A CORPUS BASED MORPHOLOGICAL DISAMBIGUATION TOOL FOR BASQUE

Alegria I., Arriola J. M.[1], Artola X., Díaz de Ilarraza A., Gojenola K.,
Maritxalar M. (*)
Aduriz I. (**)
jibaregj@si.ehu.es
(*) Informatika Fakultatea, 649 P. K.,
20080 Donostia (Euskal Herria)
(**) UZEI, Aldapeta 20, 20009 Donostia (Euskal Herria)

Abstract

This paper presents the methodology followed in the construction of a surface-based morphosyntactic parsing grammar as well as the results obtained. It is based on the Constraint Grammar formalism which we find suitable for our project of analysing unrestricted texts. Besides, we will present a description of the main types of morphosyntactic ambiguity that we have identified for Basque and the disambiguation rules designed for their treatment. This work is the first step in the computational treatment of syntax.

Keywords: basque language, disambiguation, Constraint Grammar.

1. Introduction

This paper describes the design of morphological disambiguation rules as a first step to develop a robust grammar of Basque, conceived as a general basis for different applications, such as a lemmatiser/tagger (Aduriz et al., 96) and a syntactic corrector (Gojenola and Sarasola, 94).

We have chosen the Constraint Grammar (CG) formalism (Karlsson et al., 95; Voutilainen, 94; Tapanainen and Voutilainen, 94), which was designed with the aim of being a language-independent and robust tool to disambiguate and analyse unrestricted texts. The CG grammar statements are close to real text sentences and directly address some notorious parsing problems, especially ambiguity. Far from the rigidity imposed by other formalisms, and despite some problems, we think that it is satisfactory for languages like Basque, with some degree of free order of sentence constituents and rich morphology.

The fact that it is based on morphological analysis makes this formalism adequate for our objective. It works on a text where all the possible morphological interpretations have been assigned

[1] This work is supported by a grant of the Basque Government.

to each word-form by the morphological analyser (Alegria et al., 95). The basic parsing strategy is to profit from the existing morphological information. Every relevant structure is assigned directly via lexicon, morphology and mappings from morphology to syntax. The role of the CG system is to apply a set of linguistic constraints that discard as many alternatives as possible, leaving at the end almost fully disambiguated sentences, with one morphological/syntactic interpretation for each word-form.

There are four major steps in the CG morphosyntactic treatment of texts: morphological analysis, morphological disambiguation, determination of clause boundaries and the assignment of syntactic functions. The first step has been completed by means of our robust morphological analyser and nowadays we are specially involved in the design of rules for morphological disambiguation.

2. The morphological analyser

Basque is an agglutinative language, i. e., for the formation of words the dictionary entry takes each of the elements needed for the different functions (syntactic case included). More specifically, the affixes corresponding to the determiner, number and declension case are taken in this order and independently of each other.

One of the principal characteristics of the language is its declension system with numerous cases. The markers corresponding to definiteness, 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. 1 there is an example of the analysis of *semeArEN etxeAN* (‘in the house of the son’).

seme	A	r	EN	*etxe*	A	N
noun (‘son‘)	determiner (‘the‘)	epenthetical element	genitive case (‘of‘)	noun (‘house’)	determiner (‘the‘)	inessive case (‘in‘)

Fig. 1.- Analysis of *semeArEN etxeAN* (‘in the house of the son’)

For the morphological description, the two-level morphology (Koskenniemi, 83) was applied to Basque (Agirre et al. 92; Alegria 95) with a great coverage lexicon containing over 65,000 entries. Relating to the linguistic description used, we must say that it provides a fine-grained output. There are 20 major categories (or part of speech), each of them distinguished in more detail by the following features, among others:

* Subcategory. E.g., common and proper nouns, premodifying and postmodifying adjectives, etc.;
* Number: singular or plural;
* Indefinite/Definite distinction for nominals;

* Case: in contrast to many European languages Basque has a rich case system with 20 different cases that, due to the agglutinative nature of the language, can appear attached to almost every category;
* Mood (for verbs): indicative, imperative, etc.

The morphological analyser attaches to each input word-form all possible interpretations and its associated information. Let us illustrate it with an example:

etxeko bide hori ´the path of the house´

```
/<etxeko>/      (word-form) ´of the house´

      1. interpr. ("etxeko"          FMAINVERB)
      2. interpr. ("etxe"    NOUN C + DEC NUMS DETER + DEC GEN @NC> @<NC @ADVL + DEC ABS IND @OBJ @SUBJ)
      3. interpr. ("etxe"    NOUN C + DEC NUMS DETER + DEC GEN @NC> @<NC @ADVL)
      4. interpr. ("etxeko"          NOUN C + DEC ABS IND @OBJ @SUBJ)
      5. interpr. ("etxeko"          NOUN C )

/<bide>/        (word-form) ´path´

      1. interpr. ("bide"    NOUN C  + DEC ABS IND @OBJ @SUBJ)
      2. interpr. ("bide"    NOUN C )
      3. interpr. ("bide"    PRT)

/<hori>/´       (word-form) ´that´

      1. interpr. ("hori"    FMAINVERB)
      2. interpr. ("hori"    ADJ + DEC ABS IND @OBJ @SUBJ)
      3. interpr. ("hori"    ADJ
      4. interpr. ("hori"    DET DEM ABS NUMS DETER POST)
```

The design of the morphological analyser was performed with the main objective of being robust, that is, capable of treating both standard and non-standard forms in real texts. For this reason, as Fig. 2 shows, this morphological analyser has been extended in two ways:

- The treatment of linguistic variants (dialectal variants and typical errors) (Aldezabal et al., 94).

- A two-level mechanism for lemmatisation without lexicon to deal with unknown words, based on an idea used in speech synthesis (Black et al., 91).

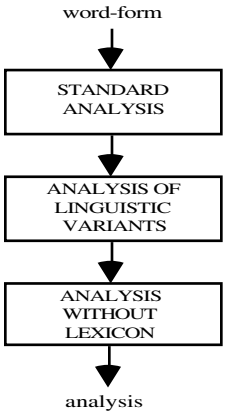

Fig. 2.- Different steps of morphological analysis

3. Morphological ambiguity

As the ambiguity rate depends on the granularity of the linguistic description, we can expect the input to the morphological disambiguator to be highly ambiguous. Furthermore, due to the late standardisation of Basque (it began in the late 60s and it is still going on), we find a number of non-standard phenomena in corpora, like variants and unknown words. As we will explain in detail later, they account for more than a tenth of the total number of interpretations. The number of morphological interpretations for these non-standard words is higher than for standard words; therefore, as a side-effect, when morphological disambiguation is applied in CG the results are not as good as we would like in the case of unknown surrounding words.

Table 1 contains data taken from a text consisting of 8,000 word-forms. It shows how the global ambiguity rate is of 2.65 analyses per word, with an average of 7.05 interpretations in the case of unknown words. The table also reveals that a relatively high percentage (7%) of the word-forms found in the text cannot be analysed by the standard morphological processor. Over 64% of the word-forms are ambiguous. This poses a hard disambiguation problem.

	N. of analyses per word	% of total word-forms	% of ambiguous word-forms
Standard forms	2.43	93%	62.7%
Linguistic variants	2.61	2%	84.44%
Unknown words	7.05	5%	99.57%
Total	2.65	100%	64.88%

Table 1: general ambiguity in the output of the morphological analyser.

Concerning the main types of ambiguity, we can distinguish, among others:

a) Categorial ambiguity, like Noun/Verb, Verb/Adjective/Adverb, etc. As table 2 shows, the ambiguity decreases about a half with respect to the previous data, when the annotation is reduced to the basic 20 categories in the lexicon, giving an average of 1.55 interpretations for each word-form.

	N. of analyses per word	% of total word-forms	% of ambiguous word-forms
Standard forms	1.44	93%	33.38%
Linguistic variants	1.36	2%	34.44%
Unknown words	3.83	5%	99.57%
Total	1.55	100%	36.54%

Table 2: ambiguity with respect to the main POS categories.

Relating to categorial ambiguity problems we have to point out that in some cases the ambiguity is due to a too "strict" linguistic distinction. For example, with Adjective/Adverb and Determiner/Pronoun pairs. In both cases the nature of the distinction is much more syntactic than categorial. The disambiguation practice shows us that it would be more adequate to have a unique category or to create a composed category that covers both of them.

• Morphosyntactic ambiguity. There are several possible morphosyntactic interpretations attached to each input word-form. In agglutinative languages morphology and syntax are tightly related to each other ; this is the reason for grouping them together in our description. This kind of ambiguity gives the difference from the 1.55 analyses per word in the case of categorial POS distinction to the 2.65 analyses in the general output of the analyser.

For instance:

gizonak	Absolutive Plural	'the men'
	Ergative Singular	'the man'
ikusiaz	Instrumental case (Morphological level)	'by (Noun) seen'
	Modal case (Syntactic level)	'seeing'

However, we must also consider that there are some cases in which the ambiguity only concerns to syntax. For example, in the case of subordinative elements, there are two aspects related to syntactic ambiguity. Firstly, we have to decide which kind of relationship is established by the subordinative elements and, on the other hand, which is the syntactic function of these elements. For instance:

	Syntactic relationship.	Syntactic function Amb[2]
-la	Completive	@SUBJ @OBJ
-la	Adverbial	@ADJUNT

As can be seen, the resolution of these ambiguities needs a deep morphosyntactic information to be resolved (the verb complementation pattern, among other aspects). The fact that we are dealing with morphosyntactic ambiguity makes the ambiguity resolution harder.

4. The design of disambiguation rules

In this section we will describe the steps followed in the design of the rules. We will show an example, the results obtained and the types of problems found during the process.

4.1. Methodology

Here we will focus on the methodology currently followed for the design of constraint rules. The process to formulate the rules has been carried out in different steps:

1) Establish the linking between the morphological analyser and the parser. Due to the fact that the morphological information given for each word-form is very detailed, we are developing a word-grammar to combine these morphological features in order to give an adequate description of each interpretation. This word-grammar will give a more optimized output, in the sense that it will facilitate the design of the rules.

2) Study of the phenomenon of morphological ambiguity. For that reason, we examined firstly the morphological ambiguity in the entries of the lexical database (Agirre et al. 95), taking into account the percentage of lexical entries for each ambiguity type (e.g. ADJECTIVE/ADVERB, VERB/ADJECTIVE/NOUN, etc), and secondly the frequency of those ambiguity types in the corpus recorded in the EEBS[3] project (Urkia and Sagarna, 91), and identified different types of morphological ambiguity, which have been presented in section 3.

3) Manual disambiguation of the corpus. Part of the corpus (about 22,000) words has been morphologically disambiguated by hand. The given morphological description will have its effect in the process of manual text disambiguation. It has been performed on the output of the morphological analyser. The corpus has been disambiguated by two different persons and the results were compared, applying the "double blind" method described in (Voutilainen and Järvinen, 95a). This manually disambiguated text serves two purposes:

- the obtention of a common definition of the tagging scheme (a grammatical representation),

- as a test for evaluating the results obtained with automatic taggers.

[2] In the CG formalism, the syntactic functions are preceded by @ character.
[3] The EEBS project is been carried out by the Language Academy in collaboration with UZEI (Center for the Lexical Standardisation of Basque). Its aim is to record and lemmatise written corpus (about 3 million words) for the elaboration of a unified dictionary.

In our case, the richness of the description gave, at the beginning, an error rate of about 5% between the two different annotators disambiguating the same text separately. After some discussions, less than 1% of the errors were left unresolved. In the case of the resolved ones, the two linguists had different linguistic perspectives, mainly as a consequence of the standardisation process of the language.

4) Design of rules adequate for disambiguating the cases established before. These rules were formulated, implemented, and tested using the corpus of 22,000 words. The detection of differences will produce the reformulation of the rules and the addition of new ones. This process will continue until the treatment of the types of morphological ambiguity considered is successful.

5) Test of the rules designed in the fourth step, using a new corpus of 1,000 words. In case the disambiguation rate is similar to the manually annotated corpus, new phenomena of morphological ambiguity will be chosen and we will start again with the fourth step. However, if the disambiguation rate is not satisfactory for the 1,000 words corpus, the design of constraint rules will go back to the fourth step in order to implement a new version of the last designed rules.

4.2. Rules

At the moment, there are 250 morphological disambiguation rules: 227 of them use unbounded context conditions and 23 are limited to one or two words around the ambiguous word. The most frequently used operators are the ones that discard the target reading.

Considering the phenomenon they wanted to resolve, the rules formulated can be classified in four different types:

1) Those which treat some verbal forms that can have two readings: finite synthetic and finite auxiliary (as the one shown in the example below).

2) Those which treat ambiguity due to declension cases and other morphological features.

3) Those formulated for the resolution of categorial ambiguity.

4) 65 rules that deal with specific word-forms.

Let us show an example:

[4] (@w=0 (NOTPART) (*-1 SD *R)
(NOT * R FAUXVERB1)
(*-1 SD *L)
(NOT *L FAUXVERB1)

This is an example of a sentence where the rule applies:

Jonek egin zuen lan guztia 'John did all the work'

[4] NOTPART- infinitive verbs; FAUXVERB1- finite auxiliary verbs: subjunctive,imperative and potential mood; SD- sentence delimiter.

This constraint is concerned with some verbal forms that can have two readings as a finite verb: infinitive and participle (in this example *egin*). This rule discards the infinitive reading if the following context conditions are satisfied: the absence of a finite auxiliary verb (subjunctive, imperative and potential mood) into the sentence.

4.3. Results

As we are in the process of developing the full grammar for morphological disambiguation, we are able to present the first results.

Table 3 gives an overview of the results of the disambiguation applied to the full output of the morphological analyser. These results are taken from a 8,000 word text, that was neither previously examined nor used for the development of the rules. This experiment gives us an idea of the potential robustness of the tool for the coverage of real texts.

	N. of analyses per word	% of ambiguous word-forms	% of correct interpretations
General input	2.65	64.88%	100%
Output	1.45	25.85%	96.51%

Table 3: results of morphological disambiguation.

The table presents the disambiguation performed on general texts. The number of interpretations is reduced to about a half, maintaining more than 96.5% of the correct interpretations. We consider the reduction of the ambiguity (from 2.65 to 1.45) satisfactory, even more if we take into account that the disambiguation work is still in process, and also that the ambiguity rate is very high, compared with other works (Voutilainen, 95). About a fourth of the word-forms are still ambiguous.

As could be expected, the results are worse for unknown words, with 3.8 analyses per word left in the case of disambiguating the full morphological output. This also adds to errors in the disambiguation performed on surrounding words. Even when the agglutinative nature of the language offers a big number of alternatives for unknown words, we have estimated that with heuristics based on capitalisation and word endings, about 60% of the ambiguities could be safely discarded. We can anticipate that this will have a positive effect on the other measures.

	N. of analyses per word	% of ambiguous word-forms	% of correct interpretations
General input	1.55	36.54%	100%
Output	1.09	7.57%	98.20%

Table 4: results of disambiguation with respect to the main categories.

When only the 20 main categories are considered, we get 1.09 interpretations for each word-form. We must also add that the ambiguity rate of the input was considerably lower, with 1.55

analyses per word-form. In the same way, the results have improved for the remaining correct interpretations, reaching to 98.2%.

(Elworthy, 95) performed an experiment to question the idea that smaller tagsets give better results in disambiguation. He tried the same statistical tagger with different size tagsets, concluding that in most of the cases the higher granularity of the tagset gives better results. As our results seem to contradict his view, we believe that this can be in part due to the high ambiguity rate of the input (about 64% of the word-forms are ambiguous), while in his experiment the highest ambiguity rate with any of the tagsets was no more than 50%. On the other hand, our results were taken after applying the disambiguation rules to the full output of the morphological analyser and then filtering the results to the main categories and so, in our opinion, the high granularity of the tagset helped to give good results.

However, there is also a number of harder constructions to deal with, where syntax only can not exactly determine the correct interpretation, that we expect to leave at least unresolved.

4.4. Pending problems

Some of the unresolved ambiguities are of syntactic nature, that is, the word-form has already been morphologically disambiguated but it has different syntactic readings. For instance, in the case of subordinative sentences, there are 178 syntactically ambiguous words with this kind of ambiguity in the 1000 words text that we have tested. For example:

```
"<den>"
    "izan" FAUXVERB A1 SG3 CS IQ @+FMAINV_MP @+FAUXV_MP
    "izan" FAUXVERB XV A1 SG3 CS REL @+FMAINV_CN> @+FAUXV_CN>
    "izan" FMAINVERB A1 SG3 CS IQ @+FMAINV_MP @+FAUXV_MP
    "izan" FMAINVERB A1 SG3 CS REL @+FMAINV_CN> @+FAUXV_CN>
```

This verb can be used as a main (FMAINVERB) or auxiliary (FAUXVERB) verb. Each of them has an interpretation as a part of a relative clause (REL), and another as part of an indirect question (IQ). All interpretations have the subordinating conjunction label (CS).

This kind of ambiguity needs an exhaustive formalisation of core elements of the grammar such as verb subcategorization. In order to cope with this problem we have created some sets that reflect the complementation pattern of some verbs. This work is being carried out but it is not still finished.

This kind of problems also requires the determination of clause boundaries which, after being determined accurately, help to delimit the internal structure of clauses.

Most of these problems show us the difficulty of separating morphology and syntax. We have established the main syntactic function labels which are included in the lexical database and attached to each interpretation by the morphological analyser. This has the effect of increasing considerably the ambiguity rate. We plan to work on rules for syntactic function disambiguation.

5. Conclusions

A surface-based morphosyntactic parsing grammar, currently under development, has been described. We have presented the design, implementation and test of rules for morphological disambiguation. Even when the grammar is still under development, the results are satisfactory in the case of disambiguating the full morphological description, where a 96.5% accuracy is given at the cost of maintaining part of the ambiguity in the analysis of new corpus not studied previously. The results improve considerably when only categorial disambiguation is performed, because of the better accuracy and the lower ambiguity rate of the output. The work also shows us the tight relationship between morphological disambiguation and syntax.

Acknowledgements

We are in debt with the research-team of the General Linguistics Department of the University of Helsinki for their permission to use the Constraint Grammar Parser.

This research was supported by the Basque Government, the University of the Basque Country and the Department of Economy of the Gipuzkoako Diputazioa

References

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N., Urizar R. "EUSLEM: A lemmatiser/tagger for Basque" EURALEX´96,.Gothenburg. 1996

Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M. "XUXEN: A spelling Checker/Corrector for Basque Based on Two-Level Morphology" Proc. of the 3rd Conference on ANLP (ACL), Trento, 1992

Agirre E., Arregi X., Arriola J. M., Artola X., Díaz de Ilarraza A., Insausti J. M., Sarasola K. "Different Issues in the Design of a General-Purpose Lexical Database for Basque" First Workshop on Application of Natural Language to Databases, 1995

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

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

Alegria I., Artola X., Sarasola K. "Improving a robust morphological analyzer using lexical transducers" RANLP, Bulgaria. 1995

Black A., van de Plassche J., Williams B. "Analysis of Unknown words through Morphological Decomposition" Proceedings. of the 5th Conf. of the EACL, 1991

Elworthy D. "Tagset Design and Inflected Languages" From Texts to Tags: Issues in Multilingual Text Analysis. ACL SIGDAT Workshop, Dublin, 1995

Gojenola K., Sarasola K. "Aplicaciones de la relajación gradual de restricciones para la detección y corrección de errores sintácticos" SEPLN, Córdoba, 1994

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

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

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

Urkia M., Sagarna A. "Terminología y Lexicografía Asistida por Ordenador. La experiencia de UZEI" Actas del VII Congreso SEPLN, 1991

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

Voutilainen A, Järvinen T. “Specifying a shallow grammatical representation for grammatical purposes” Proceedings of the 7th Conference of EACL, Dublin, 1995a

Voutilainen A. “A syntax-based part-of-speech analyser” Proceedings of the 7th Conference of the EACL, Dublin, 1995b