A Local Maxima method and a Fair Dispersion Normalization for extracting *multi-word* units from corpora

Joaquim Ferreira da Silva
Universidade Nova de Lisboa, FCT/DI
Quinta da Torre, 2725, Monte da Caparica
jfs@di.fct.unl.pt

Gabriel Pereira Lopes
Universidade Nova de Lisboa, FCT/DI
Quinta da Torre, 2725, Monte da Caparica
gpl@di.fct.unl.pt

Abstract

The availability of *multi-word* units (MWUs) in NLP lexica has important applications: enhances parsing precision, helps on attachment decision and enables more natural interaction of non-specialists users with information retrieval engines, among other applications.

Most statistical approaches to MWUs extraction from corpora measure the association between two words, define thresholds for deciding which bigrams may be elected as possible units and use complex linguistic filters and language specific morpho-syntactic rules for filtering those units. In this paper we present:

- A new algorithm (*LocalMaxs*) for extracting complex units made up of 2 or more adjacent words (*n*-grams, with $n \geq 2$).
- A new measure of "glue" or association between the words of any size *n*-gram.
- An exhaustive comparison of our association measure with other known measures (*Loglike*, ϕ^2, etc.).
- A new normalization, *fair dispersion point normalisation,* for current statistical measures (*Loglike*, ϕ^2, etc.) that enhances the precision and recall of the MWUs extracted by these measures.

Introduction

Multi-word units (MWUs) such as compound nouns (*Zimbabwean minister of foreign affairs*, *bacalhau à braz -a portuguese dish-*, *World Trade Centre*) compound verbs *(entrar em vigor -to come into force)*, compound prepositions *(a partir de -after, since),* compound conjunctions *(a fim de -in order to)* or frozen forms (*raining cats and dogs, plus ou moins*), etc., actually appear in real texts. Usually, most of these units are not available in current dictionaries, and so they should be automatically extracted from corpora, in order to enable their incorporation in NLP specialized lexica. Such lexica will enable parsers to be more effective and efficient. Moreover those MWUs can be used for refining information retrieval searches, enhancing precision, recall and the naturalness of the resulting interaction of those searches.

1 Finding units in the text

The classification of a *n*-gram as a lexical unit is a challenge. Let us take the illustrative example of a frequently co-occurring multi-word unit: *Margaret Thatcher*. It is clear for us that this *2*-gram is a compound proper name i.e., a lexical unit. In fact, when the word *Margaret* appears in a text, it is likely that the word *Thatcher* will follow it, and the probability of the word *Margaret* appearing in the position immediately prior to *Thatcher* is very high too. And so we may think of a "glue" sticking those 2 words together.

However, it is also possible to find sentences like "*…and Thatcher declared…*" or "*…the young Margaret used to play…*", where the words *Margaret* and *Thatcher* do not appear together. This means that there is some word "dispersion" in the next position after word *Margaret* as well as there is some kind of "dispersion" for the word that may precede *Thatcher*. However these two words still have a strong "glue" sticking them together as they frequently co-occur. Moreover we claim that this "glue" has a sufficiently high value for *Margaret Thatcher* to be considered a *2*-

gram *unit*. So, let us take for granted that we have a function *g(.)* that measures the "glue" sticking together every two adjacent words within a candidate *n*-gram. Let then *g(.)* be the *n*-gram's "glue". Let the range of *g(.)*[1] be an interval of values bound by real values. So, in order to consider the bigram w_1,w_2 as *a 2*-gram unit, it seems reasonable to demand a relatively high value for $g((w_1,w_2))$[2]. Usual statistical approaches for classifying bigrams define a threshold for various measures: *simple frequency* (Smadja (1991)), *mutual information* and *Dice coefficient* (Smadja & Hatzivassiloglou & McKeown (1996)). Above that threshold a bigram is selected for further pruning (namely by using morpho-syntactic information). However thresholds pose important empirical problems related to its value that depends on the *corpora* size, and other factors. Our approach (*LocalMaxs)* bypasses those problems, since there is no need for a threshold. It relies on local maxima for the association ("glue") measure.

In section 2 we study the *n*-gram's "glue". Concepts such as *n*-gram's *dispersion points*, *hole-free n*-gram and *pseudo-bigram transformation*, are defined in order to create a conceptual "glue" measure *g(.)* that transforms every *n*-gram into a pseudo-bigram. Section 3 describes *LocalMaxs* algorithm. In section 4 we present different kinds of "glue" measures currently used: *Mutual information* (*SI*), ϕ^2 *coefficient, Dice coefficient* and *Loglike coefficient*. In this section we introduce the concept of *fair dispersion point normalisation* and show how it can be applied in order to enhance the various "glue" measures for *n*-grams of size greater than 2. A new "glue" measure is also introduced: *SCP*. In section 5 we present the results obtained and assess them. In this section we also discuss related work in this area. In the last section we present conclusions and further work.

[1] In section 4 we will discuss several measures for calculating *g(.)*.

[2] To keep *g(.)* as a one-argument function we write $g((w_1 \ldots w_n))$ for the *g(.)* value of the *n*-gram $w_1 \ldots w_n$. There will come other functions ex: *SI(.)*, that will be a one-argument function too, so we can write for example *SI(W)*, $SI((w_1,w_2,w_3))$, $SI((w_1 \ldots w_n))$, etc.

2 Definitions

Definition 1: *N*-gram's *dispersion points*

As we can conclude from the previous section every *2*-gram has one *dispersion point*: that we may "locate" between the positions of the words of the *2*-gram. After that point and before it, in a *corpus* several words may appear: after "*Margaret*" in the previous example we found "*Thatcher*" and "*used*", and before "*Thatcher*" we found "*Margaret*" and "*and*". We may generalize and say that a *n*-gram has *n - 1 dispersion points* (the first *dispersion point* after the first word, the second *dispersion point* after the second word, … the *n - 1 dispersion point* after the $n - 1^{th}$ word).

Definition 2: *Hole-free / uninterrupted n*-gram.

We say that a *n*-gram is a *hole-free* or *uninterrupted n*-gram if every physical position of the *n*-gram is occupied by just one possible word, i.e., within a *hole-free n*-gram there is no physical position corresponding to a "hole" that can be occupied by any word of a set of two or more different words.

Definition 3: *Pseudo-bigram transformation*

Although every *n*-gram has *n - 1 dispersion points,* we may see the *n*-gram as having just one *dispersion point* "located" between a left and a right part of the *n*-gram: $w_1 \ldots w_p$ and $w_{p+1} \ldots w_n$, where *p* can be any value such that $1 \leq p \leq n - 1$. Reflecting this *transformation,* as we shall see in section 4, function *g(.)* will assign values of the same magnitude to different *n*-grams whatever the size each *n*-gram has, since it may be "seen" as a pseudo-bigram. This enables us to compare the "glue" values assigned to different size *n*-grams, and therefore the study of the evolution of the *n*-gram's "glue" when the *n*-gram's size changes. The information obtained from this evolution is very important for the selection of a *n*-gram as a MWU.

3 The algorithm

LocalMaxs is an algorithm that works with a *corpus* as input and produces MWUs from that *corpus*. In the context of *LocalMaxs,* we define:

- An *antecedent* (in size) of the *hole-free n*-gram $w_1,w_2 \ldots w_n$, $ant((w_1 \ldots w_n))$, is a *hole-free* sub-*n*-gram of the *n*-gram $w_1 \ldots w_n$, having size *n - 1*. i. e., the *(n-1)*-gram $w_1 \ldots w_{n-1}$ or $w_2 \ldots w_n$.

- A *successor* (in size) of the *hole-free n*-gram $w_1,w_2 \ldots w_n$, *succ(M)*, is a *hole-free* (*n+1*)-gram *N* such that *M* is an *ant(N)*. i. e., *succ(M)* contains the *n*-gram *M* and an additional word before (on the left) or after (on the right) *M*.
- Let *W* be a *hole-free n*-gram; we say that *W* is a MWU if[3]: (3.1)

$$g(W) \geq g(ant(W)) \wedge g(W) > g(succ(W)) \quad \forall_{ant(W),succ(W)} \text{ (if } W\text{'s size} \geq 3)$$

$$g(W) > g(succ(W)) \quad \forall_{succ(W)} \text{ (if } W\text{'s size} = 2).$$

Where *g(.)* is the "glue" function that will be presented in the next section.

For example suppose we have the following *n*-grams[4]:

A= O Hospital Distrital (The Regional Hospital);
B= O Hospital (The Hospital);
C= Hospital Distrital (Regional Hospital);
D= Hospital Distrital de (Regional Hospital of);
E= Hospital Distrital de Aveiro (Aveiro's Regional Hospital);
F= Hospital Distrital de Aveiro declarou (Aveiro's Regional Hospital declared);
G= Hospital Distrital de Santarém (Santarém's Regional Hospital);
H= Hospital Distrital de Santarém está (Santarém's Regional Hospital is).

Let us also supose that each *n*-gram has the following *g(.)* values[5]: *g(A) = .25; g(B) = .20; g(C) = .60; g(D) = .30; g(E) = .50; g(F) = .20; g(G) = .50* and *g(H) =.15*.

Fig. 1 depicts, the relative values of "glue" for these *n*-grams. It shows how *LocalMaxs* algorithm requires that the *g(.)* function works. For example, the *2*-gram *C* (in bold) would be considered a MWU by *LocalMaxs*, since it has an higher *g(.)* value than every *3*-grams for which *C* is an *antecedent*: *3*-grams *A* and *D*. Analogously, by definition 3.1, the *4*-grams *E* and *G* would also be MWUs because: $g(E) \geq g(D) \wedge g(E) > g(F)$ and $g(G) \geq g(D) \wedge g(G) > g(H)$. No other *n*-gram in Fig. 1 can be considered MWU by this algorithm.

Frequency alone is not important; it affects the "glue" value. For the above example and the *corpus* at stake, the real frequencies obtained are: *f(A) = 1, f(B) = 11, f(C) = 7, f(D) = 6, f(E) = 2, f(F) = 2, f(G) = 2* and *f(H) = 2*. Has we have seen, *G* is elected because it conforms the requirements of *Localmaxs* algorithm. However the higher value of *D* frequency does not prevent its unselectedness.

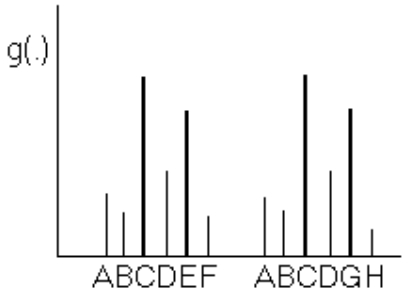

Fig. 1

4 Measuring the *n*-gram's "glue". The statistical approaches.

In this section we will review existing statistics-based measures for extraction of bigrams. Our measure (*SCP*) will also be presented. By applying the *pseudo-bigram transformation*, we generalize the application of each statistics-based measure to *n*-grams ($n \geq 2$). We will also present the *fair dispersion point normalization* for each of the association measures.

4.1 The Mutual information (SI) measure

In information theory, the mutual information *I(X,Y)* between two binary random variables *X* and *Y* is defined as:

$$I(X,Y) = \sum_{X \in \{x,\neg x\}} \sum_{Y \in \{y,\neg y\}} p(X,Y) \log\left(\frac{p(X;Y)}{p(X) \cdot p(Y)}\right) \quad (4.1)$$

[3]Since a MWU must be a relevant *n*-gram, the *LocalMaxs* does not produce MWUs with just one occurrence in the *corpus*. This criterion was applied just because it reduces drastically the processing time. The algorithm implements also a more subjective criterion: A MWU will be considered as such, only if it does not contain any punctuation sign ("," "." "!" ":" ";" etc.). However, this prevents us from selecting units such as U.S.A.

[4]By default in this paper, the *n*-grams working with our algorithm (*LocalMaxs)* are *hole-free n*-grams.

[5]These values were chosen for a better understanding of the *LocalMaxs* algorithm.

Where $p(X,Y)$, $p(X)$ and $p(Y)$ are the joint and marginal probability mass functions of the variables X and Y.

(Church & Hanks (1990)) introduced a measure of association ratio between two words, close to the concept of *Mutual information*. This measure is widely used e.g., (Daille (1995)), (Bahl et al. (1986)), (Church & Hanks (1990)), (Church et al.(1991)) and (Dagan & Marcus & Markovitch (1993)). It contains only a part of the above sum, the term where $X=x$ and $Y=y$, where x and y denote words, (not weighted by the joint probability distribution $p(X,Y)$). Although the term *Mutual Information* has been used in computational linguistics to refer this alternative measure based on *mutual information*, we prefer to call this measure *Specific mutual information* $SI((X,Y))$, as it is done in (Smadja (1996)).

$$SI((x,y)) = \log\left(\frac{p(x,y)}{p(x)\cdot p(y)}\right) \tag{4.2}$$

Where $p(x,y)$, $p(x)$ and $p(y)$ are the probabilities of the bigram x,y and unigrams x and y to be found in a given *corpus*.

4.1.1 SI n-gram generalization

Applying the *pseudo-bigram transformation*, we generalize *SI(.)* for *n*-grams. Then, we can choose for example the *dispersion point* between $w_1 \ldots w_{n-1}$ and w_n:

$$SI(((w_1..w_{n-1}),w_n)) = \log\left(\frac{p(w_1...w_n)}{p(w_1...w_{n-1})\cdot p(w_n)}\right) \tag{4.3}$$

In sub-section 5.3, we will present the results obtained for $SI((w_1,(w_2 \ldots w_n)))$ too, i.e., choosing the *dispersion point* between w_1 and $w_2 \ldots w_n$.

4.1.2 The fair dispersion point normalization. Its application to SI measure

If we take the *SI(.)* value as the value of *n*-gram's "glue", $g((w_1 \ldots w_n)) = SI(((w_1 \ldots w_{n-1}),w_n))$ from equation (4.3), it is expected that this value depends on the particular *dispersion point* considered above. As a matter of fact we would expect a different value if we had considered another *dispersion point*, for example between $w_1 \ldots w_{n-2}$ and w_{n-1},w_n. Then, one problem we must solve is related with which *dispersion point* must be chosen in order to have a fair measure of the *n*-gram's "glue". This problem can be solved by calculating the arithmetic average of the products determined by each *dispersion point* along the *n*-gram. In this way we can have a fair measure of the *n*-gram's "glue" as if the *n*-gram was made of a left and a right parts determined by a virtual *fair dispersion point* reflecting the whole *n*-gram's "glue". This is what we call *fair dispersion point normalization* or simply *fair dispersion*. We add this concept in equation (4.3) and name it *SI_f (f* for *fair)* in the context of this paper*:*

$$SI\_f((w_1..w_n)) = \log\left(\frac{p(w_1...w_n)}{Avp}\right) \tag{4.4}$$

Where:

$$Avp = \frac{1}{n-1}\cdot\sum_{i=1}^{i=n-1} p(w_1..w_i)\cdot p(w_{i+1}..w_n) \tag{4.5}$$

4.2 The ϕ^2 measure

The ϕ^2 *coefficient* was introduced by (Gale & Church (1991)) and has been widely used e.g., (Daille (1995)) and (Dunning (1993)). So, assuming we want to measure the "glue" of a bigram, we consider the following contingency table:

Table 1 - contingency table for the observed counts of each bigram

	Distribution of Y when X is present	Distribution of Y when X is not present
Distribution of X when Y is present	f(x,y)	f(¬x,y)
Distribution of X when Y is not present	f(x,¬y)	f(¬x,¬y)

Where $f(x,y)$ represents the absolute frequency of the bigram in which the first word is word x and the second word is y; $f(\neg x,y)$ represents the absolute frequency of the bigram in which the first word is not word x and the second is word y; etc..

So, considering this contingency table, we apply the ϕ^2 *coefficient*:

$$\phi^2((x,y)) = \frac{\left[f(x,y)\cdot N - f(x)\cdot f(y)\right]^2}{f(x)\cdot f(y)\cdot (N - f(x))\cdot (N - f(y))} \quad (4.6)$$

Where *f(x)* and *f(y)* are the absolute frequencies of the *1*-grams *x* and *y*. *N* is the number of words in the *corpus*.

4.2.1 Generalization for n-grams

For a generic *n*-gram, we apply it the *pseudo-bigram transformation* and we choose for example the *dispersion point* between $w_1 \ldots w_{n-1}$ and w_n.

$$\phi^2(((w_1 \ldots w_{n-1}), w_n)) = \frac{\left[f(w_1 \ldots w_n)\cdot N - P\right]^2}{P\cdot (N - f(w_1 \ldots w_{n-1}))\cdot (N - f(w_n))} \quad (4.7)$$

Where

$$P = f(w_1 \ldots w_{n-1})\cdot f(w_n)$$

4.2.2 The fair ϕ^2 measure

Acting as we did for *SI(.)* we can obtain a *n*-gram's "glue" value independent of any particular *dispersion point* by introducing the *fair dispersion* concept:

$$\phi^2\_f((w_1 \ldots w_n)) = \frac{\left[f(w_1 \ldots w_n)\cdot N - Avp\right]^2}{Avp\cdot (N - Avx)\cdot (N - Avy)} \quad (4.8)$$

where *Avp* is determined according to (4.5)

$$Avx = \frac{1}{n-1}\cdot \sum_{i=1}^{i=n-1} f(w_1 \ldots w_i)$$

$$Avy = \frac{1}{n-1}\cdot \sum_{i=2}^{i=n} f(w_i \ldots w_n) \quad (4.9)$$

Considering *f(x)* and *f(y)* in (4.6), in (4.9), we take the "average of *f(x)*" and the "average of *f(y)*", i.e. the absolute frequency of the *n*-gram's "average left part" and the absolute frequency of the *n*-gram's "average right part". The criterion we use is based on the arithmetic average.

4.3 The *Loglike* measure

The *Loglike coefficient* was introduced by (Dunning (1993)). In Dunning's work, the detection of *composite terms* is made by applying the likelihood ratio, phrasing the null hypothesis that *x* and *y* are independent as $p(x|y) = p(x|\neg y) = p(x)$ and using the binomial distribution.

$$Loglike((x,y)) = 2\cdot(\log l(p1,k1,n1) + \log l(p2,k2,n2) - \log l(p,k1,n1) - \log l(p,k2,n2)) \quad (4.10)$$

Where

$$\log l(P,K,M) = K\cdot\ln(P) + (M-K)\cdot\ln(1-P)$$

$$k1 = f(x,y) \quad k2 = f(x,\neg y) = f(x) - k1 \quad n1 = f(y)$$

$$n2 = N - n1 \quad p1 = p(x|y) = \frac{k1}{n1} \quad p2 = p(x|\neg y) = \frac{k2}{n2}$$

$$p = p(x) = \frac{k1 + k2}{N}$$

N still is the number of words in the *corpus*.

4.3.1 Generalization for n-grams

As we have done for $\phi^2(.)$, we may *transform* the *n*-gram in a *pseudo-bigram* substituting for example, *x* by $w_1 \ldots w_{n-1}$ and *y* by w_n in (4.10).

4.3.2 The fair Loglike measure

As we did previously, we can now introduce the *fair dispersion* concept in *Loglike(.)*:

$$Loglike\_f((w_1 \ldots w_n)) = 2\cdot(\log l(pf1,kf1,nf1) + \log l(pf2,kf2,nf2) - \log l(pf,kf1,nf1) - \log l(pf,kf2,nf2)) \quad (4.11)$$

Where

$$kf1 = f(w_1...w_n) \quad kf2 = Avx - kf1$$

$$pf = \frac{kf1 + kf2}{N} = Avx \quad nf1 = Avy$$

$$nf2 = N - nf1 \quad pf1 = \frac{kf1}{nf1} \quad pf2 = \frac{kf2}{nf2}$$

Avx and *Avy* are defined in (4.9)

4.4 The *Dice* measure

The *Dice coefficient* (Dice 1945)) is also widely used e.g., (Sörensen (1948)), (Salton & McGill (1983)), (Smadja (1996)) and (Frakes & Baeza-Yates (1992)). The results presented in (Smadja (1996)), led us to test this statistics-based measure in our algorithm. This measure of correlation is defined as:

$$Dice((x,y)) = \frac{2 \cdot f(x,y)}{f(x) + f(y)} \qquad (4.12)$$

4.4.1 Generalization for n-grams

For a generic *n*-gram we *transform* the *n*-gram in a *pseudo-bigram* again. We may substitute *x* by $w_1 \ldots w_{n-1}$ and *y* by w_n in (4.12).

4.4.2 The fair Dice measure

Once again, we can apply the *fair dispersion* concept and obtain a *Dice measure* independent of any particular *dispersion point*:

$$Dice((w_1...w_n)) = \frac{2 \cdot f(w_1...w_n)}{Avx + Avy} \qquad (4.13)$$

Avx and *Avy* are defined in (4.9).

4.5 A *Symmetrical Conditional Probability* measure

Let us consider the bigram *x,y*. We propose a new measure that tests the "correlation" between *x* and *y* by taking the conditional probabilities of each one given the other and multiply both terms. Let us call it *SCP (Symmetrical Conditional Probability)*:

$$SCP(x,y) = p(x \mid y).p(y \mid x) =$$

$$\frac{p(x,y)}{p(y)} \cdot \frac{p(x,y)}{p(x)} = \frac{p(x,y)^2}{p(x) \cdot p(y)} \qquad (4.14)$$

4.5.1 Generalization for n-grams

Transforming the *n*-gram in a *pseudo-bigram* and considering for example the *dispersion point* between $w_1 \ldots w_{n-1}$ and w_n, we can apply *SCP(.)* to a generic *n*-gram $w_1 \ldots w_n$:

$$SCP(((w_1...w_{n-1}),w_n)) = \frac{p(w_1...w_n)^2}{p(w_1...w_{n-1}) \cdot p(w_n)}$$

4.5.2 The fair SCP measure

Analogously, as we have done for *Si(.)*, $\phi^2(.)$ and *Loglike(.)*, we may change *SCP(.)* and get a new measure applying the *fair dispersion* concept:

$$SCP\_f((w_1...w_n)) = \frac{p(w_1...w_n)^2}{Avp} \qquad (4.15)$$

Avp is defined in (4.5).

5 Results, evaluation and discussion

In this section we will discuss and compare the results of the statistics-based measures presented in previous section by using our algorithm *LocalMaxs*.

5.1 The data

We have tested our algorithm with a *corpus* of 919,254 words. This *corpus* corresponds to the news from the *Lusa* News Agency[6] that have been broadcasted during January 1994. From this *corpus* we counted the following distinct *n*-grams:

[6] *Lusa* is the Portuguese news agency.

Table 2 - distribution of the distinct *n*-grams in *Lusa corpus*

1-grams	*2*-grams	*3*-grams	*4*-grams	*5*-grams	*6*-grams	*7*-grams	*8*-grams	Total
55408	325480	627061	772752	826246	849341	862421	871430	5190139

5.2 The results

We run the algorithm *LocalMaxs* taking into account the different kinds of "glues" we define taking each measure we have surveyed in the last section. Then according to table 2 and the algorithm definition in section 3, we have obtained MWUs from *2*-grams to *7*-grams.

5.3 The evaluation criterion

Although *LocalMaxs* produces many obvious collocations containing verbs (*tomar …decisão -to take…decision,* and others), we intend to take this type of units in future work for further refinements. For the purpose of validation of the results obtained we have concentrated our attention on *hole-free* multi-word nouns and frozen forms. Therefore, the evaluation of the *LocalMaxs* was done from all the MWUs produced which did not contain any verb. Each one of these was considered correct if it was a multi-word noun or a frozen form. Otherwise it was considered incorrect. In order to do this evaluation, we randomly took several hundreds of *n*-grams from the set of the selected *n*-grams by each statistics-based measure working with the *LocalMaxs* and checked each *n*-gram by hand.

Table 3 shows the results with the alternative statistics-based measures presented in section 4.

Table 3 - the scores for the statistics-based measures

Statistic-based measure: *g(.)=*	Precision (average)	Occurrences per MWU[7] (avg. count)	Extracted MWUs (count)
$\phi^2(.)^a$	47.50%	4.99	13023
$\phi^2(.)^b$	49.18%	5.33	12117
ϕ^2_f(.)	**83.33%**	**4.49**	**24741**
Loglike(.)[a]	41.33%	8.17	28457
Loglike(.)[b]	50.81%	8.22	28734
LogLike_f(.)	**51.66%**	**5.23**	**40614**
SI(.)[a]			0
SI(.)[b]			0
SI_f(.)	**81.80%**	**2.78**	**20918**
Dice(.)[a]	56.90%	5.18	15338
Dice(.)[b]	62.12%	5.48	18973
Dice_f(.)	**76.27%**	**4.89**	**32357**
SCP(.)[a]	64.17%	6.33	9437
SCP(.)[b]	59.64%	7.15	8932
SCP_f(.)	**84.90%**	**4.73**	**24431**

[7]This column gives the average number of occurrences per selected MWU for this particular *corpus (Lusa)*. Of course, this number changes with the size of the *corpus*.

[a]In this test we have used the *dispersion point* between $w_1…w_{n-1}$ and w_n. For example $\phi^2(.)^a$ means $\phi^2(((w_1…w_{n-1}),w_n))$.

[b]In this test we have used the *dispersion point* between w_1 and $w_2…w_n$. For example $\phi^2(.)^b$ means $\phi^2((w_1,(w_2…w_n)))$.

The first column of table 3 represents the statistics-based measure tested. The other columns contain data about each tested measure:

The *Precision* column means the average percentage of correct nouns obtained.

The number of *Occurrences per MWU (avg. count)* informs us how frequent (in average) are the MWUs extracted by the considered measure.

In this context, it is not possible to calculate the exact value of *recall* for each statistics-based measure, because we are not facing the problem of counting the multi-word nouns of a *corpus*, but counting the relevant ones, which seems to be a difficult task. So, considering that there is not a practical way for counting the relevant multi-word nouns in the *corpus*, the column *Extracted MWUs (count),* which gives the number of MWUs extracted by the considered measure, works as an indirect measure of *recall*. (Remind that we have discarded every MWU that occurred just once).

5.4 Discussion of usual and newly introduced statistics-based measures

As we can see from table 3, changing from $\phi^2(.)^a$ to $\phi^2(.)^b$ we note a little improvement in *precision* (from 47.50% to 49.18%), but we can not say that the *dispersion point* associated to $\phi^2(.)^b$ (between w_1 and $w_2 \ldots w_n$) is a "good" *dispersion point*. As a matter of fact, 49.18% is a relatively low score and *Extracted MWUs (count)* has decreased a little (from 13023 to 12117) -a very low score, as we can conclude comparing it with other scores in this column. We could also present another test ($\phi^2(.)^c$) based on another *dispersion point* somewhere in the middle of each *n*-gram, but the scores would be as low as for $\phi^2(.)^a$ and $\phi^2(.)^b$. By using the *dispersion points*, one between $w_1 \ldots w_{n-1}$ and w_n. *(a)* and another between w_1 and $w_2 \ldots w_{n,,}$ *(b)*, we enhance the utility of the *fair dispersion point normaliztion,* as it will be seen in this section. The *fair dispersion* is applicable to any syntactic pattern, compound nouns, compound prepositions, etc..

Changing from $\phi^2(.)^a$ or $^2(.)^b$ to $\phi^2\_f(.)$ (introducing the *fair dispersion),* we reached *83.33% precision* -a good score- and *Extracted MWUs (count)* also changes from 13023 reaching a relative good value: 24741.

We must have in mind that the average number of occurrences for each distinct *n*-gram in this *corpus* -given the *n* values we are considering: $2 \leq n \leq 7$- is 1.29[8]. So, for $\phi^2(.)^a$, $\phi^2(.)^b$ and $\phi^2\_f(.)$ measures, the *Occurrences per MWU (avg. count)* scores (4.99, 5.33 and 4.49) are good, meaning that the MWUs extracted by using these measures are relevant.

The introduction of the *fair dispersion* works as a general improvement for all the statistics-based measures: it improves a lot the E*xtracted MWUs* (*count*), and it presents a significant improvement in *Precision* for those measures except for *Loglike(.)* for which just a little increase (from 41.33% and 50.81%, to 51.66%) has been recorded. The small decrease of *Occurrences per MWU (avg. count)* value is due to the selection of new and less frequent MWUs when the *fair dispersion normalization* is introduced.

With Specific Mutual Information $SI(.)^a$ or $SI(.)^b$ we have got no MWUs; only the introduction of the *fair dispersion normalization* in this measure enabled the selection of MWUs. We must also remark the *Precision* value (81.8%) and a medium *Extracted MWUs count* (20918) obtained by this normalized measure *SI_f(.)*. The extracted MWUs it enabled are not as relevant / frequent as they are for other measures (2.78). In fact, the mutual information based measures tend to overestimate the significance of rare events (Dunning (1993)).

On the contrary, the relative frequency of the *n*-grams is highly valued in the case of *Loglike* based measures: 8.17, 8.22 and 5.23 for its *Occurrences per MWUs (avg. count)* scores. In fact, not all the frequent *n*-grams are so significant as compound nouns. For example *Loglike_f(.),* as well as $Loglike(.)^a$ and $Loglike(.)^b$, produce the following "MWUs":

ministro zimbabueano (Zimbabwean minister)
primeiro-ministro israelita Yitzack Rabin (israeli prime-minister Yitzack Rabin)
o ministro (the minister)
ministro dos (minister of the)

As a matter of fact, the 3rd and 4th *n*-grams are very frequent in the *corpus* but should not be considered MWUs per se.

[8]Computed for *Lusa* corpus.

Loglike measures (with or without *fair dispersion*) extracts a relatively higher number of MWUs (28457, 28734 and 40614), but *Precision* is low: 41.33%, 50.81% and 51.66%. Somehow, this result contradicts the claim made in (Dunning (1993)) about higher suitability of the *Loglike* measure to extract *composite terms* when compared with other measures. As a matter of fact, in the context of *LocalMaxs* using this medium size (*Lusa) corpus*, the ϕ^2 among others, gives better results.

The *Dice_f(.)* measure presents good values working with *LocalMaxs*: compared with ϕ^2*_f(.)* it has a lower *Precision* (76.27% instead of *83*.33%) but the relevance / occurrences per MWU is a little higher (4.89 instead of 4.49) as well as the number of *Extrated MWUs* (32357 instead of 24741). However, with this *Precision* (76.27%) we begin to see some produced "MWUs" such as "ministro dos" which is not a MWU. A higher *Precision* is desirable.

By introducing the *fair dispersion* in *SCP* measure (*SCP_f),* we reach the highest *Precision* (84.9%); the significance of the extracted MWUs (*Occurrences per MWU*) is kept high (4.73) and it points out too an interesting value for the indirect measure of *recall* (24431).

Considering the performance of these measures which use the *fair dispersion* in the context of the *LocalMaxs,* we can distinguish two groups: a group A having the *SCP_f,* ϕ^2*_f* and *SI_f* measures and a group B having the *Loglike_f* and *Dice_f* measures. The measures of the group A has the ability to assign high "glue" values just to the *n*-grams whose words belong exclusively or almost exclusively to the *n*-gram where those words are. In other words, the measures of the group A tends to penalize *n*-grams containing many words that are very common anywhere in *corpus*. In the context of the *LocalMaxs* this ability enables the algorithm to extract MWUs with good precision. The measures of the group B have not the same ability, mainly the *Loglike_f,* which assigns also high "glue" values to frequent *n*-grams, but many of those *n*-grams are not MWUs. This division was deeply tested in (Silva & Lopes & Xavier & Vicente (1999b)).

Considering the *LocalMaxs* algorithm definition (definition 3.1) and the fact that the *log* function is monotonously increasing, if the *log* was excluded from the *SI_f* measure (see definition 4.4), we would get the same results using that function with the *LocalMaxs* as we got using *SI_f*. So, in the context of the *LocalMaxs* we can see some similarity in the measures *SI_f* and *SCP_f* (see definition 4.15). However, the square of the probability of the *n*-gram's occurrence in *SCP_f* - instead of the simple probability of the *n*-gram's occurrence in *SI_f*- enables the extraction of smaller units, more frequent ones, and improves *Precision* and the number of extracted MWUs.

Despite the good scores of ϕ^2*_f(.),* the slightly higher values for *Precision* and *Occurrences per MWU* obtained with *SCP_f(.)* lead us to prefer this later. Considering the importance of these values in table 3, we believe that *SCP_f(.)* is the most suitable measure to work with our algorithm - among the measures presented here.

Some testes were made introducing the geometric average in the *fair dispersion* but the results were not so good as in the case of the arithmetic average. Appendix A presents a small sample of the *SCP_f* extracted MWUs.

5.5 Assessment of related work

In (Daille (1995)), a combined approach for automatic term extraction from a terminology *corpus* was presented. Starting with lemma pairs representing candidate terms selected by using morpho-syntactic patterns such as *N Adj*, *N1 de N2*, etc., some statistical scores are then applied. The terms of length greater than 2 are mainly created recursively from *base-terms* (*2*-grams or *3*-grams). This is done through the introduction of some morpho-syntactic operations: *overcomposition by juxtaposition and substitution*, *modification by insertion of modifiers and post-modification* (Daille (1995)). Clearly, these operations as well as the morpho-syntactic structure of the base terms depend on the language considered. The attained precision and recall are not presented in this Daille's work.

We believe that by using statistics to measure the whole *n*-gram's "glue", whatever the *n*-gram length is, rather than just measuring the 2-grams's "glue", perhaps the same or better results could have been obtained in a more comfortable way. With our algorithm we do not have to worry about the specificities (typology, etc.) of the language we are working with (French / Portuguese /

whatever), in order to obtain terms of length greater than 2: Those terms (MWUs) are obtained simply by using statistics without any complex morpho-syntactic operation. This was confirmed by the results we obtained for English, French and German in the European Parliament Discussions *corpora*.

In this work (Daille (1995)), the *Loglike* criterion is selected as the best. As Daille, we also think that this result is due to the fact that the statistics is applied after the linguistic filters. As a matter of fact *Loglike and Dice* measures gave rise to worst precision results for various *corpora* and for various languages, we have worked on.

In a paper for improving and evaluating system *Xtract* (Smadja (1991)), collocations are extracted from a *corpus*. This work is weakly related to ours, given that the collocations are based on pairs (not necessarily *hole-free* bigrams). Keeping the mixed approach (parsing and statistical) of the early version of *Xtract* (Smadja & McKeown (1990)), statistical measures are made just for bigrams. However, apart from the 80% precision and 94% recall, *Xtract* produces syntactic information associated to each collocation, which is a useful feature (for example, it adds the label "verb-object" to the collocation "make-decision").

In a paper for retrieving collocations (Shimohata & Sugio & Nagata (1997)), a co-occurrence measure for *n*-grams is presented. The approach is based on an entropy based measure applied to every size *n*-grams (not just bigrams) followed by some filtering based on thresholds. As it is said in section 1, we believe that avoiding thresholds is a good policy. The results produced by this approach tends to extract collocations terminating (and / or beginning) in very frequent words such as articles, prepositions, etc.: "For example, the", "The default is", ",you can use the", "These are", etc.. The precision obtained in the final stage of this work is 66.9%. We believe these results are important, but they show that this approach would get lower precision if it had been used for extracting just compound nouns and frozen forms.
In (Dunning (1993)), it is shown how the application of *Loglike* measure yields good results for relatively small samples, when compared with other statistical measures. However, our experiments show exactly the opposite. When the *corpus* size augments, the *Loglike* precision degrades. But Dunning's work is weakly related to ours, since the statistical measures are made only for *2*-grams and, despite the selection he makes for the most significant *2*-grams he obtains, we do not know which criterion he used for identifying the non significant *2*-grams.

In (Smadja & Hatzivassiloglou McKeown (1996)), a method for translating collocations implemented in *Champalion*, is presented. The *Champalion* takes parallel collocations previously extracted by *Xtract*, and translates the English collocations to French ones. In this work, the *Dice* statistical measure gave better results than *SI* measure for finding the association / co-occurrence between collocations in different languages. Once again we may observe that there is not an absolute statistical measure for all the purposes. In (Dias & Guilloré & Lopes (1999)) the algorithm presented here (*LocalMaxs)* is applied to the extraction of units with holes using another glue measure "Mutual Expectation". The results obtained for French are better than the results obtained by using the *Dice coefficient*, and confirm some of our conclusions.

In non-statistical approaches, patterns of occurrence that generally enable the retrieval of adequate compounds are searched. Generally do not go beyond *3*-grams. For example (Barkema (1993)) and (Barkema (1994)) search for those patterns in partially parsed corpora (treebanks), but recognize that the occurrence of a pattern does not necessarily mean that compound terms have been found. (Jacquemin & Royauté (1994)) also use this kind of pattern matching and then generate variations by inflection or by derivation and check if those possible units do appear in the *corpus* used. These approaches soon fall short of available tagged *corpora* and the precision their approaches enable are surely very low. They rely mostly on human selection.

6 Conclusion

Work on extraction of multi-word nouns, frozen forms, collocations, and / or multi-word indexes has been based mostly on statistical measurement of the *2*-grams frequency and more or less complex linguistic filters and specific morpho-syntactic operations. We propose a more robust approach, reducing the commitments associated to that complexity and take a greater advantage of

statistics. We have developed an algorithm (*LocalMaxs*) that extracts MWUs, by using just statistics. In this paper we have assessed the precision of extraction of multi-word nouns and frozen forms. Some other collocations such as compound verbs and compound prepositions are also produced but we did not elaborate on this subject matter here. Moreover we have just looked at MWUs that occur more than once.

Our approach measures the "glue" of each *n*-gram considering the whole *n*-gram whatever length it has. The algorithm detects the *n*-grams that must be elected under the criterion of the *LocalMaxs,* being an alternative to thresholds based approaches. Through a *pseudo-bigram transformation* we "normalize the *n*-gram's length" and achieve a "glue" value comparable for different size *n*-grams.

Through the introduction of the *fair dispersion point normalization,* we have obtained a fair measure of the *n*-gram's "glue", since the "glue" between every two adjacent words within the *n*-gram is reflected on it. This concept was tested for different statistical measures and gave rise to great improvements to every previously used glue measure.

In order to measure every *n*-gram's "glue" sticking its words together we investigated alternative statistical measures: *Specific mutual information (SI), Loglike, Dice* and ϕ^2, with and without the *fair dispersion normalization*. Despite of good results from *SI_f* and specially ϕ^2*_f,* in the context of *LocalMaxs* the best results were obtained with a measure we developed: *SCP_f*. With this measure, in *Lusa corpus* we reached 84.90% *precision* -higher than any other. The *relevance* / relative frequency of the extracted MWUs is also high (not far from the *Loglike_f* score). A good value for the indirect measure of *recall* was also obtained for this measure.

Recently we have been working also on an information retrieval project (PGR[9]) where our algorithm has produced good results by extracting the MWUs from the *PGR corpus,* and providing a more natural and comfortable interaction with the novice user, not knowing which descriptors have been assigned to each opinion of the Portuguese General Attorney. The same method has also been used in the General Chronicle of Spain written in Medieval Portuguese, and the results once again showed how this methodology can be an important support for lexicographers.

As future work we intend to apply the same methods to automatically tagged *corpora* (Marques & Lopes (1996)) and (Marques (1999)), and we will also work on the *LUSA* treebank that meanwhile we are constructing (Rocio & Lopes (1999)). We intend also to improve the performance of the *LocalMaxs* algorithm. One possibility is to compare the *fair dispersion point* smoothing with more standard smoothing methods and study how does *LocalMaxs* criterion hold in those cases. Other possibility will result from a comparative / differential study of the results obtained using the various methods reported.

Appendix A

A small sample of the *SCP_f* extracted MWUs (correct or incorrect) containing the word *Universidade*[10]:

Universidade Autodidacta (Self-taught University)
Universidade Nova (New University)
Universidade Tecnica (Technical University)
Universidade Técnica (Technical University)
Universidades Portuguesas (Portuguese Universities)
Associacao de Estudantes da Universidade do Algarve (Students Association of Algarve's University)
** cento dos estudantes da Universidade de Coimbra (cent of the students of the Coimbra's University)*
reitor da Universidade Nova de Lisboa (Rector of the New University of Lisbon)
Faculdade de Economia da Universidade Nova (New University's Faculty of Economics)
** académica da Universidade da Beira Interior (academic of the Beira Interior's University)*

[9]PGR means *Procuradoria Geral da República* (Portuguese Republic General Attorney). The application is accessible in http://www.pgr.pt)

[10]This sample presents ortographically misspelled terms. They appear in the *corpus* and are due to the fact that sometimes the available keyboards had not the ability to write diacritics. The extracted *n*-grams signed by an "*" are considered incorrect MWUs, dispite of its obvious strong glue sticking them together.

criação de uma Universidade de Bragança (creation of a Bragança's University)
dirigente da associação académica da Universidade (University's leader of the academic association)
reitor da Universidade de Aveiro (Rector of the Aveiro's University)
Associacao de Estudantes da Universidade (University's Students Association)
Associação de Estudantes da Universidade (University's Students Association)
Estudantes da Universidade do Algarve (Students of Algarve's University)
Hospitais da Universidade de Coimbra (Hospitals of Coimbra's University)
Reitoria da Universidade de Lisboa (Main office of the Lisbon's University)
** cento dos estudantes da Universidade (cent of the University's students)*
** uma Universidade de Bragança (a Bragança's University)*
**Economia da Universidade Nova (Economics of the New University)*
Universidade Clássica de Lisboa (Classic University of Lisbon)
Universidade Nova de Lisboa (New University of Lisbon)
Universidade da Beira Interior (Beira Interior's University)
associação académica da Universidade (University's academic association)
criação de uma Universidade (creation of a University)
Estudantes da Universidade (University's Students)
Hospitais da Universidade (University's Hospitals)
Reitores de Universidades (Universitys Presidents)
Universidade Católica Portuguesa (Portuguese Catholic University)
Universidade de Aveiro (Aveiro's University)
Universidade de Coimbra (Coimbra's University)
Universidade de Edimburgo (Edinburgh's University)
Universidade de Evora (Évora's University)
Universidade do Algarve (Algarve's University)
reitor da Universidade (University's rector)

Acknowledgements

We thank to Gaël Dias for many helpful discussion. We also thank *Agência Noticiosa Lusa* for the *corpus*. The work reported in this paper has been possible due to the funding received through the projects: "Córpora de Português Medieval, Etiquetagem e Segmentação Automática", Ref. PRAXIS 2/2.1/CSH/778/95; "DIXIT – Multilingual Intentional Dialog Systems", Ref. PRAXIS XXI 2/2.1/TIT/1670/95 and "PGR – Acesso Selectivo aos Pareceres da Procuradoria Geral da República" Ref. LO59-P31B-02/97. The information about this projects can be accessed by http://kholosso.di.fct.unl.pt/~di/people.phtml?it=CENTRIA&ch=gpl.

References

Bahl, Latif R. & Brown, Peter F. & de Sousa, Peter V. & and Mercer, Robert L. (1986), in *Maximum Mutual Information of Hidden Markov Model Parameters for Speech Recognition*. In Proceedings, International Conference on Acoustics, Speech, and Signal Processing Society, Institute of Electronics and Communication Engineers of Japan, and Acoustical Society of Japan.

Barkema H. (1993) *Idiomaticy in english Nps*, in Aarts J., de Haan P., Oostdijk N. (eds.): English language corpora: design, analysis and exploitation, Rodopi, Amsterdam, 1993, 257-278.

Barkema H. (1994) *Determining the syntactic flexibility of idioms*, in Fries U., Tottie G., Shneider P. (eds.): Creating and using English language corpora, Rodopi, Amsterdam, 1994, 39-52.

K. Church & K. Hanks (1990) in *Word Association Norms, Mutual Information and Lexicography*. Computational Linguistics, 16(1):22-29.

Church, Kenneth W. & Gale, William A. & Hanks, Patrick & Hindle, Donald. (1991) in *Using Statistical Linguistics in Lexical Analysis*. In Lexical Acquisition: Using On-line Resources to Build a Lexicon, edited by Uri Zernik. Lawrence Erlbaum, Hilldale, New Jersey, 115-165.

Dagan, Ido & Church, Kenneth W. & Gale, William A. (1993) in *Robust Bilingual Word*

Alignment for Machine-Aided Translation. In Proceedings, Workshop on Very Large Corpora: Academic and Industrial Perspectives, Columbus, Ohio, 1-8. Association for Computational Linguistics.

Daille, Béatrice (1995) in *Study and Implementation of Combined Techniques from Automatic Extraction of Terminology*. Chap. 3 of "The Balancing Act" – combining symbolic and statistical approaches to language, edited by Judith L. Klavans and Philip Resnik.

Gale, William A. & Church, Kenneth W. (1991) in *Concordance for parallel texts*. In Proceedings of the Seventh Annual Conference of the UW Centre of the new OED and Text Research, Using Corpora, pp.40-62, Oxford, 1991.

Dias, G. Guilloré, S. and Lopes, J. G. P. (1999) "*Language Independent Automatic Acquisition of Rigid Multiword Units from Unrestricted Text Corpora*", acceped to the Traitement Automatique des Langues Naturelles (TALN'99)

Dice, Lee (1945) in *Measures of the Amount of Ecologic Association between Species*. Journal of Ecology, 26: 297-302.

Dunning, Ted (1993) in *Accurate Methods for the Statistic of Surprise and Coincidence*. Association for Computational Linguistics, 19(1): 61-76 1993.

Frakes, Willian B. and Baeza-Yates, Ricardo, eds. (1992). *Information Retrieval: Data Structures and Algorithms*. Prentice Hall, Englewood Cliffs, New Jersey.

Jacquemin C., Royauté J. (1994) *Retrieving terms and their variants in a lexicalized unification-based framework*, in: SIGIR'94, Dublin, 1994 132-141.

Marques, Nuno Miguel & Lopes, J. G. P. (1996). *Using Neural Nets for Portuguese Part-of-Speech Tagging*. In Proceedings of the Fifth International Conference on The Cognitive Science of Natural Language. Dublin City University, Ireland September 1996.

Marques, Nuno Miguel (1999) "*Metodologia para a modelação estatística da subcategorização verbal*". Ph.D. Thesis. Universidade Nova de Lisboa, Faculdade de Ciências e Tecnologia, Lisbon, Portugal, Previewed presentation June 1999 (In Portuguese).

Rocio, V. J. & Lopes, J. G. P. (1999). *An infra-structure for diagnosing causes for partially parsed natural language input. 1999*. In "Actas of the VI Simposio Internacional de comunicacion Social Santiago de Cuba", Jan 99. Vol. 1 p. 550-554.

(Salton & McGill (1983)) Gerard Salton and McGill, Michael J. (1983). *Introduction to Modern Information Retrieval. MaGraw-Hill, New York*.

Sayori Shimohata, Sayori & Sugio, Toshiyuki & Nagata, Junji (1997) in *Retrieving Collocations by Co-occurrences and Word Order Constraints*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics.

Silva, J. Ferreira da & Lopes, G.P. L. & Xavier, M.F. & Vicente, G. (1999b) *Relevant Expressions in Large Corpora*. Accepted to "Atelier-TALN99" 1999.

Smadja, Frank A. & McKeown, Kathleen R. (1990) in *Automatically Extracting and Representing Collocations for Language Generation*. Proceedings, 28h Annual Meeting of the ACL, Berkeley, California, 279-284. Association for Computational Linguistics.

Smadja, Frank A. & Hatzivassiloglou, Valiseios & McKeown, Kathleen R. (1996) in *Translating Collocations for Bilingual Lexicons: A Statistical Approach*. Association for Computational Linguistics.

Smadja Frank A. (1991) in *From N-grams to Collocations: An evaluation of Extract*. In Proceedings, 29th Annual Meeting of the ACL, Berkeley, California, 279-284. Association for Computational Linguistics.

Sörensen Thorvald J. (1948). *A Method of Establishing Groups of Equal Amplitude in Plant Sociology Based on Similarity of Species Content and its application to Analysis of the Vegetation of Danish Commons*. Biologiske Skrifter, 5(4):1-34.