Representation and Treatment of Multiword Expressions in Basque

Iñaki Alegria, **Olatz Ansa**, **Xabier Artola**
Nerea Ezeiza, **Koldo Gojenola** and **Ruben Urizar**
Ixa Group
University of the Basque Country
649 pk E-20.080
Donostia. Basque Country
rubenu@sc.ehu.es

Abstract

This paper describes the representation of Basque Multiword Lexical Units and the automatic processing of Multiword Expressions. After discussing and stating which kind of multiword expressions we consider to be processed at the current stage of the work, we present the representation schema of the corresponding lexical units in a general-purpose lexical database. Due to its expressive power, the schema can deal not only with fixed expressions but also with morphosyntactically flexible constructions. It also allows us to lemmatize word combinations as a unit and yet to parse the components individually if necessary. Moreover, we describe HABIL, a tool for the automatic processing of these expressions, and we give some evaluation results. This work must be placed in a general framework of written Basque processing tools, which currently ranges from the tokenization and segmentation of single words up to the syntactic tagging of general texts.

1 Introduction

Most texts are rich in multiword expressions, which must be necessarily processed if we want any NLP tool to perform accurately. Jackendoff (1997) estimates that their number in the speakers' lexicon "is of the same order of magnitude as the number of single words".

There is no agreement among authors about the definition of the term Multiword Expression. However, in this article, Multiword Expressions (hereafter MWE) refer to *any* word combinations ranging from idioms, over proper names, compounds, lexical and grammatical collocations… to institutionalized phrases. MWEs comprise both semantically compositional and non-compositional combinations, and both syntactically regular and idiosyncratic phrases, including complex named entities such as proper nouns, dates and number expressions (see section 2).

In contrast, Multiword Lexical Units (hereafter MWLU) comprise lexicalized phrases — semantically non-compositional or syntactically idiosyncratic word combinations— which are represented and stored in the lexical database of Basque (EDBL).

The remaining sections are organized as follows. Section 2 presents the main features of MWEs in Basque, and defines which are currently considered for automatic processing. Section 3 describes the representation of MWLUs in the lexical database. Section 4 is devoted to the description and evaluation of the automatic treatment of MWEs by means of HABIL. Section 5 summarizes future work. And, finally, section 6 outlines some conclusions.

2 Multiword Expressions in the processing of real texts in Basque

The definition of the term Multiword Expression and the types of such MWEs to be treated in NLP may vary considerably depending on the purposes or "the depth of processing being undertaken" (Copestake *et al.,* 2002). *Multiword* itself is a vague term. At text level, a word could be defined as "any string of characters between two blanks" (Fontenelle *et al.*, 1994). This is not applicable to languages as Japanese, which are typically written without spaces. Besides, a great number of MWEs that in uninflected languages would be multiword, constitute a single typographic unit in agglutinative languages such as Basque (*ziurrenik* 'most probably', *aurrerantzean* 'from now on', *aurretiaz* 'in advance'). Therefore, we consider them single words and they are included in the lexical database as such (or recognized by means of morphological analysis).

In our case, when deciding which Basque MWEs to include in the database, we mostly rely on lexicographers' expertise since we consider *lexicalized phrases* have a top priority for both lemmatizing and syntactic purposes. So, the MWEs dealt with in the database comprise *fixed expressions*, which admit no morphosyntactic or internal modification —including foreign expressions such as *in situ*, *a priori*, *strictu sensu*, etc.—, *idioms,* both decomposable and non-decomposable, and lexicalized *compounds*. We also consider light verb constructions when they are syntactically idiosyncratic.

However, currently we do not treat open collocations, proverbs, catch phrases and similes. Mostly, we don't include proper names in the database either, since complex named entities are given a separate treatment. Apart from proper nouns, also dates and number expressions are treated separately (see 4.1).

So far we have described 2,270 MWLUs in our database. This work has been carried out in two phases. For the first phase, we made use of the *Statistical Corpus of 20th Century Basque* (http://www.euskaracorpusa.net) that contains about 4.7 million words. As a starting point, we chose the MWLUs that occurred more than 10 times in this manually lemmatized corpus. This amounted to about 1,300 expressions. For the second phase, this list has been enlarged using the *Hiztegi Batua*, a dictionary of standard Basque that the Basque Language Academy updates regularly (http://www2.euskaltzaindia.net/hiztegibatua).

2.1 Main features of lexicalized phrases

Many of the lexicalized phrases are semantically non-compositional (or partially compositional), i.e. they can hardly be interpreted in terms of the meaning of their constituents (*adarra jo* 'to pull someone's leg', literally 'to play the horn').

Often, a component of these sequences hardly occurs in any other context and it is difficult to assign it a part of speech. For example, the word *noizik* is an archaism of modern *noiztik* 'from when', which occurs just in the expressions *noizik behin*, *noizik behinean*, *noizik noizera*, and *noizik behinka* all meaning 'once in a while'. Besides, it is not clear which is the part of speech of the words *laprast* in *laprast egin* 'to slip' or *dir-dir* in *dir-dir egin* 'to shine'.

From a syntactic point of view, many of these MWEs present an unusual structure. For example, many complex verbs in Basque are light verb constructions, being the meaning of the compound quite compositional, e.g. *lo egin* 'to sleep' literally 'to make (a) sleep' or *lan egin* 'to work' literally 'to make (a) work'. However, *lo egin* and *lan egin* can be considered 'syntactically idiomatic' since the nouns in these expressions, *lo* and *lan*, take no determiner, which would be completely ungrammatical for a noun functioning as a regular direct object (**arroz jan nuen* 'I ate rice').

Morphosyntactic flexibility, being significant in this type of constructions in Basque, may vary considerably. For example in *lo egin* 'to sleep' the noun *lo* admits modification (*lo* ***asko*** *egin zuen* 'he slept very much') and may take the partitive assignment (*ez dut lo****rik*** *egin* 'I haven't slept') while the verb *egin* can be subject to focalization (*egin duzu lorik bart?* 'did you sleep at all last night?'); besides, the components of the construction may change positions and some elements and phrases may be placed between them (*mendian* ***egin*** *omen zuen lasai* ***lo*** 'it is said that he slept peacefully in the mountain'). In contrast, *alde egin* 'to escape' is morphosyntactically quite rigid. In all the cases, the verb *egin* can take any inflection.

For our database, we have worked out a single representation that covers all MWLUs ranging from fixed expressions to these of highest morphosyntactic flexibility.

3 Representation of MWLUs in the lexical database

In this section we explain how MWLUs are represented in EDBL (Aldezabal *et al.*, 2001), a lexical database oriented to language processing that currently contains more than 80,000 entries, out of which 2,270 are MWLUs. Among these:

- ~69% are always unambiguous. The average number of Surface Realization Schemas (SRS, see section 3.2) is 1.02.
- ~23% are sometimes unambiguous and have 3.6 SRSs in average, half of them ambiguous.
- ~8% are always ambiguous and have 1.2 SRSs in average.

We want to point out that almost all of the unambiguous MWLUs have only one SRS, their components appearing in contiguous positions and always in the same order. About half of them are inflected, so, even if we discard the interpretations of the components, there is still some morphosyntactic ambiguity left. However, the identification of these MWLUs helps in disambiguation, as the input of tagging is more precise.

The description of MWLUs within a general-purpose lexical database must include, at least, two aspects (see Figure 1): (1) their *composition*, i.e. which the components of the MWLU are, whether each of them can be inflected or not, and according to which one-word lexical unit (OWLU [1]) it inflects; and (2), what we call the *surface realization*, that is, the order in which the components may occur in the text, the mandatory or optional contiguousness of components, and the inflectional restrictions applicable to each one of the components.

3.1 Composition

As it has just been said, the description of the composition of MWLUs in EDBL gathers two aspects: on the one side, it depicts which the individual components of a MWLU are; on the other side, it links the inflectable components of a MWLU to the corresponding OWLU according to which each of them inflects.

In Figure 1, we can see that the `composed of` relationship links every MWLU to up to 9 individual components (`MWLU_Components`). Each component is characterized by the following attributes:

- `Component_Position`: this indicates the position of the component word-form in the canonical form of the MWLU.
- `Component_Form`: i.e. the word-form itself as it appears in the canonical form of the MWLU.
- `Conveys_Morph_Info?`: this is a Boolean value, indicating whether the component inflection conveys the morphological information corresponding to the whole MWLU or not[2].

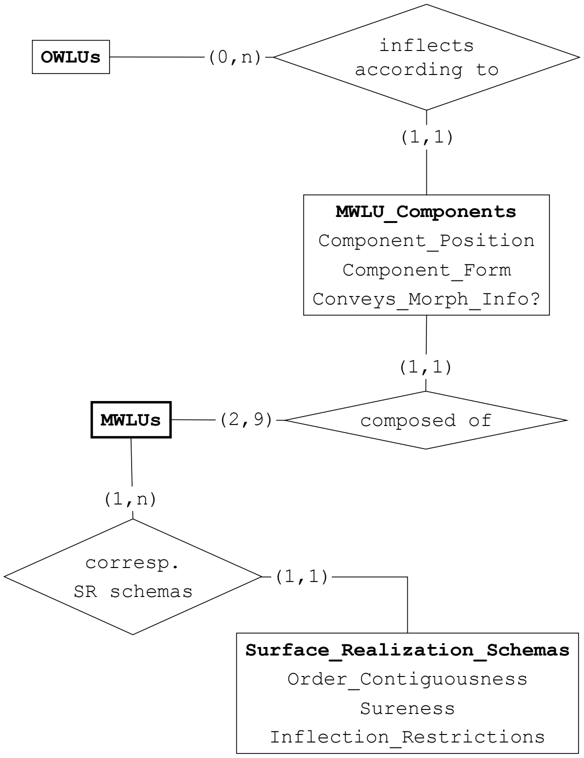

Figure 1. Composition and surface realization of MWLUs.

Moreover, the components of a MWLU are linked to its corresponding OWLU (according to which it inflects). This is represented by means of the `inflects according to` relationship (see Figure 1).

[1] We consider OWLUs lexical units with no spaces within its orthographical form; so, we also take hyphenated compounds as OWLUs.

[2] The morphological information that the attribute refers to is the set of morphological features the inflection takes in the current component instance.

These two aspects concerning the composition of a MWLU are physically stored in a single table of the relational database in which EDBL resides.

The columns of the table are the following: Entry, Homograph_Id, Component_Position, Component_Form, Conveys_Morph_Info?, OWLU_Entry, and OWLU_Homograph_Id. In the example below, the composition of the MWLU *begi bistan egon* 'to be evident' is described. Note that one row is used per component:

```
<begi bistan egon, 0, 1, begi, -, begi, 2>
<begi bistan egon, 0, 2, bistan, -, bista, 1>
<begi bistan egon, 0, 3, egon, +, egon, 1>
```

This expression allows different realizations such as *begi bistan dago* 'it is evident' (literally 'it is at reach of the eyes')*, begi bistan daude* 'they are evident'*, begien bistan egon*, 'to be evident', etc. In the table rows above, it can be seen that the last component *egon 1* 'to be' conveys the morphological information for the whole MWLU (+ in the corresponding column).

3.2 Surface realization

As for surface realization, we have already mentioned that the components of a MWLU can occur in a text either contiguously or dispersed. Besides, the order of the constituents may be fixed or not, and they may either inflect or occur in an invariable form. In the case of inflected components, some of them may accept any inflection according to its corresponding OWLU, whilst others may only inflect in a restricted way. Moreover, some MWLUs are unambiguous and some are not, since it cannot be certainly assured that the very same sequence of words in a text corresponds undoubtedly to a multiword entry in every context. For example, in the sentence *Emilek buruaz baiezko keinu* ***bat egin*** *zuen* 'Emile nodded his head' the words *bat* and *egin* do not correspond to the complex verb *bat egin* 'to unite' but to two separate phrases.

According to these features, we use a formal description where different realization patterns may be defined for each MWLU. The corresp. SR schemas relationship in Figure 1 links every MWLU to one or more Surface_Realization_Schemas. Each SRS is characterized by the following attributes:

- Order_Contiguousness: an expression that indicates both the order in which the components may appear in the different instances of the MWLU and the contiguousness of these components. In these expressions the position of the digits indicate the position each component takes in a particular SRS, * indicates that 0 or more words may occur between two components, and ? indicates that at most one single word may appear between two given components of the MWLU.
- Unambiguousness: a Boolean value, indicating whether the particular SRS corresponds to an unambiguous MWLU or not. It expresses whether the sequence of words matching this SRS must be unambiguously analyzed as an instance of the MWLU or, on the contrary, may be analyzed as separate OWLUs in some contexts.
- Inflection_Restrictions: an expression that indicates the inflection paradigm according to which the MWLU may inflect in this specific SRS. In these expressions each component of the MWLU is represented by one list component (in the same order as the components of the MWLU appear in its canonical form): % indicates that the whole inflection paradigm of the corresponding inflectable component may occur; the minus sign (-) is used for non-inflectable components (no inflection at all may occur); finally, a logical expression (and, or, and not are allowed) composed of attribute-value pairs is used to express the inflectional restrictions and the morphotactics the component undergoes in this particular SRS of the MWLU (in brackets in the examples below).

In the examples below, it can be seen that one row is used per SRS. The columns of the table are the following: Entry, Homograph_Id, Order_Contiguousness, Unambiguousness, and Inflection_Restrictions:

```
<begi bistan egon, 0, 123, +,
  (((CAS=ABS) and (DEF=-)) or
  ((CAS=GEN) and (NUM=PL)), -, %)>
```

```
<begi bistan egon, 0, 312, +,
 (((CAS=ABS) and (DEF=-)) or
  ((CAS=GEN) and (NUM=PL)), -, %)>
<begi bistan egon, 0, 3?12, +,
 (((CAS=ABS) and (DEF=-)), -, %)>
```

The first SRS matches occurrences such as ***begi bistan dago*** *hau ez dela aski* 'it is evident that it is not enough' or ***begien bistan zegoen*** *honela bukatuko genuela* 'it was evident that we would end up this way', where the components are contiguous and the analysis as an instance of the MWLU would be unambiguous. This SRS allows the inflection of the first component as absolutive case (non-definite) or as genitive (plural), and the whole set of inflection morphemes of the third one.

The third SRS matches occurrences such as *ez* ***dago*** *horren* ***begi bistan*** 'it is not so evident', where the components are not contiguous (at most one word is allowed between the “third” component and the “first one”) and they occur in a non-canonical order: `3?12`. In this case, the interpretation as an instance of the MWLU would also be unambiguous. However, this SRS only allows the inflection of the first component as absolutive case (non-definite).

3.3 Different information requirements in lemmatization and syntax processing

The first prototype for the treatment of MWEs in Basque HABIL (Ezeiza *et al.*, 1998; Ezeiza, 2003) was built for lemmatization purposes. However, we are nowadays involved in the construction of a deep syntactic parser (Aduriz *et al.*, 2004) and the MWEs seem to need a different treatment. The fact that many MWEs may be syntactically regular but, above all, that an external element may have a dependency relation with one of the constituents, forces us to analyze the elements independently. For example, in the verb *beldur izan* 'to be afraid (of)' an external noun phrase may have a modifier-noun dependency relation with *beldur* 'fear' as in *sugeen beldur naiz* 'I'm afraid of snakes'. In *loak hartu* 'to fall asleep' there is a subject-verb relation as in *loak hartu nau* 'I have fallen asleep', literally 'sleep has caught me'; therefore subject-auxiliary verb agreement would fade if both components were analyzed as one.

The MWLU representation we have adopted allows us to lemmatize the word combination as a unit and yet to parse the components individually whenever necessary. In order to do so, when describing each MWLU, we specify whether the elements in the MWLU must be analyzed separately or not[3].

4 Treatment of multiword expressions

MWEs could be treated at different stages of the language process. Some approaches treat them at tokenization stage, identifying fixed phrases, such as prepositional phrases or compounds, included in a list (Carmona *et al.,* 1998; Karlsson *et al.,* 1995). Other approaches rely on morphological analysis to better identify the features of the MWE using finite state technology (Breidt *et al.,* 1996). Finally, there is another approach that identifies them after the tagging process, allowing the correction of some tagging errors (Leech *et al.,* 1994).

All of these approaches are based on the use of a closed set of MWLUs that could be included in a list or a database. However, some groups of MWEs are not subject to be included in a database, because they comprise an open class of expressions. That is the case of collocations, compounds or named entities. The group of collocations and compounds should be delimited using statistical approaches, such as Xtract (Smadja, 1993) or LocalMax (Silva *et al.,* 1999), so that only the most relevant—those of higher frequency— are included in the database.

Named entity recognition task has been solved for a large set of languages. Most of these works are linked to the Message Understanding Conference (Chinchor, 1997). There is a variety of methods that have been used in NE recognition, such as HMM, Maximum Entropy Models, Decision Trees, Boosting and Voted Perceptron (Collins, 2002), Syntactic Structure based approaches and WordNet-based approaches (Magnini *et al.,* 2002; Arévalo, 2002). Most references on NE task might be accessed at http://www.muc.saic.com.

4.1 Processing MWEs with HABIL

We have implemented HABIL, a tool for the treatment of multiword expressions (MWE), based

[3] Currently we are studying the MWLUs in the lexical database in order to determine which of them deserve to be parsed as separate elements. We have not defined yet how this will be formally represented in the database.

on the features described in the lexical database. The most important features of HABIL are the following:

- It deals with both contiguous and split MWEs.
- It takes into account all the possible orders of the components (SRS).
- It checks that inflectional restrictions are complied with.
- It generates morphosyntactic interpretations for the MWE.

This tool has two different components: on the one hand, there is a searching engine that identifies MWEs along the text, and, on the other hand, there is a morphosyntactic processor that assigns the corresponding interpretations to the components of the MWE.

The morphosyntactic processor generates the interpretations for MWEs using category and subcategory information in the lexical database. When one of the components adds information to the MWE, the processor applies pattern-matching techniques to extract the corresponding morphological features of the analyses of that component, and these features are included in the interpretation of the MWE. Then, it replaces all the morphosyntactic interpretations of the components of unambiguous MWEs with the MWE interpretations. When MWEs are ambiguous, the new interpretations are added to the existing ones.

HABIL also identifies and treats dates and numerical expressions. As they make up an open class, they are not obviously included in the lexical database. Furthermore, their components are always contiguous, have a very strict structure, and use a closed lexicon. Thus, it is quite easy to identify them using simple finite state transducers. For the morphosyntactic treatment of dates and numerical expressions, we use the morphosyntactic component of HABIL. These expressions may appear inflected and, in this case, the last component adds morphosyntactic features to the MWE. Finally, as they are unambiguous expressions, the processor discards the interpretations of the components and assigns them all the interpretations of the whole expression.

4.2 Evaluation

We performed several experiments using 650 unambiguous, contiguous and ordered MWEs. We treated a reference corpus of around 36,000 tokens and there were 386 instances of 149 different MWEs. We also applied this process to a small test corpus of around 7,100 tokens in which there were 87 instances of 45 MWEs. Taking both corpora into account, there were 473 instances of 167 different MWEs, which amounted to 25% of the expressions considered, and 50% of the instances were ambiguous. Besides, only 14 dates and 12 numerical expressions were found in the reference corpus, and 18 dates and 9 numerical expressions in the test corpus.

		Ambiguity Rate	Interpretations per Token	Recall
word-forms:	before	81.78%	3.37	99.31%
	after	79.83%	3.30	99.31%
all tokens:	before	67.47%	2.96	99.43%
	after	65.86%	2.89	99.43%

Table 1. Results of HABIL.

The ambiguity measures of the test corpus are shown in Table 1. The ambiguity rate of word-forms decreases by 2% and the average ambiguity rate by 1.5% after the processing of MWEs. It is important to point out that no error is made along the process. Furthermore, some important MWEs, more specifically, some complex sentence connectors that have highly ambiguous components, are correctly disambiguated.

Bearing in mind the proportion of words treated by HABIL, these results help significantly in improving precision results of tagging and avoiding almost 10% of the errors, as shown in Table 2.

	Precision	Error
before MWE processing	94.96%	5.04%
after MWE processing	95.42%	4.58%

Table 2. Tagging results.

5 Future work

After confirming the viability of the system and the good results in POS tagging, our main goal is to increase the number of MWLUs in the database, which will improve the identification of MWEs in corpora.

A remaining difficulty that we are facing is the problem of ambiguous split MWEs. At present, we are creating a disambiguation grammar that will discard or select the multiword interpretations in ambiguous MWLUs. We are developing similar rules using both the Constraint Grammar formalism and finite state transducers (XFST tools, Kartunnen *et al*. 1997). The very first rules seem to be quite effective. Soon, we will be assessing the first results, and then we will be able to choose the method that performs best with a lesser effort. Once we have chosen the best formalism, we intend to develop a comprehensive grammar that will disambiguate as many ambiguous MWLUs as possible.

In addition, we are developing new processes after POS tagging in order to identify complex named entities and terminological units. These units constitute an open class and so their exhaustive inclusion in a database would not be viable.

6 Conclusion

In this paper we have described a whole framework for the representation and treatment of MWEs, which is being currently used at the IXA Research Group to process this kind of expressions in general texts. Although it has been conceived and so far used for Basque, a highly inflected language, we think that it is general enough to be applied to other languages.

A general representation schema for MWLUs at the lexical level has been proposed. This schema allows us to state which components a MWLU has and to formally encode all the different surface realizations it can adopt in the text.

The problems that diverse information requirements in lemmatization and syntactic processing can eventually pose have been explained, and a possible solution for the representation of these phenomena has also been outlined.

As for the processing aspects, we have described HABIL, the tool for the treatment of MWEs. HABIL processes MWEs based on their description in the lexical database, dealing also with some types of open class MWEs.

One of the remaining problems when split and ambiguous MWEs are to be tagged is related with disambiguation procedures using Hidden Markov Models, which are not able to manage different paths with variable lengths. This problem can be solved using rule-based methods or lattice structures for tagging.

7 Acknowledgements

This research is being partially funded by the European Commission (MEANING project, IST-2001-34460) and the Basque Government (Etortek-Hizking, Saiotek-Ihardetsi).

References

Aduriz I., Aranzabe M., Arriola J., Díaz de Ilarraza A., Gojenola K., Oronoz M., Uria L. 2004. A cascaded syntactic analyser for Basque. *Fifth International Conference on Intelligence Text Processing and Computational Linguistics* (CICLing2004). Seoul, Korea.

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza N., Hernández G., Lersundi M. 2001. EDBL: a General Lexical Basis for the Automatic Processing of Basque. *IRCS Workshop on Linguistic Databases.* Philadelphia.

Arévalo M. 2002. MICE, un recurso para la resolución de la anáfora. *International Workshop on Computational Linguistics*. *http://www.lsi.upc.es/~nlp/iwcl02*.

Breidt E., Segond F., Valetto G. 1996. Local grammars for the description of multi-word lexemes and their automatic recognition in texts. *Proceedings of COMPLEX'96*, 19-28. Budapest.

Carmona J., Cervell S., Màrquez L., Martí M.A., Padró L., Placer R., Rodríguez H., Taulé M., Turmo J. 1998. An environment for morphosyntactic processing of unrestricted Spanish text. *Proceedings of LREC'98*. 915-922.

Chinchor N. 1997. MUC-7 Named Entity Task Definition. Version 3.5. *http://www.itl.nist.gov/iaui/894.02/related_projects/muc/*

Collins M. 2002. Ranking Algorithms for Named-Entity Extraction: Boosting and the Voted Percetron. *Proceedings of ACL-2002.*

Collins M., Singer Y. 1999. Unsupervised Models for Named Entity Classification. *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Workshop on Very Large Corpora (EMNLP-VLC-99).*

Copestake A., Lambean F., Villavicencio A., Bond F., Baldwin T., Sag I., Flickinger D. 2002. Multiword Expressions: linguistic precision and reusability.

Proceedings of the Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, pp. 1941-7.

Ezeiza N. 2003. *Corpusak ustiatzeko tresna linguistikoak. Euskararen etiketatzaile sintaktiko sendo eta malgua*. PhD thesis, University of the Basque Country.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998. Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *COLING-ACL'98*, Montreal (Canada).

Fontenelle T., Adriaens G., De Braekeleer G. 1994. *The Lexical Unit in the Metal® MT System*, *MT*. The Netherlands. v9. 1-19.

Jackendoff R. 1997. *The Architecture of the Language Faculty.* Cambridge, MA MIT Press.

Karlsson F., Voutilainen A., Heikkila J,. Anttila A. 1995. *Constraint Grammar: A Language-independent System for Parsing Unrestricted Text*. Mouton de Gruyter.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A. 1997. *Regular expressions for language engineering.* Natural Language Engineering, Cambridge University Press.

Leech G., Garside R., Bryan M. 1994. CLAWS4: The tagging of the British National Corpus. *Proceedings of COLING-94*, 622-628.

Magnini B., Negri M., Prevete R., Tanev H. 2002. A WordNet Approach to Named Entities Recognition. *Proceeding of the Workshop SemaNet'02: Binding and Using Semantic Networks*.

Silva J., Dias G., Guilloré S., Lopes G. 1999. Using localmaxs algorithm for the extraction of contiguous and non-contiguous multiword lexical units. *Proceedings of 9th Portuguese Conference in Artificial Inteligence*, 21-24.

Smadja F. 1993. Retrieving Collocations from Text: Xtract. *Computational Linguistics,* 19(1), 143-177.