Multiword Unit Hybrid Extraction

Gaël Dias
Centre of Mathematics
Beira Interior University
Covilhã, Portugal
ddg@di.ubi.pt

Abstract

This paper describes an original hybrid system that extracts multiword unit candidates from part-of-speech tagged corpora. While classical hybrid systems manually define local part-of-speech patterns that lead to the identification of well-known multiword units (mainly compound nouns), our solution automatically identifies relevant syntactical patterns from the corpus. Word statistics are then combined with the endogenously acquired linguistic information in order to extract the most relevant sequences of words. As a result, (1) human intervention is avoided providing total flexibility of use of the system and (2) different multiword units like phrasal verbs, adverbial locutions and prepositional locutions may be identified. The system has been tested on the *Brown Corpus* leading to encouraging results.

1 Introduction

Multiword units (MWUs) include a large range of linguistic phenomena, such as compound nouns (e.g. *interior designer*), phrasal verbs (e.g. *run through*), adverbial locutions (e.g. *on purpose*), compound determinants (e.g. *an amount of*), prepositional locutions (e.g. *in front of*) and institutionalized phrases (e.g. *con carne*). MWUs are frequently used in everyday language, usually to precisely express ideas and concepts that cannot be compressed into a single word. As a consequence, their identification is a crucial issue for applications that require some degree of semantic processing (e.g. machine translation, summarization, information retrieval).

In recent years, there has been a growing awareness in the Natural Language Processing (NLP) community of the problems that MWUs pose and the need for their robust handling. For that purpose, syntactical (Didier Bourigault, 1993), statistical (Frank Smadja, 1993; Ted Dunning, 1993; Gaël Dias, 2002) and hybrid syntaxico-statistical methodologies (Béatrice Daille, 1996; Jean-Philippe Goldman *et al.* 2001) have been proposed.

In this paper, we propose an original hybrid system called HELAS[1] that extracts MWU candidates from part-of-speech tagged corpora. Unlike classical hybrid systems that manually pre-define local part-of-speech patterns of interest (Béatrice Daille, 1996; Jean-Philippe Goldman *et al.* 2001), our solution automatically identifies relevant syntactical patterns from the corpus. Word statistics are then combined with the endogenously acquired linguistic information in order to extract the most relevant sequences of words i.e. MWU candidates. Technically, we conjugate the Mutual Expectation (ME) association measure with the acquisition process called GenLocalMaxs (Gaël Dias, 2002) in a five step process. First, the part-of-speech tagged corpus is divided into two sub-corpora: one containing words and one containing part-of-speech tags. Each sub-corpus is then segmented into a set of positional ngrams i.e. ordered vectors of textual units. Third, the ME independently evaluates the degree of cohesiveness of each positional ngram i.e. any positional ngram of words and any positional ngram of part-of-speech tags. A combination of both MEs is then used to evaluate the global degree of cohesiveness of any sequence of words associated with its respective part-of-speech tag sequence. Finally, the GenLocalMaxs retrieves all the MWU candidates by evidencing local maxima of association measure values thus avoiding the definition of global thresholds. The overall architecture can be seen in Figure 1.

Compared to existing hybrid systems, the benefits of HELAS are clear. By avoiding human intervention in the definition of syntactical patterns, it provides total

[1] HELAS stands for *Hybrid Extraction of Lexical ASsociations*.

flexibility of use. Indeed, the system can be used for any language without any specific tuning. HELAS also allows the identification of various MWUs like phrasal verbs, adverbial locutions, compound determinants, prepositional locutions and institutionalized phrases. Finally, it responds to some extent to the affirmation of Benoît Habert and Christian Jacquemin (1993) that claim that "*existing hybrid systems do not sufficiently tackle the problem of the interdependency between the filtering stage* [*the definition of syntactical patterns*] *and the acquisition process* [*the scoring and the election of relevant sequences of words*] *as they propose that these two steps should be independent*".

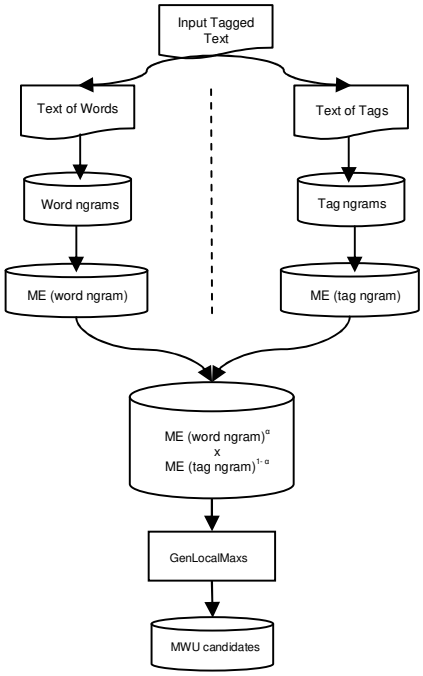

Figure 1: Global architecture of HELAS

The article is divided into five main sections: (1) we introduce the related work; (2) we present the text corpus segmentation into positional ngrams; (3) we define the Mutual Expectation and a new combined association measure; (4) we propose the GenLocalMaxs algorithm as the acquisition process; Finally, in (5), we present some results over the *Brown Corpus*.

2 Related Work

For the purpose of MWU extraction, syntactical, statistical and hybrid syntaxico-statistical methodologies have been proposed. On one hand, purely linguistic systems (Didier Bourigault, 1993) propose to extract relevant MWUs by using techniques that analyse specific syntactical structures in the texts. However, these methodologies suffer from their monolingual basis as the systems require highly specialised linguistic techniques to identify clues that isolate possible MWU candidates.

On the other hand, purely statistical systems (Frank Smadja, 1993; Ted Dunning, 1993; Gaël Dias, 2002) extract discriminating MWUs 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, these methodologies can only identify textual associations in the context of their usage. As a consequence, many relevant structures can not be introduced directly into lexical databases as they do not guarantee adequate linguistic structures for that purpose.

Finally, hybrid syntactico-statistical systems (Béatrice Daille, 1996; Jean-Philippe Goldman *et al.* 2001) define co-occurrences of interest in terms of syntactical patterns and statistical regularities. Thus, such systems reduce the searching space to groups of words that correspond to *a priori* defined syntactical patterns (e.g. *Adj+Noun*, *Noun+Prep+Noun*) and apply statistical scores to identify the most relevant sequences of words. One major drawback of such systems is that they do not deal with a great proportion of interesting MWUs (e.g. phrasal verbs, prepositional locutions). Moreover, they lack flexibility as the syntactical patterns have to be revised whenever the targeted language changes.

In order to overcome these difficulties, we propose an original architecture 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 assess that most of the MWUs evidence well-known morpho-syntactic structures (Gaston Gross, 1996). On the other hand, MWUs are recurrent combinations of words. Indeed, according to Benoît Habert and Christian Jacquemin (1993), the MWUs may represent a fifth of the overall surface of a text. Consequently, it is reasonable to think that the syntactical patterns embodied by the MWUs may be endogenously identified by using statistical scores over texts of part-of-speech tags exactly in the same manner as word dependencies are identified in corpora of words. So, the global degree of cohesiveness of any sequence of words may be evaluated by a combination of its degree of cohesiveness of words and the degree of cohesiveness of its associated part-of-speech tag sequence (See Figure 1).

Compared to existing systems, the benefits of our architecture are clear. By avoiding human intervention in the definition of syntactical patterns, (1) HELAS provides total flexibility of use being independent of the targeted

language and (2) it allows the identification of various MWUs like phrasal verbs, adverbial locutions, compound determinants, prepositional locutions and institutionalized phrases.

3 Text Segmentation

Positional ngrams are nothing more than ordered vectors of textual units which principles are introduced in the next subsection.

3.1 Positional Ngrams

The original idea of the positional ngram model (Gaël Dias, 2002) comes from the lexicographic evidence that most lexical relations associate words separated by at most five other words (John Sinclair, 1974). As a consequence, lexical relations such as MWUs can be continuous or discontinuous sequences of words in a context of at most eleven words (i.e. *5* words to the left of a pivot word, *5* words to the right of the same pivot word and the pivot word itself). In general terms, a MWU can be defined as a specific continuous or discontinuous sequence of words in a (*2.F+1*)-word size window context (i.e. *F* words to the left of a pivot word, *F* words to the right of the same pivot word and the pivot word itself). This situation is illustrated in Figure 2 for the multiword unit *Ngram Statistics* that fits in the window context of size *2.3+1=7*.

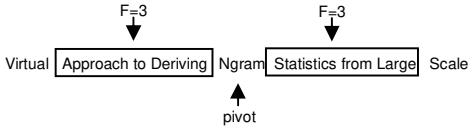

Figure 2: *7*-word size window context

Thus, any substring (continuous or discontinuous) that fits inside the window context and contains the pivot word is called a positional word ngram. For instance, [Ngram Statistics] is a positional word ngram as is the discontinuous sequence [Ngram ___ from] where the gap represented by the underline stands for any word occurring between Ngram and from (in this case, Statistics). More examples are given in Table 1.

Positional word 2grams	Positional word 3grams
[Ngram Statistics]	[Ngram Statistics from]
[Ngram ___ from]	[Ngram Statistics ___ Large]
[Ngram ___ ___ Large]	[Ngram ___ from Large]
[to ___ Ngram]	[to ___ Ngram ___ from]

Table 1: Possible positional ngrams

Generically, any positional word ngram may be defined as a vector of words $[p_{11}\ u_1\ p_{12}\ u_2\ \ldots\ p_{1n}\ u_n]$ where u_i stands for any word in the positional ngram and p_{1i} represents the distance that separates words u_1 and u_i[2]. Thus, the positional word ngram [Ngram Statisitcs] would be rewritten as [0 Ngram +1 Statistics]. More examples are given in Table 2.

Positional word ngrams	Algebraic notation
[Ngram ___ from]	[0 Ngram +2 from]
[Ngram ___ ___ Large]	[0 Ngram +3 Large]
[to ___ Ngram]	[0 to +2 Ngram]
[Ngram Statistics ___ Large]	[0 Ngram +1 Statisitcs +3 Large]

Table 2: Algebraic Notation

However, in a part-of-speech tagged corpus, each word is associated to a unique part-of-speech tag. As a consequence, each positional word ngram is linked to a corresponding positional tag ngram. A positional tag ngram is nothing more than an ordered vector of part-of-speech tags exactly in the same way a positional word ngram is an ordered vector of words. Let's exemplify this situation. Let's consider the following portion of a part-of-speech tagged sentence following the Brown tag set:

Virtual /JJ Approach /NN to /IN Deriving /VBG Ngram /NN Statistics /NN from /IN Large /JJ Scale /NN Corpus /NN

It is clear that the corresponding positional tag ngram of the positional word ngram [0 Ngram +1 Statisitcs] is the vector [0 /NN +1 /NN]. More examples are in Table 3. Generically, any positional tag ngram may be defined as a vector of part-of-speech tags $[p_{11}\ t_1\ p_{12}\ t_2\ \ldots\ p_{1n}\ t_n]$ where t_i stands for any part-of-speech tag in the positional tag ngram and p_{1i} represents the distance that separates the part-of-speech tags t_1 and t_i.

Positional word ngrams	Positional tag ngrams
[0 Ngram +2 from]	[0 /NN +2 /IN]
[0 Ngram +3 Large]	[0 /NN +3 /JJ]
[0 to +2 Ngram]	[0 /IN +2 /NN]
[0 Ngram +1 Statisitcs +3 Large]	[0 /NN +1 /NN +3 /JJ]

Table 3: Positional tag ngrams

So, any sequence of words, in a part-of-speech tagged corpus, is associated to a positional word ngram and a corresponding positional tag ngram. In order to introduce the part-of-speech tag factor in any sequence of words of part-of-speech tagged corpus, we present an alternative notation of positional ngrams called positional word-tag ngrams.

In order to represent a sequence of words with its associated part-of-speech tags, a positional ngram may be represented by the following vector of words and part-

[2] By statement, any p_{ii} is equal to zero.

of-speech tags [p_{11} u_1 t_1 p_{12} u_2 t_2… p_{1n} u_n t_n] where u_i stands for any word in the positional ngram, t_i stands for the part-of-speech tag of the word u_i and p_{1i} represents the distance that separates words u_1 and u_i. Thus, the positional ngram [Ngram Statistics] can be represented by the vector [0 Ngram /NN +1 Statistics /NN] given the text corpus in section (3.1). More examples are given in Table 4.

Positional ngrams	Alternative notation
[Ngram ___ from]	[0 Ngram /NN +2 from /IN]
[Ngram ___ ___ Large]	[0 Ngram /NN +3 Large /JJ]
[to ___ Ngram]	[0 to /IN +2 Ngram /NN]

Table 4: Alternative Notation

This alternative notation will allow us to defining, with elegance, our combined association measure, introduced in the next section.

3.2 Data Preparation

So, the first step of our architecture deals with segmenting the input text corpus into positional ngrams. First, the part-of-speech tagged corpus is divided into two sub-corpora: one sub-corpus of words and one sub-corpus of part-of-speech tags. The word sub-corpus is then segmented into its set of positional word ngrams exactly in the same way the tagged sub-corpus is segmented into its set of positional tag ngrams.

In parallel, each positional word ngram is associated to its corresponding positional tag ngram in order to further evaluate the global degree of cohesiveness of any sequence of words in a part-of-speech tagged corpus. Our basic idea is to evaluate the degree of cohesiveness of each positional ngram independently (i.e. the positional word ngrams on one side and the positional tag ngrams on the other side) in order to calculate the global degree of cohesiveness of any sequence in the part-of-speech tagged corpus by combining its respective degrees of cohesiveness i.e. the degree of cohesiveness of its sequence of words and the degree of cohesiveness of its sequence of part-of-speech tags.

In order to evaluate the degree of cohesiveness of any sequence of textual units, we use the association measure called Mutual Expectation.

4 Cohesiveness Evaluation

The Mutual Expectation (ME) has been introduced by Gaël Dias (2002) and evaluates the degree of cohesiveness that links together all the textual units contained in a positional ngram ($\forall n, n \geq 2$) based on the concept of Normalized Expectation and relative frequency.

4.1 Normalized Expectation

The basic idea of the Normalized Expectation (NE) is to evaluate the cost, in terms of cohesiveness, of the loss of one element in a positional ngram. Thus, the NE is defined in Equation 1 where the function $k(.)$ returns the frequency of any positional ngram[3].

$$NE([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n]) = \frac{k([p_{11}u_1 \ldots p_{1i}u_i \ldots p_{1n}u_n])}{\frac{1}{n}\left(k([p_{22}u_2 \ldots p_{2i}u_i \ldots p_{2n}u_n]) + \sum_{i=2}^{n} k\left(\left[p_{11}u_1 \ldots \hat{p}_{1i}\ \hat{u}_i \ldots p_{1n}u_n\right]\right)\right)}$$

Equation 1: Normalized Expectation

In order to exemplify the NE formula, we present in Equation 2 its development for the given positional ngram [0 A +2 C +3 D +4 E] where each letter may represent a word or a part-of-speech tag.

$$NE([0\,A, 2\,C, 3\,D, 4\,E]) = \frac{k([0\,A, 2\,C, 3\,D, 4\,E])}{\frac{1}{4}\begin{pmatrix} k([0\,A, 2\,C, 3\,D]) + \\ k([0\,A, 2\,C, 4\,E]) + \\ k([0\,A, 3\,D, 4\,E]) + \\ k([0\,C, 1\,D, 2\,E]) \end{pmatrix}}$$

Equation 2: Normalized Expectation example

However, evaluating the average cost of the loss of an element is not enough to characterize the degree of cohesiveness of a sequence of textual units. The Mutual Expectation is introduced to solve this insufficiency.

4.2 Mutual Expectation

Many applied works in Natural Language Processing have shown that frequency is one of the most relevant statistics to identify relevant textual associations. For instance, in the context of multiword unit extraction, (John Justeson and Slava Katz, 1995; Béatrice Daille, 1996) assess that the comprehension of a multiword unit is an iterative process being necessary that a unit should be pronounced more than one time to make its comprehension possible. Gäel Dias (2002) believes that this phenomenon can be enlarged to part-of-speech tags. From this assumption, they pose that between two positional ngrams with the same NE, the most frequent positional ngram is more likely to be a relevant sequence.

So, the Mutual Expectation of any positional ngram is defined in Equation 3 based on its NE and its relative frequency embodied by the function $p(.)$.

[3] The "^" corresponds to a convention used in Algebra that consists in writing a "^" on the top of the omitted term of a given succession indexed from 1 to n.

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

Equation 3: Mutual Expectation

We will note that the ME shows interesting properties. One of them is the fact that it does not sub-evaluate interdependencies when frequent individual textual units are present. In particular, this allows us to avoid the use of lists of stop words. Thus, when calculating all the positional ngrams, all the words and part-of-speech tags are used. This fundamentally participates to the flexibility of use of our system.

As we said earlier, the ME is going to be used to calculate the degree cohesiveness of any positional word ngram and any positional tag ngram. The way we calculate the global degree of cohesiveness of any sequence of words associated to its part-of-speech tag sequence, based on its two MEs, is discussed in the next subsection.

4.3 Combined Association Measure

The drawbacks shown by the statistical methodologies evidence the lack of linguistic information. Indeed, these methodologies can only identify textual associations in the context of their usage. As a consequence, many relevant structures can not be introduced directly into lexical databases as they do not guarantee adequate linguistic structures for that purpose.

In this paper, we propose a first attempt to solve this problem without pre-defining syntactical patterns of interest that bias the extraction process. Our idea is simply to combine the strength existing between words in a sequence and the evidenced interdependencies between its part-of-speech tags. We could summarize this idea as follows: the more cohesive the words of a sequence and the more cohesive its part-of-speech tags, the more likely the sequence may embody a multiword unit.

This idea can only be supported due to two assumptions. On one hand, a great deal of studies in lexicography and terminology assess that most of the MWUs evidence well-known morpho-syntactic structures (Gaston Gross, 1996). On the other hand, MWUs are recurrent combinations of words capable of representing a fifth of the overall surface of a text (Benoît Habert and Christian Jacquemin, 1993). Consequently, it is reasonable to think that the syntactical patterns embodied by the MWUs may endogenously be identified by using statistical scores over texts of part-of-speech tags exactly in the same manner as word dependencies are identified in corpora of words. So, the global degree of cohesiveness of any sequence of words may be evaluated by a combination of its own ME and the ME of its associated part-of-speech tag sequence. The degree of cohesiveness of any positional ngram based on a part-of-speech tagged corpus can then be evaluated by the combined association measure (CAM) defined in Equation 4 where α stands as a parameter that tunes the focus whether on words or on part-of-speech tags.

$$CAM([p_{1 1} u_1 t_1 ... p_{1 i} u_i t_i ... p_{1 n} u_n t_n]) =$$
$$ME([p_{1 1} u_1 ... p_{1 i} u_i ... p_{1 n} u_n])^{\alpha} \times ME([p_{1 1} t_1 ... p_{1 i} t_i ... p_{1 n} t_n])^{1-\alpha}$$

Equation 4: Combined Association Measure

We will see in the final section of this paper that different values of α lead to fundamentally different sets of multiword unit candidates. Indeed, α can go from a total focus on part-of-speech tags (i.e. the relevance of a word sequence is based only on the relevance of its part-of-speech sequence) to a total focus on words (i.e. the relevance of a word sequence is defined only by its word dependencies). Before going to experimentation, we need to introduce the used acquisition process which objective is to extract the MWUs candidates.

5 The Acquisition Process

The GenLocalMaxs (Gaël Dias, 2002) proposes a flexible and fine-tuned approach for the selection process as it concentrates on the identification of local maxima of association measure values. Specifically, the GenLocalMaxs elects MWUs from the set of all the valued positional ngrams based on two assumptions. First, the association measures show that the more cohesive a group of words is, the higher its score will be. Second, MWUs are localized associated groups of words. So, we may deduce that a positional word-tag ngram is a MWU if its combined association measure value is higher or equal than the combined association measure values of all its sub-groups of (n-1) words and if it is strictly higher than the combined association measure values of all its super-groups of (n+1) words. Let *cam* be the combined association measure, *W* a positional word-tag ngram, Ω_{n-1} the set of all the positional word-tag (n-1)-grams contained in *W*, Ω_{n+1} the set of all the positional word-tag (n+1)-grams containing *W* and *sizeof(.)* a function that returns the number of words of a positional word-tag ngram. The GenLocalMaxs is defined as:

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

Definition 1: GenLocalMaxs Algorithm

Among others, the GenLocalMaxs shows one important property: it does not depend on global thresholds. A

direct implication of this characteristic is the fact that, as no tuning needs to be made in order to acquire the set of all the MWU candidates, the use of the system remains as flexible as possible. Finally, we show the results obtained by applying HELAS over the *Brown Corpus*.

6 The Experiments

In order to test our architecture, we have conducted a number of experiments with 11 different values of α for a portion of the *Brown Corpus* containing 249 578 words i.e. 249 578 words plus its 249 578 part-of-speech tags. The limited size of our corpus is mainly due to the space complexity of our system. Indeed, the number of computed positional ngrams is huge even for a small corpus. For instance, 21 463 192 positional ngrams are computed for this particular corpus for a 7-word size window context. As a consequence, computation is hard. For this experiment, HELAS has been tested on a personal computer with 128 Mb of RAM, 20 Gb of Hard Disk and an AMD 1.4 Ghz processor under Linux Mandrake 7.2. On average, each experiment (i.e. for a given α) took 4 hours and 20 minutes. Knowing that our system increases proportionally with the size of the corpus, it was unmanageable, for this particular experiment, to test our architecture over a bigger corpus. Even though, the whole processing stage lasted almost 48 hours[4].

We will divide our experiment into two main parts. First, we will do a quantitative analysis and then we will lead a qualitative analysis. All results will only tackle contiguous multiword units although non-contiguous sequences may be extracted. This decision is due to the lack of space.

6.1 Quantitative Analysis

In order to understand, as deeply as possible, the interaction between word cohesiveness and part-of-speech tag cohesiveness, we chose eleven different values for α, i.e. $\alpha \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, going from total focus on words (α = 1) to total focus on part-of-speech tags (α = 0).

First, we show the number of extracted contiguous MWU candidates by α in table 5. The total results are not surprising. Indeed, with α = 0, the focus is exclusively on part-of-speech tags. It means that any word sequence, with an identified relevant part-of-speech sequence, is extracted independently of the words it contains. For instance, all the word sequences with the pattern [/JJ /NN] (i.e. *Adjective* + *Noun*) may be extracted independently of their word dependencies! This obviously leads to an important number of extracted sequences. The inclusion of the word factor, by increasing the value of α, progressively leads to a decreasing number of extracted positional ngrams. In fact, the word sequences with relevant syntactical structures are being filtered out depending on their word statistics. Finally, with α = 1, the focus is exclusively on words. The impact of the syntactical structure is null and the positional ngrams are extracted based on their word associations. In this case, the word sequences do not form classes of morpho-syntactic structures being the reason why less positional ngrams are extracted.

alpha	0	0.1	0.2	0.3	0.4	0.5
2gram	23146	21890	20074	17689	15450	13461
3gram	297	467	567	351	1188	1693
4gram	86	108	127	163	225	326
5gram	79	81	81	82	77	82
6gram	62	57	56	57	56	58
TOTAL	23670	22603	20905	18342	16996	15620
alpha	0.6	0.7	0.8	0.9	1.0	
2gram	11531	9950	9114	8650	8465	
3gram	2147	2501	2728	2828	2651	
4gram	428	557	679	740	484	
5gram	93	112	128	161	145	
6gram	58	58	60	64	60	
TOTAL	14257	13178	12709	12443	11805	

Table 5: Number of extracted MWU candidates

A deeper analysis of table 5 reveals interesting results. The smaller the values of α, the more positional 2grams are extracted. This situation is illustrated in Figure 3.

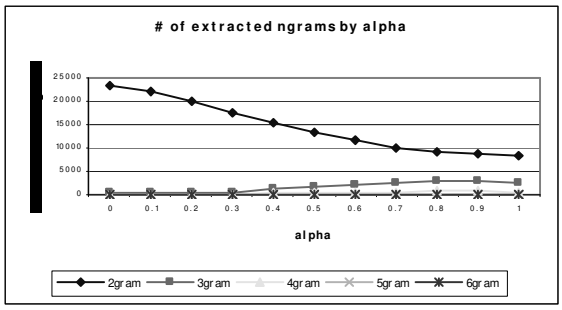

Figure 3: Number of extracted MWU candidates

Once again these results are not surprising. The Mutual Expectation tends to give more importance to frequent sequences of textual units. While it performs reasonably well on word sequences, it tends to over-evaluate the part-of-speech tag sequences. Indeed, sequences of two part-of-speech tags are much more frequent than other types of sequences and, as a consequence, tend to be over-evaluated in terms of cohesiveness. As small values of α focus on syntactical structures, it is clear that in this case, small sequences of words are preferred over longer sequences.

[4] We are already working on an efficient implementation of HELAS using suffix-arrays and the concept of masks.

By looking at Figure 3 and Table 5, we may think that a great number of extracted sequences are common to each experiment. However, this is not true. In order to assess this affirmation, we propose, in Table 6, the summary of the *identical ratio*.

alphas	0	0.1	0.2	0.3	0.4	0.5
0		14.64	5.74	2.99	1.73	1.17
0.1			9.99	3.77	2.08	1.35
0.2				6.2	2.83	1.69
0.3					4.89	2.36
0.4						5.31
0.5						
alphas	0.6	0.7	0.8	0.9	1.0	
0	0.83	0.63	0.54	0.49	0.47	
0.1	0.93	0.70	0.59	0.54	0.52	
0.2	1.11	0.81	0.68	0.61	0.59	
0.3	1.42	0.98	0.81	0.72	0.69	
0.4	2.34	1.44	1.13	0.97	0.90	
0.5	4.77	2.26	1.62	1.33	1.17	
0.6		5.06	2.82	2.10	1.73	
0.7			7.21	3.99	2.81	
0.8				9.45	4.50	
0.9					7.71	
1.0						

Table 6: Identical Ratio

The identical ratio calculates, for two values of α, the quotient between the number of identical extracted sequences and the number of different extracted sequences. Thus, the first value of the first row of table 6, represents the identical ratio for α=0 and α=0.1, and means that there are 14.64 times more identical extracted sequences than different sequences between both experiments.

Taking α=0 and α=1, it is interesting to see that there are much more different sequences than identical sequences between both experiments (identical ratio = 0.47). In fact, this phenomenon progressively increases as the word factor is being introduced in the combined association measure to reach α=1. This was somewhat unexpected. Nevertheless, this situation can be partly decrypted from Figure 3. Indeed, figure 3 shows that longer sequences are being preferred as α increases. In fact, what happens is that short syntactically well-founded sequences are being replaced by longer word sequences that may lack linguistic information. For instance, the sequence [Blue Mosque] was extracted with α=0, although the longer sequence [the Blue Mosque] was preferred with α=1 as whenever [Blue Mosque] appears in the text, the determinant [the] precedes it.

Finally, a last important result concerns the frequency of the extracted sequences. Table 7 gives an overview of the situation. The figures are clear. Most of the extracted sequences occur only twice in the input text corpus. This result is rather encouraging as most known extractors need high frequencies in order to decide whether a sequence is a MWU or not. This situation is mainly due to the GenLocalMaxs algorithm.

alpha	0	0.1	0.2	0.3	0.4	0.5
Freq=2	13555	13093	12235	11061	10803	10458
Freq=3	4203	3953	3616	3118	2753	2384
Freq=4	1952	1839	1649	1350	1166	960
Freq=5	1091	1019	917	743	608	511
Freq>2	2869	2699	2488	2070	1666	1307
TOTAL	23670	22603	20905	18342	16996	15620
alpha	0.6	0.7	0.8	0.9	1.0	
Freq=2	10011	9631	9596	9554	9031	
Freq=3	2088	1858	1730	1685	1678	
Freq=4	766	617	524	485	468	
Freq=5	392	276	232	202	189	
Freq>2	1000	796	627	517	439	
TOTAL	14257	13178	12709	12443	11805	

Table 7: Number of extracted MWUs by frequency

6.2 Qualitative Analysis

As many authors assess (Frank Smadja, 1993; John Justeson and Slava Katz, 1995), deciding whether a sequence of words is a multiword unit or not is a tricky problem. For that purpose, different definitions of multiword unit have been proposed. One of the most successful attempts can be attributed to Gaston Gross (1996) that classifies multiword units into six groups and provides techniques to determine their belonging. As a consequence, we intend as multiword unit any compound noun (e.g. interior designer), compound determinant (e.g. an amount of), verbal locution (e.g. run through), adverbial locution (e.g. on purpose), adjectival locution (e.g. dark blue) or prepositional locution (e.g. in front of).

The analysis of the results has been done *intramuros* although we are aware that an external independent cross validation would have been more suited. However, it was not logistically possible to do so and by using Gaston Gross's classification and methodology, we narrow the human error evaluation as much as possible. Technically, we have randomly extracted and analysed 200 positional 2grams, 200 positional 3grams and 100 positional 4grams for each value of α. For the specific case of positional 5grams and 6grams, all the sequences have been analysed.

Precision results of this analysis are given in table 8 and show that word dependencies and part-of-speech tag dependencies may both play an important role in the identification of relevant sequences. Indeed, values of α between 0.4 and 0.5 seem to lead to optimum results. Knowing that most extracted sequences are positional 2grams or positional 3grams, the global precision results approximate the results given by 2grams and 3grams. In these conditions, the best results are for α=0.5 reaching an average precision of 62 %. This would mean that

word dependencies and part-of-speech tags contribute equally to multiword unit identification.

alpha	0	0.1	0.2	0.3	0.4	0.5
2gram	29 %	22 %	30 %	44 %	53 %	60 %
3gram	52 %	77 %	74 %	73 %	80 %	85 %
4gram	38 %	32 %	32 %	46 %	47 %	41 %
5gram	34 %	28 %	29 %	31 %	33 %	34 %
6gram	29 %	22 %	18 %	24 %	31 %	38 %
alpha	0.6	0.7	0.8	0.9	1.0	
2gram	45 %	23 %	25 %	18 %	30 %	
3gram	43 %	35 %	46 %	51 %	36 %	
4gram	41 %	45 %	39 %	44 %	37 %	
5gram	27 %	27 %	29 %	38 %	38 %	
6gram	32 %	37 %	26 %	29 %	29 %	

Table 8: Precision in % by alpha

A deeper look at the results evidences interesting regularities as shown in figure 4. Indeed, the curves for 4grams, 5grams and 6grams are reasonably steady along the X axis evidencing low results. This means, to some extent, that that our system does not seem to be able to tackle successfully multiword units with more than three words. In fact, neither a total focus on words or on part-of-speech tags seems to change the extraction results. However, the importance of these results must be weakened as they represent a small proportion of the extracted structures.

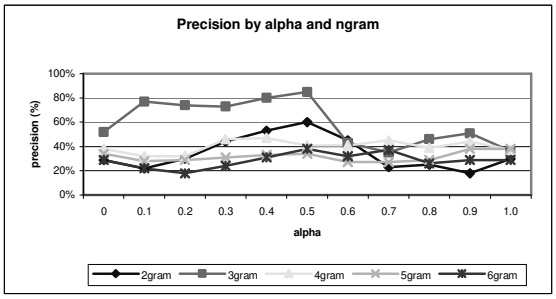

Figure 4: Precision by alpha and ngram

On the other hand, the curves for 2grams and 3grams show different behaviours. For the 3gram case, it seems that the syntactical structure plays an important role in the identification process. Indeed, precision falls down drastically when the focus passes to word dependencies. This is mainly due to the extraction of recurrent sequences of words that do not embody multiword unit syntactical structures like [been able to] or [can still be]. As 2grams are concerned, the situation is different. In fact, it seems that too much focus on either words or part-of-speech tags leads to unsatisfactory results. Indeed, optimum results are obtained for a balance between both criteria. This result can be explained by the fact that there exist many recurrent sequences of two words in a corpus. However, most of them are not multiword units like [of the] or [can be]. For that reason, only a balanced weight on part-of-speech tag and word dependencies may identify relevant two word sequences. However, not-so-high precision results show that two-word sequences still remain a tricky problem for our extractor as it is difficult to filter out very frequent patterns that embody meaningless syntactical structures.

7 Conclusion

This paper describes an original hybrid system that extracts multiword unit candidates by endogenously identifying relevant syntactical patterns from the corpus and by combining word statistics with the acquired linguistic information. As a result, by avoiding human intervention in the definition of syntactical patterns, (1) HELAS provides total flexibility of use being independent of the targeted language and (2) it allows the identification of various MWUs like compound nouns, compound determinants, verbal locutions, adverbial locutions, prepositional locutions and adjectival locutions without defining any threshold or using lists of stop words. The system has been tested on the *Brown Corpus* leading to encouraging results evidenced by a precision score of 62 % for the best configuration. The system will soon be available on http://helas.di.ubi.pt.

References

Béatrice Daille. 1996. *Study and Implementation of Combined Techniques for Automatic Extraction of Terminology*. The balancing act combining symbolic and statistical approaches to language, MIT Press, 49-66.

Benoît Habert and Chistian Jacquemin. 1993. *Noms composés, termes, dénominations complexes: problématiques linguistiques et traitements automatiques*. Traitement Automatique des Langues, vol. 34(2), 5-41.

Didier Bourigault. 1993. *Analyse syntaxique locale pour le repérage de termes complexes dans un texte*. Traitement Automatique des Langues, vol. 34 (2), 105-117.

Frank Smadja. 1993. *Retrieving collocations from text: XTRACT*. Computational Linguistics, vol. 19(1), 143-177.

Gaël Dias. 2002. *Extraction Automatique d'Associations Lexicales à partir de Corpora*. PhD Thesis. DI/FCT New University of Lisbon (Portugal) and LIFO University of Orléans (France).

Gaston Gross. 1996. *Les expressions figées en français*. Paris, Ophrys.

Jean-Philippe Goldman, Luka Nerima and Eric Wehrli. 2001. *Collocation Extraction using a Syntactic Parser*. Workshop of the 39th Annual Meeting of the Association for Computational Linguistics on Collocation: Computational Extraction, Analysis and Exploitation, Toulouse, France, 61-66.

John Justeson and Slava Katz. 1995. Technical Terminology: some linguistic properties and an algorithm for identification in text. Natural Language Engineering, vol. 1, 9-27.

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

Ted Dunning. 1993. *Accurate Methods for the Statistics of Surprise and Coincidence*. Computational Linguistics, vol. 19(1), 61-74.