (-*transport of dangerous substances*-). A potential solution to this problem is to consider the term *transporte de __ perigosas* (-*transport of dangerous __*-) as indexing term. On the other hand, the terms obtained by modification allow the identification of discriminating hapaxes that enable to recover a great number of pertinent texts. Indeed, a non-contiguous term corresponds to a set of specialised terms that can occur only once in the text collection. For example, the following multiword term *proposta de __ do Conselho* (-*proposal of __ of the Council*-) corresponds exactly to two hapaxes: *proposta de Directiva do Conselho* (-*proposal of Directive of the Council*-) and *Proposta de Regulamento do Conselho* (-*proposal of Regulation of the Council*-). So, the non-contiguous terms used as indexing terms allow the identification of specific terms independently of their frequency.

ME	Freq.	Base Terms
0.000107062	12	*Comunicação Social* (-*the media*-)
0.000105067	3	*Oliveira Ascensão* (-*Oliveira-Ascenção*-)
0.000105067	3	*cobrar taxas* (-*to tax*-)
0.000105519	5	*Liberdade de Informação* (-*Press freedom*-)
0.000100499	2	*oferta e procura* (-*supply and demand*-)
0.000100499	2	*entrar em vigor* (-*to come into force*-)
0.000100499	2	*Associação dos Arquitectos Portugueses* (-*Portuguese Union of Architects*-)
0.000102732	4	*Regulamento Municipal de Edificações Urbanas* (-*Municipal Regulation for Building Constructions*-)
0.000100499	2	*rendimento familiar anual ilíquido per capita* (-*unearned annual family income per capita*-)

Table 6: Base terms.

[22] An indexing term is either an indexing key (indexing documents) or a term that is proposed to the user in his search for information (user-friendly environment).

ME	Freq.	Terms obtained by composition
0.000102	3	*[Direcção Geral] dos Desportos* (*-Sport [General Board]-*)
0.000102	4	*Teoria Geral do [Direito Civil]* (*-General Theory of the [Civil Law]-*)
0.000128	2	*licenciamento municipal de obras particulares* (*-[municipal licensing] for particular works-*)
0.000102	2	*[estabelecimentos oficiais] não militares* (*-non military [official buildings]-*)
0.000102	2	*[pessoas colectivas] de [utilidade pública]* (*-[public utility] [collectives]-*)
0.000102	2	*[artigo 35º] da [Lei de Imprensa]* (*-[Article nº 35] of the [Press Law]-*)
0.000102	2	*[Liga Nacional] de [Futebol Profissional]* (*-[National League] of [Professional Football]-*)

Table 7: Terms obtained by composition.

ME	Freq.	Terms obtained by modification	Modifiers
0.000105	2	*controle _____ fronteiras* (*-control of the border- Border control*)[23]	*de* (*-of-*) *das* (*-of the-*)
2.708e-05	2	*transporte de _____ perigosas* (*-transport of dangerous _____-*)	*matérias* (*-materials-*) *substâncias* (*-substances-*)
2.708e-05	4	*artigo _____ do regulamento* (*-article _____ of the Regulation-*)	*3º* *6º* *32º* *45º*
0.000102	2	*proposta de _____ do Conselho* (*-proposal of _____ of the Council-*)	*Directiva* (*-Directive-*) *Regulamento* (*-Regulation-*)

Table 8: Terms obtained by modification.

5.1.4. Limits of the Statistical Process In the context of Information Retrieval, we are particularly interested in the extraction of multiword terms. However, statistical methodologies extract multiword lexical units that can not be considered terms (Habert & Jacquemin, 1993). Our methodology does not avoid this problem. Indeed, the detailed analysis of the results shows that adverbial, adjectival, prepositional and conjunctive locutions are also retrieved. Furthermore, statistical works based on the study of text corpora identify textual associations in the context of their usage. As a consequence, many terminologically relevant structures can not be introduced directly into lexical databases as they do not guarantee adequate linguistic structures for that purpose. For example, the LocalMaxs associated to the Mutual Expectation extracts the multiword lexical unit *a assembleia municipal* (*-the municipal assembly-*) as whenever the multiword unit *assembleia municipal* (*-municipal assembly-*) occurs in the text collection, the determinant *a* (*-the-*) also occurs. In that case, we should reasonably consider the multiword term *assembleia municipal* (*-municipal assembly-*) as the only relevant unit. One final drawback when extracting multiword lexical units based only on statistical measures in a

[23] This multiword unit embodies a fixation process. The term Control of the borders is becoming Border Control.

non-continuous environment is the election of incorrect non-contiguous lexical units. Indeed, the associations retrieved by the system may not embody correct lexical associations but instead may characterize some local spurious associations. For example, the following unit *a __ __ técnicos responsáveis* (*-the__ __ responsible technicians-*) is elected by the system as whenever the unit *técnicos responsáveis* (*-responsible technicians-*) occurs, the determinant *a* (*-the-*) also incidentally occurs three positions ahead as shown in Table (9).

oficial	***a***	*exigir*	*aos*	***técnicos***	***responsáveis***	*pelos*
(*-oficial-*)	(*-for-*)	(*-demanding-*)	(*-to the-*)	(*-technicians-*)	(*-responsible-*)	(*-for-*)
cancelada	***a***	*inscrição*	*dos*	***técnicos***	***responsáveis***	*pelo*
(*-cancelled by-*)	(*-the-*)	(*-inscription-*)	(*-of the-*)	(*-technicians-*)	(*-responsible-*)	(*-for-*)
obdecer	***a***	*qualificação*	*dos*	***técnicos***	***responsáveis***	*por*
(*-obey to-*)	(*-the-*)	(*-qualification-*)	(*-of the-*)	(*-technicians-*)	(*-responsible-*)	(*-for-*)

Table 9: Concordances for *técnicos responsáveis* (*-responsible technicians-*).

One way to tackle this problem is to introduce linguistic information in the acquisition process. Indeed, syntactical regularities have to be taken into account in order to filter incorrect lexical units. That is the goal of our second experiment.

5.2. Hybrid Approach: 2º Experiment

In order to overcome the problems evidenced by most of the statistical approaches, hybrid linguistic-statistical methods define co-occurrences of interest in terms of syntactical patterns and statistical regularities. Some approaches reduce the searching space to groups of words that correspond to *a priori* defined syntactical patterns (Noun+Adj, Noun+Prep+Noun etc...) and then apply statistical measures to classify the pertinent sequences (Justeson, 1993; Daille, 1995; Heid, 1999). Other approaches first identify statistical word regularities and then apply *a priori* defined syntactical filters to extract multiword term candidates (Smadja, 1993). However, such systems do not sufficiently tackle the problem of the interdependency between the filtering stage and the acquisition process as they propose that these two steps should be independent. Moreover, by defining *a priori* syntactical filters they do not extract a great deal of multiword terms that embody non-nominal structures like compound verbs. In order to overcome these difficulties, we propose an original experiment that combines word statistics with endogenously acquired linguistic information. We base our study on two assumptions. On one hand, a great deal of studies in lexicography and terminology access that most of the multiword terms evidence well-known morpho-syntactic structures (Noally, 1990; Gross, 1996). On the other hand, multiword terms are recurrent combinations of words[24]. Consequently, it is reasonable to think that the syntactical patterns embodied by the multiword terms should be endogenously acquired by using statistical scores over "texts of tags". So, in parallel to the evaluation of the Mutual Expectation value of each word n-gram, all the words of the input text corpus are pruned out and the Normalised Expectation is applied to each tag n-gram previously calculated from the “text of tags”[25]. Finally the LocalMaxs is applied over the set of all the word n-grams associated to their new association measure value that is the product between their Mutual Expectation value and the Normalised Expectation value of their associated tag n-gram.

5.2.1. What do we gain ? Some of the problems evidenced by purely statistical methods are partly solved. On one hand, the introduction of endogenously acquired linguistic information in the acquisition process guarantees (especially, for the case of the 3-grams and the 4-grams) the extraction of multiword terms with adequate linguistic structures that allows their direct introduction into lexical databases. For example, the system extracts the multiword term *concessão de bolsas* (*-granting of*

[24] According to Habert and Jacquemin (1993), the multiword terms may represent a fifth of the overall surface of the text corpus.

[25] We did the experience with the Mutual Expectation measure but the results were not satisfactory as there are no linguistic evidence, contrarily to the case of the word n-grams, that the more frequent tag n-grams are more discriminating than the less frequent ones.

scholarships-) although the sequence always occurs in the corpus with the preposition *de* (*-of-*) concatenated to its end (i.e. *concessão de bolsas de* (*-granting of scholarships of-*)). On the contrary, the purely statistical approach elects the whole unit *concessão de bolsas de* (*-granting of scholarships of-*). The same scenario applies to the term *assembleia municipal* (*-municipal assembly-*) that is elected without its associated determinant. On the other hand, correct multiword units that were not extracted by exclusively relying on statistical word regularities were retrieved benefiting from the identification of relevant syntactical patterns. For example, *concessão de auxílios* (*-granting of help-*), *jogo de futebol* (*-football game-*) and *establecimentos de ensino* (*-scholar establishments-*) were retrieved as they embody the idiosyncratic syntactical pattern Noun+Prep+Noun. More results are presented in Table (10).

Terms obtained by hybrid experiment	**Corresponding term obtained by statistical experiment**
BAPTISTA MACHADO	*. ___ ___ BAPTISTA MACHADO*
Direitos Conexos (*-Associated Rights-*)	*De autor e Direitos Conexos* (*-from autors and Associated Rights-*)
imagens recolhidas (*-extracted pictures-*)	*de imagens recolhidas* (*-from extracted pictures-*)
espectáculos cinematográficos (*-movie show-*)	*espectáculos cinematográficos ,* (*-movie show ,-*)
Direitos do homem (*-Human Rights-*)	*dos Direitos do homem* (*-of the Human Rights-*)
Ministério da Justiça (*-Justice Ministry-*)	*Boletim do Ministério da Justiça* (*-Justice Ministry Jornal-*)
Direito a Informação Desportiva (*-Sport Information Right-*)	*Direito a Informação* (*-Information Right-*)
Direito de crónica (*-Chronicle Right-*)	*o Direito de crónica e* (*-the Chronicle Right and-*)
Ministro das Obras Públicas (*-Minister of the Public Works -*)	*das Obras Públicas* (*-of the Public Works -*)
elaborar projectos (*-to elaborate projects-*)	--
ensino primário (*-primary school-*)	--
proceder a (*-to procede-*)	*devem proceder a* (*-must procede-*)
campo de aplicação (*-application field-*)	*seu campo de aplicação* (*-his application field-*)

Table 10: Comparative results between both experiments.

However, the introduction of linguistic information also leads to incoherence and noise in the retrieval process as we access in the following subsection.

5.2.2. What do we loose ? The introduction of linguistic information in the acquisition process evidences three major drawbacks. At this point, we must stress that the quality of the output strongly depends on the task being tackled and a precision measure should be calculated in relation with a particular task. Indeed, a translator and a lexicographer may evaluate the same results in a different manner. However, in order to define some "general" rule to measure the precision of the system, we propose that a word n-gram is a multiword term if it is grammatically appropriate. By grammatically appropriate, we refer to compound nouns/names and compound verbs. As a consequence, the results of precision are calculated using the quotient between the number of grammatically correct units and the number of all the extracted units.

On one hand, poor precision results are obtained for the case of 2-grams. In fact, many irrelevant syntactical patterns evidence high association scores that introduce noise in the acquisition process. For example, the high scores associated to the syntactical patterns Det+Noun and Prep+Noun result in electing uninteresting word n-grams like *o concurso* (*-the contest-*), *da carreira* (*-of the career-*). This situation is illustrated in Table (11).

On the other hand, most of the elected non-contiguous units reveal interruptions that rarely correspond to the occurrence of different modifiers. For example, the following multiword lexical unit *Medalha do ___ militar* (*-Medal of the military ___-*) is elected by the LocalMaxs although the probability to occur the single word *mérito* (*-merit-*) is one (i.e. *mérito* (*-merit-*) is the only token that fulfil the interruption in the corpus).

Finally, the syntactical pattern Noun+Prep+Noun that legitimately receives a high association score, as it corresponds to a well-known morpho-syntactic sequence in the context of multiword terms, introduces noise in the election process. For example, the multiword lexical unit *Ministro dos Negócios* (*-Minister of the Affairs-*) is preferably elected in comparison to the correct multiword term *Ministro dos Negócios Estrangeiros* (*-Minister of the Foreign Affairs-*). The same applies to the sequence *Direito de Informação Jornalística* (*-Right to Press Information-*) that is not extracted while the system elects the unsatisfactory unit *Direito de Informação* (*-Right to Information-*). This result particularly brings into question the results obtained by the systems that define *a priori* syntactical patterns. Indeed, they always count with the problematic Noun+Prep+Noun unit (Justeson, 1993; Daille, 1995; Heid, 1999).

Terms obtained by statistical experiment	**Corresponding term obtained by hybrid experiment**
isenção de (*-non-payment of-*)	*isenção de propinas* (*-non-payment of scholarship -*)
acesso a (*-acess to-*)	*acesso a categoria* (*-acess to the category-*)
anos de serviço (*-years of service-*)	*anos de serviço efectivo* (*-years of effective service-*)
Estatuto social do Bombeiro (*-social status of the fireman-*)	"*Estatuto social do Bombeiro*" (-"*social status of the fireman*"-)
licenciamento municipal de obras (*-city hall funding for works-*)	*licenciamento municipal de obras públicas* (*-city hall funding for public works-*)
a primeira (*-the first-*)	--

Table 10: Comparative results between both experiments.

In Figure (4), we present the comparative results of precision rate between both experiments. The precision rates have been calculated following the guidelines mentioned above (i.e. a multiword lexical unit is a correct multiword term if it is a compound name/noun or a compound verb). The results access interesting phenomena. Indeed, the reading of the curves suggests that the purely probabilistic method performs better for the case of the extraction of multiword terms containing two words than the hybrid methodology. On the other side, the hybrid method over-performs the purely statistical methodology for the extraction of multiword terms containing three to four words. Finally, both methodologies tend to elect the same set of multiword terms containing up to five words. These results evidence interesting issues for possible studies in cognitive science, psychology and linguistics. Indeed, they clearly suggest that the syntactical patterns should not be taken into account for the extraction of 2-grams while syntactical patterns seem to play a fundamental role for the extraction of terms containing between three and four words. Moreover, the syntactical patterns turn out to be less meaningful for the extraction of multiword terms containing up to five words. As a consequence, the "one methodology" paradigm for the extraction of multiword terms clearly needs to

be questioned as the results clearly suggest that different methodologies should be taken into account depending on the length of the terms being tackled.

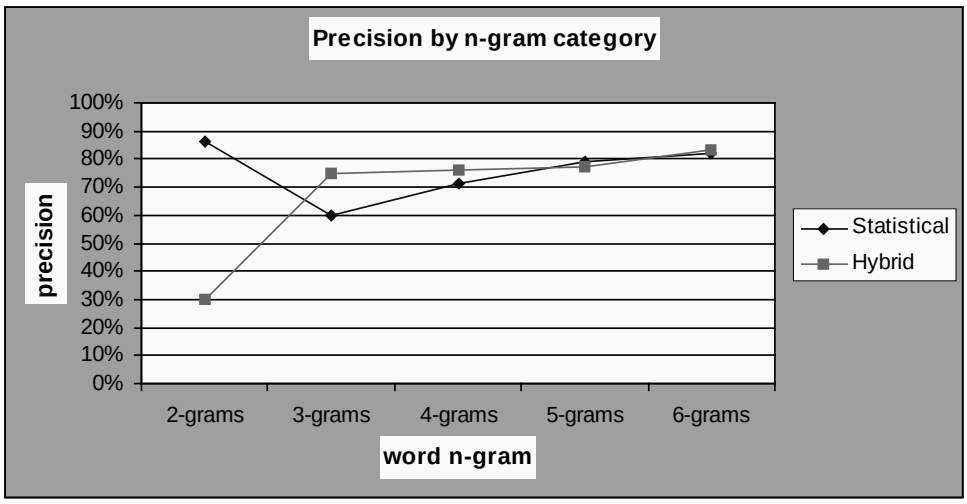

Figure 4: Precision by n-gram category.

5.3. Technical Conditions

The system has initially been developed for experimental purposes in script shell for Linux. Indeed, the *gawk* language affords interesting features for text handling. However, its lack of representation power leads to its incapacity to handle optimised programming. In order improve its efficiency, the system is currently being developed in C using optimised data structures and search algorithms.

Moreover, in order to cope with the huge amount of data resulting from the calculation of all the possible combinations of n-grams, we have designed an optimised algorithm that avoids the calculation of duplicate n-grams as the text is being processed. Duplicates are equivalent n-grams that embody different representations. For example, [*United* +1 *States*] and [*States* -1 *United*] are duplicates. Finally, in order to reduce the memory space needed, the text is previously encoded before the system is run. We use a Huffman-like encoding (i.e. the most frequent words are given the smallest possible codes) that reaches a compression rate of 52.8%.

Both experiments have been processed on a Pentium 200Mhz with 64Mbytes of Random Access Memory and 10 Gbytes of Disk storage. While the first experiment only took a couple of hours to process, the second experiment revealed itself to be much more time-consuming taking around 24 hours to elect the multiword terms. We are confident that these results will be over-performed by (1) the C version of the system that introduces optimised programming and (2) the extension of the Random Access Memory. However, we are aware that theoretical work has to be done to calculate the complexity of all the algorithms of the system.

6. Conclusion

In this paper, we presented an original methodology that allows extracting multiword terms by either (1) exclusively considering statistical word regularities or by (2) combining word statistics with endogenously acquired linguistic information. For that purpose, we proposed a new association measure called the Mutual Expectation and a new acquisition process called the LocalMaxs that, when associated, avoid the definition of global thresholds based on experimentation and do not require enticement techniques to extract relevant multiword terms. The results obtained by applying the methodology over a Portuguese tagged corpus show that the introduction of linguistic information might benefit the acquisition process although it also evidenced major drawbacks. The results also show that the "one methodology" paradigm for the extraction of multiword terms clearly needs to be questioned as the results clearly suggest that different methodologies should be taken into account depending on the length of the terms being tackled. In the context of Information Retrieval, experiments will have to be performed in order to decide whether the introduction of linguistic

information oversteps or not the purely statistical methodology for the retrieval process. However, we hardly believe that more experiments have to be performed in the same way in order to evaluate the real interdependency between the filtering stage and the acquisition process. In particular, we tested our methodology on the same corpus previously tagged with a more complete set of tags and the results proved not to lead to improvements. Instead, they evidenced worst results for precision and recall evidencing that richer information does not necessarily benefit the acquisition process.

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.

References

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

Betts, R. & Marrable D. (1991). Free Text vs controlled vocabulary, retrieval precision and recall over large databases. In *Online Inf 91* (pp 153--165). London.

Bourigault, D. (1996). Lexter, a Natural Language Processing Tool for Terminology Extraction. In *Proceedings of 7th EURALEX International Congress.*

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

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

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. & Guilloré, S. & Lopes, J.G.P. (1999*a*). 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. & Guilloré, S. & Lopes, J.G.P. (1999*b*). Multilingual Aspects of Multiword Lexical Units. In *Proceedings of Workshop on Language Technologies*, Ljubljana, Slovenia.

Dias, G. & Guilloré, S. & Vintar, S. & Lopes, J.G.P. (1999*c*). Identifying and Integrating Terminologically Relevant Multiword Units in the IJS-ELAN Slovene-English Parallel Corpus. In *Proceedings of 10th CLIN*. Utrecht Institute of Linguistics OTS, Utrecht, Netherlands.

Dias, G. & Guilloré, S. & Lopes, J.G.P. (1999*d*). Mutual Expectation: a Measure for Multiword Lexical Unit Extraction. In *Proceedings of VExTAL Venezia per il Trattamento Automatico delle Lingue*. Universitá Cá Foscari. Venezia, Italy.

Dias, G. & Guilloré, S. & Lopes, J.G.P. (1999*e*). Multiword Lexical Units Extraction. In *Proceedings of the International Symposium on Machine Translation and Computer Language Information Processing.* Beijing, China.

Dias, G. & Guilloré, S. & Lopes, J.G.P. (1999*f*). Extraction Automatique d'Associations Textuelles à Partir de Corpora Non Traités. In *Proceedings of 5es Journées Internationales d'Analyse Statistique des Données Textuelles.* Lausanne, Suisse.

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

Enguehard, C. (1993). Acquisition de Terminologie à partir de Gros Corpus. In Proceedings of *Informatique & Langue Naturelle* (pp 373--384).

Evans, D. & Lefferts, R. (1993). Design and Evaluation of the CLARIT-TREC-2 System. *TREC93* (pp 137--150).

Feldman, R. (1998). Text Mining at the Term Level. In Proceedings of *PKDD'98*. Lecture Notes in AI 1510. Springer Verlag.

Frantzi, K.T. & Ananiadou S. (1996). Retrieving Collocations by Co-occurrences and Word Order Constraint. In *Proceedings of 16th International Conference on Computational Linguistics (COLING'96)* (pp 41--46). Copenhagen, Denmark.

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

Grefenstette, G. (1994). *Explorations In Automatic Thesaurus Discovery*, Boston/Dordrecht/London, Kluwer Academic Publishers.

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

Habert, B. & Jacquemin, C. (1993). Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques. In *Traitement Automatique des Langues* 34(2). Association pour le Traitement Automatique des langues, France.

Habert, B. & Nazarenko, A. & Salem A. (1997). *Les linguistiques du Corpus*. Paris, Armand Colin.

Heid, U. (1999). Extracting Terminologically Relevant Collocations from German Technical Texts. http://www.ims.uni-stuttgart.de/~uli/.

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

Marques, N. (2000). Metodologia para a Modelação Estatística da Subcategorização Verbal. Ph.D. Thesis. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Lisbon, Portugal.

Mason, O. (1997). The Weight of Words: an Investigation of Lexical Gravity. In Proceedings of *PALC'97*.

Noally, M. (1990). *Le substantif épithète*. Paris, PUF.

Quaresma, P. & Rodrigues, I. & Lopes, J.G.P. (1998). PGR Project: The Portuguese Attorney General Decisions on the Web. In Proceedings of *The Law in the Information Society*. Instituto per la documentazione giuridica del CNR, C. Ciampi, E. Marinai (ed.), Florence, Italy.

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. In *Computational Linguistics*, 19 (1) (pp 143--177).

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