DIFFERENT ISSUES IN THE DESIGN OF A LEMMATIZER/TAGGER FOR BASQUE

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar, M.
Department of Computer Languages and Systems - University of the Basque Country
649 p.k., 20080 Donostia - Spain
jipgogak@si.ehu.es

ABSTRACT

This paper presents relevant issues that have been considered in the design of a general purpose lemmatizer/tagger for Basque (EUSLEM). The lemmatizer/tagger is conceived as a basic tool necessary for other linguistic applications. It uses the lexical data base and the morphological analyzer previously developed and implemented. Due to the characteristics of the language, the tagset here proposed is structured in four levels so that each level is a refinement of the previous one in the sense that it adds more detailed information. We will focus on the problems found in designing this tagset and on the strategies for morphological disambiguation that will be used.

1. Introduction

This paper describes the development of a general purpose lemmatizer/tagger for Basque which will lay the foundations for further applications in the field of automatic processing of Basque texts. In order to elaborate this project the following basic tools will be used:

- The Lexical Database for Basque (LDBB). This database is both source and support for the lexicons needed in different applications, and was designed with the objectives of being neutral in relation to linguistic formalisms, flexible, open and easy to use. At present it contains 60,000 entries, each with its associated linguistic features (category, subcategory, case, number, etc.).
- A general morphological analyzer/generator (Agirre et al., 92). For morphological analysis, we use morphological descriptions based on the two-level formalism proposed by Koskenniemi (Koskenniemi, 83). The morphological analyzer attaches to each input word-form all possible interpretations and its associated information. The LDBB also stores information about the two-level lexicon system.

As Fig. 1 shows, this morphological analyzer has been extended in two ways so as to make it more robust (Aldezabal et al., 94; Alegria et al., 95):

- The treatment of linguistic variants. Due to the fact that the process of normalization of Basque is still in progress, the morphological processor must deal not only with standard but also with dialectal and non-standard forms of words. For this purpose, new sublexicons and two-level rules dealing with this kind of forms have been added. With this additional information the system is able to analyze and correct competence errors.
- A two-level mechanism for lemmatization without lexicon, based on the idea used in speech synthesis (Black et al., 91). This mechanism has two main components in order 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. The obtention of at least one analysis is guaranteed, but in order to eliminate the great number of ambiguities in the analysis, a local disambiguation process is carried out.

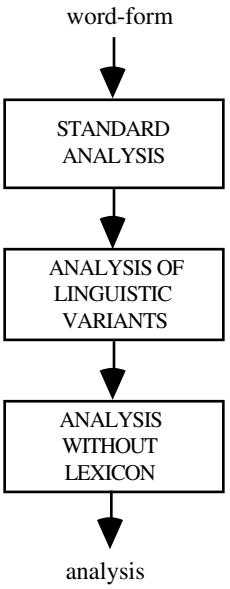

Fig. 1.- Different steps of morphological analysis

In section 2 we introduce some special characteristics of Basque morphology and the problem of ambiguity. Section 3 describes the problems found in designing a tagset for Basque texts. After that, in section 4 we will show the strategies that will be used for disambiguation. Finally, in section 5 some conclusions will be presented.

2. Brief Description of Basque

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 declension case are taken in this order and independently of each other.

One of the principal characteristics of Basque is its declension system with numerous cases. The marks corresponding to determination, number and case appear only after the last element in the noun phrase. This last element may be the noun, but also typically an adjective or a determiner. In Fig. 2 there is an example of the analysis of *semeArEN etxeAN* (in the house of the son).

seme	*A*	r	*EN*	*etxe*	*AN*
noun (son)	determinate	epenthetical element	genitive	noun (house)	inessive case

Fig. 2.- Different steps of morphological analysis

As prepositional functions are realized by case suffixes inside word-forms, Basque presents a relatively high power to generate inflected word-forms. For instance, from one noun entry a minimum of 135 inflected forms can be generated. Moreover, while 77 of them are simple

combinations of number, determination, and case marks, not capable of further inflection, the other 58 are word-forms ended with one of the two possible genitives or with a sequence composed of a case mark and a genitive mark. If the latter is the case, then by adding again the same set of morpheme combinations (135) to each one of those 58 forms a new, complete set of forms could be recursively generated. This kind of construction reveals a noun ellipsis inside a complex noun phrase and could be theoretically extended *ad infinitum*; however, in practice it is not usual to find more than two levels of this kind of recursion in a word-form.

Derivation and composition are quite productive and irregular, being widely used in neologism formation. Derivation is performed by means of more than 100 morphemes (mostly suffixes) and the categories used to create the derivatives are nouns *(kamiolari "lorry-driver")*, adjectives *(itsuskeria "ugliness")* and verbs *(ekarpena "contribution")* . Noun, adjective and verb are the most common categories in composition. Here are some examples of the most productive compound constructions in Basque:

noun + noun	*eguzki-lore "sunflower"*
adjective + adjective	*zurigorri "white and read"*
verb + noun	*idazmakina "typewriter"*

Both ways of word-formation are not as systematic as declension and, therefore, their computational treatment becomes more complex.

2.1. Morphological ambiguity in Basque

In order to deal with morphological ambiguity the following main types can be underlined:

- Categorial (or part of speech) ambiguity, like Noun/Verb, Verb/Adjective/Adverb, ... Just to have an idea about the magnitude of this kind of ambiguity, after analyzing a corpus of 2,113 word-forms we found 655 ambiguous, so the rate of categorial ambiguity is around 0.31 (1.38 analysis for each word-form).

- Morphosyntactic ambiguity. There are several possible morphosyntactic interpretations attached to each input word-form. For instance:

  gizonak Nominative Plural / Ergative Singular
  (the men / the man)

  The borderlines between morphology and syntax are not clear, because the information attached to each analysis contains features belonging to both morphology and syntax:

  ikusiaz Instrumental Case at morphological level / Modal Clause at syntactic level
  (seeing)

  Moreover, we must also point out that whenever intraword noun ellipsis occurs, genitive recursion has been applied, but the reverse is not always true. This kind of ambiguity (elliptic ambiguity) will be presented later in section three.

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 will depend on it. Before defining the tagset we have had to take some aspects into account:

a) There was not any exhaustive tagset for automatic use because manual lemmatization processes carried out on Basque texts in previous projects (Urkia and Sagarna, 91) did not include a systematically built tagset.

b) The output of the morphological analyzer, due to the characteristics of the language, is too rich and does not offer a directly applicable tagset.

After considering the main problems concerned with the this outpu of the morphological analyzer, the resulting multilevel tagset will be presented.

3.1. The treatment of the ellipsis

As we have mentioned above, noun ellipsis inside a word-form is a usual phenomenon. When noun ellipsis occurs, there is an explicit lemma plus an implicit one (it is also possible to have multiple ellipsis in a single word-form, although it is not frequent to find more than two). *Semearenak* (the ones of the son) is an example of a simple ellipsis:

seme	*a*	*r*	*en*	ε[1]	*ak*
noun (son)	determiner	epenthetical element	genitive	(elliptic noun)	nominative case

If we put it in an actual context:

> *Aitaren eskuak handiak dira eta* ***semearenak*** *txikiak*
> (The hands of the father are big and **the ones** (hands) **of the son** are small.)

In this example the word *semearenak* has an internal ellipsis referring to *hands*. Thus, there is an explicit lemma (son) and one implicit lemma (hand). There are two choices to tag this word: tagging at lemma-level and tagging at word-level.

- Tagging at lemma-level, so that each explicit or implicit lemma will have its own tag. E.g.:

  semearen-/**NOUN** ε-*ak*/**NOUN**

  ederreko-/**ADJECTIVE** ε-*etatik*/**NOUN**

  It must be added that we can not assert that whenever inflection marks appear after a genitive mark noun ellipsis occurs. In the previous cases *semearenak* is theoretically ambiguous, although the elliptic interpretation is the most likely. The output of the tagger in this alternative (lemma-level tagging) must reflect two possibilities in both cases:

  (1) *semearenak* /**NOUN** (no ellipsis)
  (2) *semearen-*/**NOUN** ε-*ak* /**NOUN**

  (1) *ederrekoetatik*/**ADJECTIVE** (no ellipsis)
  (2) *ederreko-*/**ADJECTIVE** ε-*etatik*/**NOUN**

  This kind of output poses some problems for the reuse of existing disambiguating tools and algorithms, which take word-forms as input.

---

[1] The symbol ε stands for the elided noun (implicit lemma).

- Tagging at word-level. The tag corresponds to the explicit lemma or it may be a composed tag collecting the information of all lemmas (explicit and implicit) present in the word. In this case we would use a composed tag:

  semearenak/**NOUN_WITH_NOUN_ELLIPSIS**

  ederrekoetatik/**ADJECTIVE_WITH_NOUN_ELLIPSIS**
  (from the ones of the pretty one)

  This kind of tagging implies an indexation of some morphological features, because of the ellipsis. For instance, when we say *semearenak (the ones of the son)* singular number appears, related to the explicit lemma *(the son)* and plural number appears related to the implicit lemma *(the ones)*.

After taking both choices into account, we took the second option: tagging will be at word-level but the information about the elided components will be kept and expressed by means of composed tags and indexation (see next section).

3.2. Derivation and composition

As we previously mentioned, composition and derivation are very productive. There are two main problems related to them:

- Ambiguity in compounds. As the terms appearing in a compound may be ambiguous, we decided to index the interpretations corresponding to compounds, much in the way done in (Leech et al., 94). This will also have the effect of increasing the ambiguity. For example:

plaza	*gizon*[2]	
/NOUN	/NOUN	
	/VERB	
/NOUN(1)	/NOUN(2)	(indexed tags for compounds)

- Derivation also poses some problems, because many times the derivative morphemes change the category of the stem. In some cases a composed tag will be assigned, whereas in others only the tag corresponding to the resulting category will be kept.

  E.g.: *etor* /VERB + *tze*/NOMINALIZATION_SUFFIX

  => *etortze*/VERBAL_NOUN[3] (composed tag)

Both problems have been addressed in the resulting tagset.

3.3. The Multilevel Tagset

In designing the general tagset we tried to meet the following requirements:

- It must take into account all the above mentioned problems of ellipsis, derivation and composition.

---

[2] *plaza:* square
gizon: man / to become a man
plaza gizon: public man

[3] *etortze:* (the act of) coming

- In addition, the lemmatizer/tagger has to be general, because it will be used as a step prior to other applications like parsing, information retrieval, indexing, etc.
- It must be coherent with the information provided by the morphological analyzer.

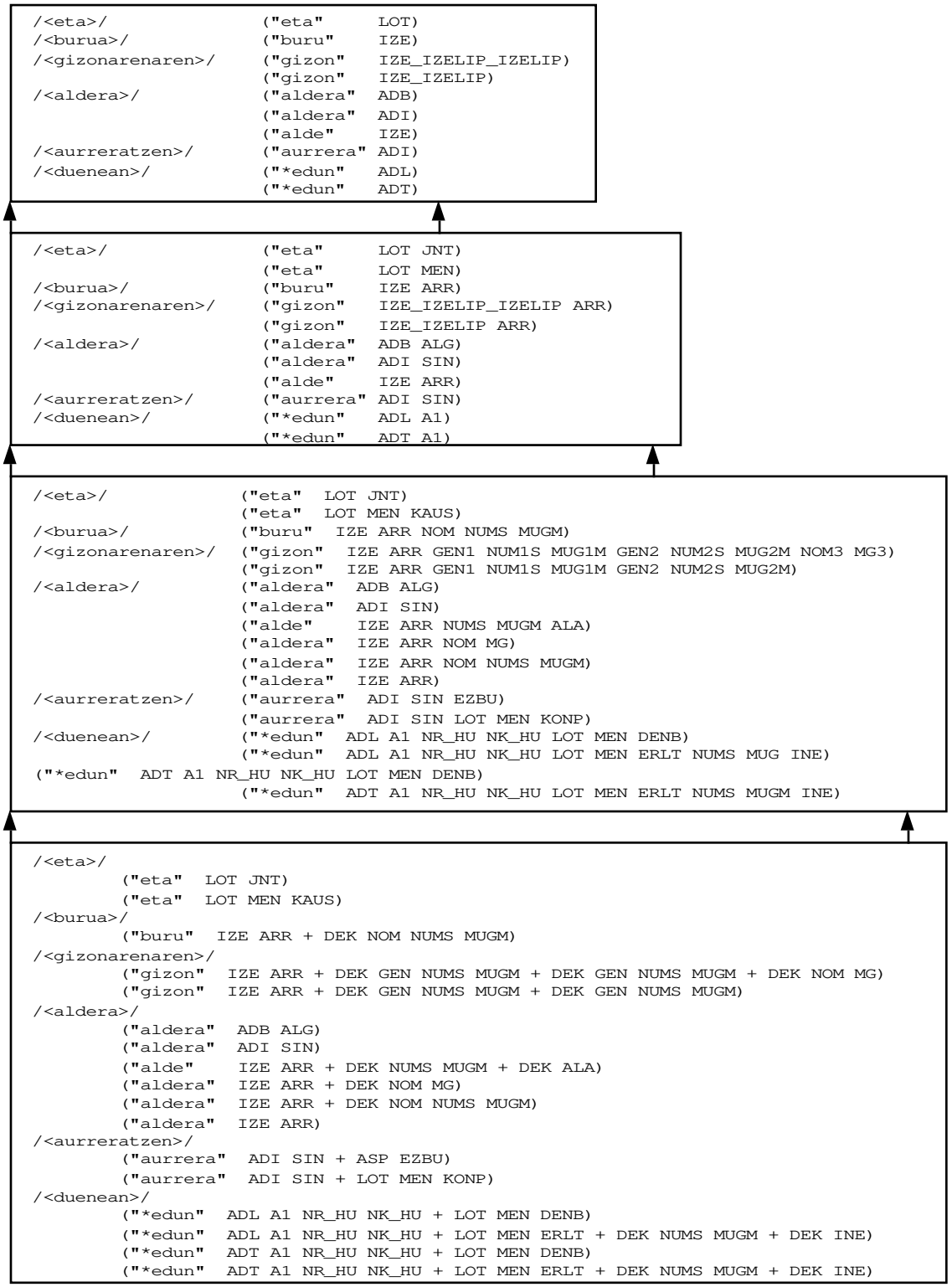

Fig. 3.- Different levels of tagging[4]

[4] In order to understand the figure number 3, some first level tags will be translated:
LOT = CONJUNCTION
IZE = NOUN
IZE_IZELIP_IZELIP = NOUN_WITH_DOUBLE_NOUN_ELLIPSIS
IZE_IZELIP = NOUN_WITH_NOUN_ELLIPSIS
ADB = ADVERB
ADI = MAIN VERB
ADT = FINITE MAIN VERB
ADL = FINITE AUXILIARY VERB

Taking all these considerations into account, the tagset has been structured in four levels (see fig. 3), ranging from the simplest part-of-speech tagging scheme up to the full morphological information:

- In the first level, seventeen general categories are included (noun, verb, etc.). This level distinguishes different tags for ellipsis and for the indexing of compounds. It is the basic tagset for ordinary lemmatization.
- In the second level each category tag is further refined by subcategory tags. For example, the verb category has two subcategories for simple and compound verbs.
- The third level may include other interesting morphological information, as declension case, number, etc. In this parametrizable level, the user is allowed to specify which information he/she wishes. This one and the previous level are suitable for the construction of tables of probabilities as well as for the design of rules to be used in disambiguation. As the structure of different morphemes is not explicit at this level, indexation of some features (e.g. number) is necessary because of the ellipsis.
- The output of the morphological analysis constitutes the last level of tagging. The only difference with the previous level is that, here, all the information given by the morphological analyzer is taken into account. This is, obviously, the level that will be the input for syntactic and other kinds of language processing.

4. The disambiguation process

In recent years a number of projects have centered on the automatic disambiguation of texts. Our lemmatizer/tagger, as others (Leech et al., 94), will try to combine the two kinds of methods most successfully used:

Methods based on linguistic knowledge. The Constraint Grammar formalism (Karlsson et al., 95; Voutilainen, 94; Tapanainen, 94) was designed with the aim of being a language-independent and robust tool to disambiguate and analyze unrestricted texts. It has been used in a project to disambiguate English corpora of about 200 million word-forms, and it claims to have a better performance than statistical methods (about 99% accurate in full disambiguation).

Although the formalism is intended to obtain a surface-syntactic annotation of texts, its main emphasis was put on solving the ambiguity problem. It works on a text where all the possible morphological interpretations have been assigned to each word-form by the morphological analyzer. 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 syntactic label for each word-form.

In our case, rules for morphological disambiguation are being devised and tested, focusing on the problem of categorial ambiguity. In designing rules we consider two types of linguistic knowledge sources: theoretic grammars and corpora. They will be tested against a previously disambiguated text, before performing a massive disambiguation process. For the moment, we have got more than 80 rules for morphological disambiguation that eliminate a third of the categorial and morphosyntactic ambiguities accurately.

Statistical techniques (DeRose, 88; Garside et al., 87; Cutting et al., 92; Elworthy, 93). The method proposed in different projects uses bigram or trigram probabilities based on knowledge obtained from corpora. Most of the systems use tables of probabilities for each tag of each word-form, but dealing with highly inflected languages implies that the set of word-forms to be considered is wider than in other languages. A possible approach is to build these tables based on the information associated to lemmas instead of word-forms, although the potential loss of information that could arise in this case has to be considered; in this case, the problem is to

recognize unambiguously the lemma. Multext (Armstrong et al., 95) proposes a flexible and modular approach that does not need the set of word-forms, but it is based on the probability of pairs of tags. We are trying to use it in our system.

We are dealing with this problem at present. A corpus has been manually disambiguated in order to extract the first tables for automatic disambiguation. The disambiguation is being done at the level of morphological information (third level) so that it will be possible to experiment with different tagsets.

In our system, once the morphological analysis has been performed, disambiguation based on linguistic knowledge will help in removing most of the incorrect interpretations. The remaining ambiguity will be dealt with by means of statistical techniques.

5. Conclusions

A lemmatizer/tagger for Basque, currently under development, has been described. It is based on a general-purpose morphological analyzer and it uses a mixed disambiguation technique that combines linguistic and statistical knowledge.

In our opinion, syntactic aspects must be considered when tagging words in agglutinative languages, where morphology and syntax are so tied. This characteristic has been taken into account when analyzing different problems —such as intraword ellipsis, derivation, and composition— arising in the design of the system. Particularly, the problem of intraword ellipsis has been described and two different tagging solutions discussed.

Along with the design of the lemmatizer, an incrementally structured four-level tagset that can be adapted to different needs has been defined.

One of the things this work has shown us is that tagsets developed for typologically very different languages are not directly applicable. This is a conclusion that should be taken into account when designing tagset standards.

6. 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. Proceedings of the Third Conference ANLP (ACL) (pp. 119-125). Trento, Italy.

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

Alegria I., Artola X., Sarasola K. (1995) Improving A Robust Morphological Analyser Using Lexical Transducers. RANLP95 199-205. Bulgary

Armstrong S., Russell G., Petitpierre P., Robert G. An Open Architecture for Multilingual Text Processing. Proc. of the ACL SIGDAT95, 30-34. Dublin.

Black A., van de Plassche,J., Williams B. (1991) Analysis of Unkown Words through Morphological Decomposition. *Proc. of 5th Conference of the EACL*, vol. 1, 101-106.

Chanod J.P., Tapanainen P. (1994) Statistical and constraint-based taggers of French. Xerox MLTT-016.

Cutting D., Kupiec J., Pedersen J., Sibun P. (1992) A practical part-of-speech tagger. Proc. of the Third Conference ANLP (ACL), 133-140.

DeRose S.. (1988) Grammatical Category Disambiguation by Statistical Optimization. Computational Linguistics, 14:1.

Elworthy D. (1993) Part-of-speech Tagging: A working paper, Acquilex WP No. 10.

Garside R., Leech G., Sampson G. (1987) The Computational Analysis of English: A corpus-based approach. Longman. London.

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

Koskenniemi K. (1993) Two-level Morphology: A general Computational Model for Word-Form Recognition and Production. Ph.D. thesis, University of Helsinki.

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

Urkia M., Sagarna A. (1991) Terminología y lexicografía asistida por ordenador. La experiencia de UZEI. Actas del VII Congreso de la SEPLN.

Voutilainen A. (1994) Three studies of grammar-based surface parsing of unrestricted English text. Ph.D. thesis. University of Helsinki. Publications no. 24.

Tapanainen P., Voutilainen A. (1994) Tagging Accurately - Don´t guess if you know. Proc. of ANLP´94, 47-52.