EUSLEM: A lemmatiser/tagger for Basque

Itziar Aduriz
UZEI
Izaskun Aldezabal / Iñaki Alegria / Xabier Artola /
Nerea Ezeiza / Ruben Urizar
Computer Science Faculty of University of the Basque Country

Abstract

This paper presents relevant issues that have been considered in the design and development of a general purpose lemmatiser/tagger for Basque (EUSLEM). The lemmatiser/tagger is conceived as a basic tool for other linguistic applications. It uses the lexical database and the morphological analyser previously developed and implemented. We will describe the components used in the development of the lemmatiser/tagger and, finally, we will point out possible further applications of this tool.

1. Introduction

An automatic lemmatiser/tagger is a basic tool for applications such as automatic indexation, documental databases, syntactic and semantic analysis, analysis of text corpora, etc. Its job is to give the correct lemma of a text-word, as well as its grammatical category.

This project is being carried out by two entities: a group of the Computer Science Faculty of The Basque Country University and UZEI (1), an association that works on Basque terminology and lexicography. It's not the first time both entities collaborate in a project, adding the Computer Science Faculty's research task on computational linguistics to UZEI's lexicographic experience. In fact, they have productively been working together during the last years and some fruits of their collaboration have already come to light, e.g. the spelling corrector for Basque (Agirre *et al.*, 1992).

The background of this project is the Systematic Compilation of Modern Basque (EEBS) project, sponsored by the Basque Government and local institutions and carried out by UZEI (Urkia, Sagarna, 1991). The aim of this project was to compile and semiautomatically lemmatise a three million word corpus of 20th century's Basque texts. During the 1987-1992 period UZEI has created a new database consisting of three million of lemmatised words, which is being annually renewed. The lemmatiser/tagger will be a great relief for the lemmatisation of this corpus, as well as for the lemmatisation of the General Basque Dictionary corpus (2).

This paper consists of three main parts: we will first make a brief description of Basque morphology and some problematic cases from the point of view of the lemmatiser. In the second part we will deal with the tagger and the need for a clear tagset. In the third and most important section we will discuss the different components in which EUSLEM consists. Eventually, we will make a brief exposition of the current situation of the project.

2. Brief description of Basque morphology

Basque is an agglutinative language, that is, for the formation of words the dictionary entry independently takes each of the elements necessary for the different functions (syntactic case included). More specifically, the affixes corresponding to the determinant, number and case are taken in this order and independently of each other (see Fig. 1).

seme	A	r	EN	*etxe*	A	N
noun ('son')	determinant	ephentetical element	genitive case	noun ('house')	determinant	inessive case

Fig. 1.- Analysis of *semearen etxean* ('in the house of the son')

<u>Morphological ambiguity in Basque</u>

In order to cope with morphological ambiguity, we distinguish the following main types:

- <u>Categorial</u> (or part-of-speech) ambiguity, like Noun/Verb, Verb/Adjective/Adverb, etc. The rate of categorial ambiguity is around 0.31 (1.38 analysis for each word-form).
- <u>Morphosyntactic ambiguity</u> There are several possible morphosyntactic interpretations attached to each input word-form (see Fig. 2).

gizonak -> Absolutive Plural 'the men' *or* Ergative Singular 'the man'

Fig. 2.- Two morphosyntactic interpretations of the word *gizonak*

The borderlines between morphology and syntax are not clear, because the information attached to each analysis contains features belonging to both morphology and syntax (see Fig. 3).

ikusiaz:	Morphological level	Syntactic level
	Instrumental case 'by (Noun) seen'	modal case 'seeing'

Fig. 3.- Morphological and syntactic information of the word *ikusiaz*

Moreover, we must also consider that if intraword noun ellipsis occurs, genitive recursion has been applied, but the reverse is not always true.

3. The design of the tagset

The choice of a tagset is a critical aspect when designing a tagger, because the usefulness of the product and the ambiguity rate depend on it. We have found two main problems while trying to define the tagset for Basque:

- It did not exist an exhaustive tagset for automatic use because manual lemmatisation processes carried out on Basque texts in previous projects (Urkia and Sagarna, 1991) did not include a systematically built tagset. Moreover, Basque printed dictionaries also lack systematisation of categories.
- The output of the morphological analyser is too rich and it does not offer a directly applicable tagset.

The tags are used both as result of the process and to establish the tables that will allow the disambiguation based on statistics. The tagset system we have chosen for Basque is a three level system which the user can parametrise when using the programme. In the first level seventeen general categories are included (noun, adjective, verb, etc.). In the second one each category tag is further refined by subcategory tags (for example, the verb category has two subcategories: compound and simple verbs). The last level includes other interesting morphological information (case, number, etc.).

We also deal with complex tags since it is vital for derivation as well as for multiword terms, idiomatic expressions, abbreviations etc. The tagset is still open, but we have defined a total of 17 categories in the first level (two of them derived), with an average of 3 subcategories for each one on the second level.

4. Components of EUSLEM

In order to elaborate the lemmatiser/tagger for Basque we have used the following components:

- A pre-processor to detect and tag figures, punctuation marks, etc. Pre-processing those elements is very useful because they don't produce ambiguous tags and, therefore, they reduce the strings of ambiguous elements.
- The general-purpose morphological analyser for Basque, based on two-level morphology (Agirre *et al.*, 1992).
- Lexicon-free lemmatisation so that the system may be robust.
- Treatment of compound lexical units.
- Disambiguation based on linguistic knowledge, completed by disambiguation based on statistics.

The basis of all this components is the Lexical Database for Basque (EDBL) (Agirre *et al.*, 1995), which is both source and support for the lexicons needed not only in this application but in many others.

4.1. The lexical database

The Lexical DataBase for Basque (EDBL) was designed so as to be neutral in relation to linguistic formalisms, flexible, open and easy to use.

The fundamental entity in the Lexical DataBase is a class called EDBL Units. This class is specialised in three subclasses: <u>Dictionary Entries</u>, that contains those entries in the EDBL that you would expect to find in an ordinary dictionary, <u>Verb Forms</u>, that contains the finite verb forms, and the subclass <u>Non-Independent Morphemes</u> (suffixes, prefixes, etc.)

The morphological aspects of the entries and their variants are described by means of two features that all the lexical units of the database have: Morphology and Variants. The feature called Morphology has as value a Feature Structure that contains the two-level form (Koskenniemi 1983) of the word and two attributes featuring the morphotactic aspects: the <u>continuation class</u>, that describes the set of morphemes that can follow a given entry word, and the <u>sublexicon</u> to which the entry belongs. The variants of the lexical entry are described also based on the two-level model and are currently employed for a more intelligent correction strategy by the spelling corrector and for the lemmatisation and tagging of non-standard Basque texts.

4.2. The general morphological analyser/generator.

(Agirre *et al.*, 1992) describe the basic part of our morphological analyser, consisting in the application of two-level morphology (Koskenniemi, 1983) for Basque. Our system consists of:

- a set of 24 morphophonological rules that describe the changes occurring between the lexical level and the surface-level.
- a lexicon made up of about 65,000 items, grouped into 120 sublexicons and supported in the general lexical database (EDBL).

The morpheme strings that can be linked up are expressed by means of the above-mentioned continuation classes, that define the set of morphemes that can follow a given morpheme.

In order to evaluate the accuracy of the analyser, we tested some corpora and obtained a coverage between 92% and 96%. Examining the results we observed that most of the non-analysed words were linguistic variants (non-standard uses or competence errors) or forms the lemmas of which were not in the general lexicon (foreign words, proper names, derived words, etc.). Due to the fact that the process of normalisation of Basque is still in progress, the morphological processor must deal not only with standard but with dialectal forms of words. For this purpose, the treatment of variant errors was carried out using a two-level subsystem made up of: 1) about 1,000 items (mostly dictionary entries) linked to the corresponding correct ones and 2) twenty rules to cover the most common competence errors. With this information the system is able to analyse linguistic variants and to distinguish between standard and non-standard lemmas. This is one of the features of our lemmatiser, the capability of discriminating standard and non-standard lemmas.

Derivation and composition in Basque are quite productive and widely used in neologism formation, but they are not as systematic as declension and, therefore, their computational treatment becomes more complex. For the moment, a database for derivation has already been designed. This database will be integrated in the general database (EDBL) (Agirre *et al.*, 1995), so that, in the future, not only inflectional but also lexical morphology can be recognised.

4.3. The lexicon-free lemmatisation.

If the lexicon-based analyser produces no valid analysis, we need a lexicon-free lemmatiser that won't let any item unlemmatised or untagged, so that the system may be robust enough. Among the different systems we studied, we chose a two-level mechanism based on the idea used in speech synthesis (Black *et al.*, 1991). This mechanism has two main components so as to be capable of treating unknown words: 1) generic lemmas corresponding to each possible open category or subcategory, and 2) two additional rules in order to express the relationship between the generic lemmas at lexical level and any acceptable lemma of Basque.

As the output of the analysis, we obtain generic lemmas and concrete affixes. A heuristic is the responsible of finding concrete possible lemmas instead of the generic ones. These lemmas are given in the output as variants and not as standard forms. In order to eliminate the great number of ambiguities in the analyses a local disambiguation is carried out in terms of the length and the last characters of the lemmas (3). With this treatment the lemmatisation/tagging process is robust and accuracy is higher than 99%.

4.4. Treatment of multiword terms

It's not always easy to decide whether an item must be lemmatised/tagged as a compound unit or not. We rely on the experience UZEI gathered in the EEBS project (Urkia and Sagarna, 1991), in which a wider perspective was taken and an extensive range of MultiWord Terms (MWT) was lemmatised as a unit.

In order to describe MWTs in Basque, we have functionally established the following features:

- <u>contiguous/dispersed</u>: We say a MWT is dispersed when its components do not necessarily occur one after another. In that case, the processing gets more complicated since we have to seek the components in subsequent words. If a MWT has more than two components, some of them may be contiguous and some others may not.
- <u>ordered/order-free</u>: If a MWT is dispersed, its elements may not necessarily keep an order. A clear example of this are verb periphrasis such as *negar egin* -to cry-, *min eman* -to hurt-, *behar izan* -to need-,... since their constituents usually shift their positions, e.g. in negative clauses (we say <u>*lo*</u> <u>*egin*</u> *dut*, 'I have slept' but *ez dut* <u>*egin*</u> <u>*lorik*</u> 'I haven't slept').
- <u>Inflected/fixed</u>: The components of a MWT may either appear in an invariable form (*hurrenez hurren*) or be inflected. For instance, both constituents of the verb periphrasis *bizi izan* are inflected and therefore the possible combinations are countless (*bizi naiz, biziko banintz, bizi izanik*,...). In the case that the components of the MWT are inflected, we make two different groups: those accepting any inflection and those

accepting just a restricted set of inflected forms. Thus, restrictions are needed for components accepting just a few inflected forms.

- Sure/ambiguous: We say a MWT is *sure* when its components can only be analysed as a whole lexical unit and therefore no other interpretation is possible (*hala eta guztiz*).

EUSLEM treats the multiword terms *ad hoc*, keeping their information in the above-mentioned lexical database (EDBL). Once the words are lemmatised, it will check the database to see if the form or lemma can be part of a compound term. In that case, it will get the features of the compound term and check if the rest of the elements are also present, eliminating the rest of the analyses, if the term turns to be non-ambiguous.

We are using word co-occurrence measures to detect in lemmatised/tagged corpora multiword terms that are not currently in the database (Church and Hanks, 1989). The multiword terms that have a high enough frequency will be used to enrich the database.

4.5. The disambiguation process

In recent years a number of projects have focused on the automatic disambiguation of texts. Our lemmatiser/tagger, as others (Chanod and Tapanainen, 1994), will try to combine the two kinds of methods most successfully used.

Methods based on linguistic knowledge. The Constraint Grammar (CG) formalism (Karlsson *et al.*, 1995; Voutilainen, 1994; Tapanainen, 1994) works on a text where all the possible morphological interpretations have been assigned to each word-form by the morphological analyser. The role of the CG system is to apply a set of linguistic constraints that discard as many alternatives as possible, leaving in the end fully disambiguated sentences, with one interpretation for each word-form. We have lately been working with that formalism, that let us, amongst other things, profit by the existing morphological information (Aduriz *et al.,* 1995). We have already got about 200 rules for morphological disambiguation that solve different problems, such as certain ambiguities of categorial disambiguation, morphosyntax, etc.

Statistical techniques (Cutting *et al.*, 1992; Elworthy, 1993; Leech *et al.* 1994). The method we propose uses bigram or trigram probabilities based on knowledge obtained from manually disambiguated corpora.

A mixed method combining both techniques will be done using statistical disambiguation only when the method based on linguistic knowledge, which is more intelligent, is not capable of disambiguating.
Our aim is to use public software, but most of them are oriented to working with a dictionary of word-forms, so they are not able to deal with the output of the morphological analyser. Therefore, we are adapting one of them (Amstrong *et al.*, 1995) to our necessities.

5. Current situation

We are continuously updating the lexical database with lexical contribution from different sources (modern dictionaries, the Basque Language Academy's lexical proposals, feedback of the XUXEN spelling corrector's users, etc.) and with the new grammatical rules the Language Academy publishes periodically. Moreover, we are polishing the category and the continuation class systems in order to meet the requirements that keep arising while disambiguating.

The process of the compound lexical units is being tested with a sample of the most used multiword terms in Basque.

A corpus of more than 20.000 words has been manually disambiguated in order to extract the first tables for statistical disambiguation as well as to test the result of the disambiguation processes. This disambiguation is being done at the level of morphological information so that it will be possible to experiment with different tagsets. As for disambiguation based on linguistic knowledge, rules for morphological disambiguation are being devised and tested. They will be tested against the manually disambiguated text, before performing a massive disambiguation process.

6. Conclusions

The most relevant features of the lemmatiser/tagger for Basque are the next:

- it is a general-purpose tool
- it manages standard and non-standard lemmas
- it deals with and lemmatises multiword terms
- it is based on morphological analyses.

Although the lemmatiser/tagger for Basque is a general-purpose tool, we intend to apply it to lexicography, and along with an interface and a corpus manager it will be part of a tool for semiautomatic creation of lexicons.

Notes

1 UZEI is a cultural association created in 1977, the aim of which is to promote Basque lexicon's modernisation within the normalisation process of this language.
2 This corpus has got 5,800,000 text-words and is being used along with the EEBS corpus by the Language Academy for the completion of the unified dictionary (UZEI is helping the Academy in that task). But unlike the EEBS corpus, this one has not been exhaustively lemmatised. Lemmatisation has just been done depending on punctual necessities.
3 The treatment of the last characters is very important because some endings can be managed as suffixes when they are part of the lemma, e.g. the ending *-iko* in the adjectives.

References

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K., Urkia, M. 1992 "XUXEN: A spelling Checker/Corrector for Basque Based on Two-Level Morphology", in: *Proceedings of the Third Conference on ANLP (ACL)*, pp. 119—125.

Agirre, E., Arregi, X., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Insausti, J.M., Sarasola, K. 1995 "Different issues in the design of a general-purpose Lexical Database for Basque", in: *First Workshop on application of Natural Language to Data Bases, NLDB'95,* pp. 299—313.

Aduriz, I., Aldezabal, I., Alegria, I., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Urkia, M. 1994 "EUSLEM: Un lematizador/etiquetador de textos en euskera", in: *Actas del X Congreso SEPLN,* Córdoba.

Aduriz, I., Alegria, I., Arriola, J.M., Artola, X., Díaz de Ilarraza, A., Ezeiza, N., Gojenola, K., Maritxalar, M. 1995 "Different issues in the design of a lemmatizer/tagger for Basque", in: *Proceedings of the EACL SIGDAT workshop,* Dublin.

Alegria, I. 1995 *Euskal morfologiaren tratamendu automatikorako tresnak.* Ph. D. thesis. University of the Basque Country. 217 pp.

Amstrong, S, Russel, G., Petitpierre, D., Robert, G. 1995 "An open Architecture for Multilingual Text Procesing", in: *Proceedings of the EACL SIGDAT workshop,* Dublin. pp. 30—34.

Black, A., van de Plassche, J., Williams, B. 1991 "Analysis of Unknown words through Morphological Decomposition", in: *Proceedings. of the 5th Conference of the EACL*, Volume 1, pp. 101—106.

Chanod, J.P., Tapanainen, P. 1994 "Statistical and constraint-based taggers of French", in: *Xerox MLTT-016.*

Church, K. W. and Hanks, P. 1989 "Word association Norms, Mutual Information, and Lexicography" in: *Proceedings of ACL*, Vancouver, pp. 76—83.

Cutting, D., Kupiec, J., Pedersen, J., Sibun, P. 1992 "A practical part-of-speech tagger", in: *Proceedings. of the Third conference ANLP (ACL)*, pp. 133—140.

Elworthy, D. 1993 "Part-of-speech Tagging: A working paper", in: *Acquilex WP* Number 10.

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

Koskenniemi, K. 1983 *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production.* Ph D. thesis, University of Helsinki, Number 11.

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

Tapanainen, P., Voutilainen, A. 1994 "Tagging Accurately - Don't guess if you know", in: *Proceedings of ANLP'94*, pp. 47—52.

Urkia, M., Sagarna, A. 1991 "Terminología y lexicografía asistida por ordenador. La experiencia de UZEI", in: *Actas del VII Congreso de SEPLN*, Volume 9, pp. 193—202.

Voutilainen, A. 1994 *Three studies of grammar-based surface parsing of unrestricted English text.* Ph. D. thesis. University of Helsinki, Number 24, 79 pp.

Keywords: Basque, Computational Lexicography, Lemmatiser, Tagger, Morphology.