MORPHOSYNTACTIC DISAMBIGUATION FOR BASQUE BASED ON THE CONSTRAINT GRAMMAR FORMALISM

Aduriz I. (**), Arriola J. M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. (*).

(*) Informatika Fakultatea, 649 P. K., 20080 Donostia (Euskal Herria)
(**) UZEI, Aldapeta 20, 20009 Donostia (Euskal Herria)
jipgogak@si.ehu.es

Abstract

This paper presents the development 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 disambiguating unrestricted texts. Besides, we will present a description of the main types of morphosyntactic ambiguity that we have identified and the disambiguation rules designed for their treatment. This work is the first step in the computational treatment of Basque syntax.

Keywords:
Morphosyntactic disambiguation
Constraint Grammar
Basque language

Word Count: 3200

1 Introduction

This paper describes the design of morphosyntactic disambiguation rules as a first step to develop a robust grammar of Basque, conceived as a general basis for different applications; for instance, 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 crucial parsing problems, especially ambiguity.

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 morphosyntactic interpretations have been assigned to each word-form by the morphological analyser (Alegria et al., 95, 96b). The basic parsing strategy is to profit from the existing morphosyntactic 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 morphosyntactic/syntactic interpretation for each word-form.

There are two major steps in the CG morphosyntactic treatment of texts: morphological analysis and morphosyntactic disambiguation. The first step has been completed by means of a morphological analyser and nowadays we are finishing the design of rules for morphosyntactic disambiguation and on the resolution of syntactic ambiguities.

2. Morphological analysis

Basque is an agglutinative language, that is, 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, which shows how the morphological analysis of a word is equivalent to the analysis of a phrase.

etxe	a	n
noun (‘house’)	determiner (‘the‘)	inessive case (‘in‘)

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

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.

2.1 The morphological analyser

The analyser attaches to each input word-form all possible interpretations and its associated information. The result is the set of possible morphosyntactic analyses of a word, where each morpheme is associated with its corresponding features in the lexicon: category, subcategory, declension case, number and definiteness, as well as the syntactic functions and some semantic features.

It was designed with the main objective of being robust, that is, capable of treating both standard and non-standard forms in real texts. Furthermore, due to the late standardisation of Basque (it began in the late 60’s and it is still going on), we find a number of non-standard phenomena in corpora, like variants and unknown words. 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).

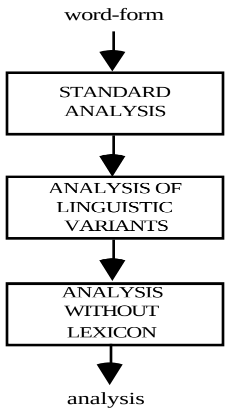


Fig. 2.- Different steps of morphological analysis

2.2 The design of the tagset

Relating to the linguistic description used, we must say that it provides a fine-grained output. The choice of a tagset is a critical aspect for disambiguation, because the usefulness of the product and the ambiguity rate depend on it. The main problem we found while defining the tagset was the absence of an exhaustive one for automatic use because manual lemmatisation processes carried out on Basque texts in previous projects (Urkia and Sagarna, 91) did not include a systematically built tagset. Moreover, Basque printed dictionaries also lacked systematisation of categories.

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

- It must take into account problems such as intraword ellipsis, derivation and composition.

- In addition, the tagset defined has to be general, because it will be the base for future developments.

- It must be coherent with the information provided by the morphological analyser.

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 morphosyntactic information. Complex tags are also dealt with since it is vital for derivation as well as for multiword terms, idiomatic expressions, abbreviations etc.

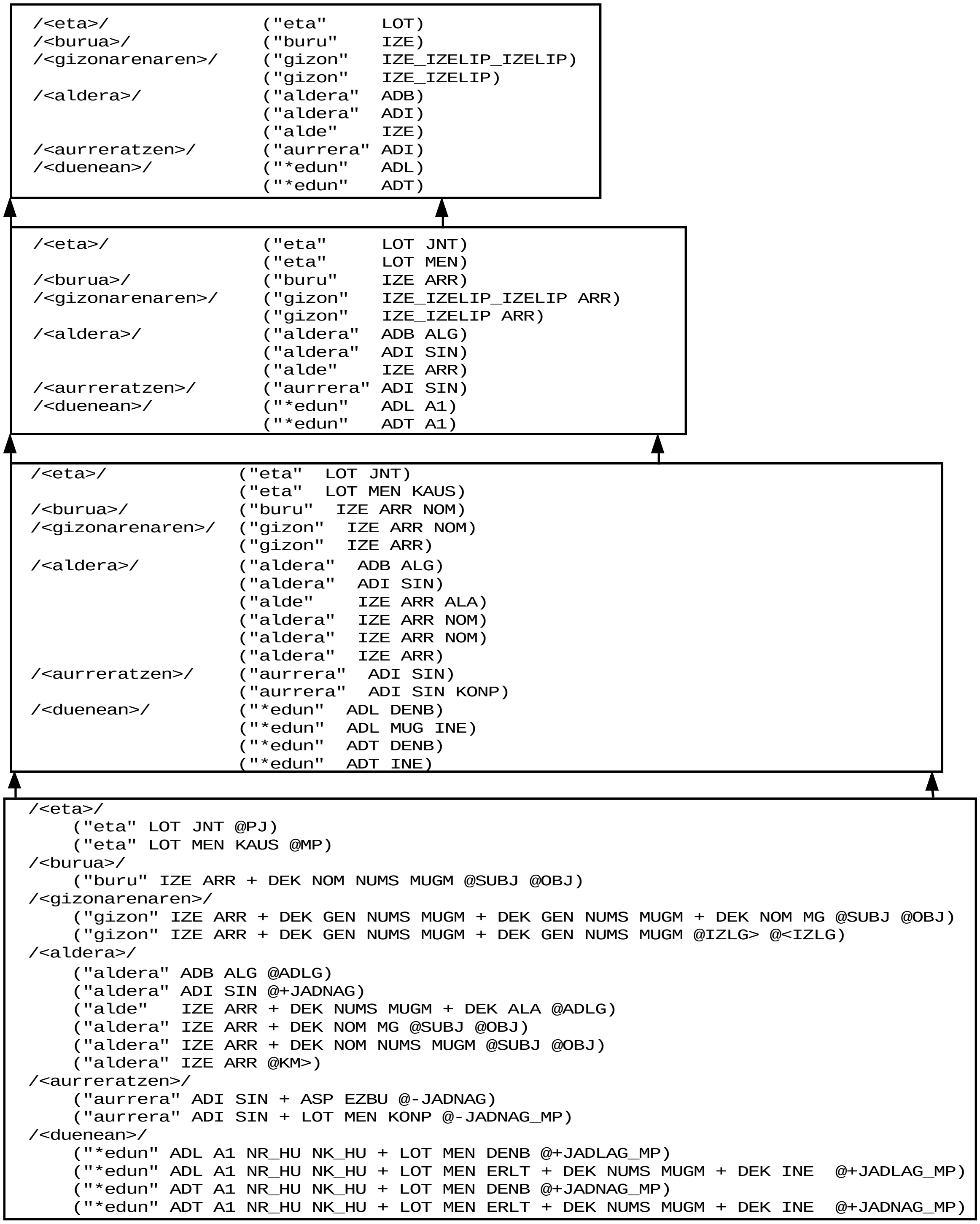
The levels defined consist of:

- In the first level, 19 general categories are included (noun, verb, etc.). It is the basic tagset for ordinary lemmatization.

- In the second level each category tag is further refined by subcategory tags. It contains 47 different tags. For example, the verb category has two subcategories for simple and compound verbs.

- The third level may include other interesting morphosyntactic information, as declension case, number, etc. (for example, taking the category, subcategory and case in a sample text 318 tags were found). In this parametrizable level, the user is allowed to specify which information is needed at wish.

- The full 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 analyser is taken into account, including tags for syntactic functions. The specification at this level is very detailed, with 2943 tags (not including syntactic functions). This level constitutes the input to the morphosyntactic disambiguation process and it will be the base for syntactic and other types of language processing.



```
/<eta>/              ("eta"     LOT)
/<burua>/            ("buru"    IZE)
/<gizonarenaren>/    ("gizon"   IZE_IZELIP_IZELIP)
                     ("gizon"   IZE_IZELIP)
/<aldera>/           ("aldera"  ADB)
                     ("aldera"  ADI)
                     ("alde"    IZE)
/<aurreratzen>/      ("aurrera" ADI)
/<duenean>/          ("*edun"   ADL)
                     ("*edun"   ADT)
```

```
/<eta>/              ("eta"     LOT JNT)
                     ("eta"     LOT MEN)
/<burua>/            ("buru"    IZE ARR)
/<gizonarenaren>/    ("gizon"   IZE_IZELIP_IZELIP ARR)
                     ("gizon"   IZE_IZELIP ARR)
/<aldera>/           ("aldera"  ADB ALG)
                     ("aldera"  ADI SIN)
                     ("alde"    IZE ARR)
/<aurreratzen>/      ("aurrera" ADI SIN)
/<duenean>/          ("*edun"   ADL A1)
                     ("*edun"   ADT A1)
```

```
/<eta>/            ("eta"  LOT JNT)
                   ("eta"  LOT MEN KAUS)
/<burua>/          ("buru"  IZE ARR NOM)
/<gizonarenaren>/  ("gizon"  IZE ARR NOM)
                   ("gizon"  IZE ARR)
/<aldera>/         ("aldera"  ADB ALG)
                   ("aldera"  ADI SIN)
                   ("alde"    IZE ARR ALA)
                   ("aldera"  IZE ARR NOM)
                   ("aldera"  IZE ARR NOM)
                   ("aldera"  IZE ARR)
/<aurreratzen>/    ("aurrera"  ADI SIN)
                   ("aurrera"  ADI SIN KONP)
/<duenean>/        ("*edun"  ADL DENB)
                   ("*edun"  ADL MUG INE)
                   ("*edun"  ADT DENB)
                   ("*edun"  ADT INE)
```

```
/<eta>/
    ("eta" LOT JNT @PJ)
    ("eta" LOT MEN KAUS @MP)
/<burua>/
    ("buru" IZE ARR + DEK NOM NUMS MUGM @SUBJ @OBJ)
/<gizonarenaren>/
    ("gizon" IZE ARR + DEK GEN NUMS MUGM + DEK GEN NUMS MUGM + DEK NOM MG @SUBJ @OBJ)
    ("gizon" IZE ARR + DEK GEN NUMS MUGM + DEK GEN NUMS MUGM @IZLG> @<IZLG)
/<aldera>/
    ("aldera" ADB ALG @ADLG)
    ("aldera" ADI SIN @+JADNAG)
    ("alde"   IZE ARR + DEK NUMS MUGM + DEK ALA @ADLG)
    ("aldera" IZE ARR + DEK NOM MG @SUBJ @OBJ)
    ("aldera" IZE ARR + DEK NOM NUMS MUGM @SUBJ @OBJ)
    ("aldera" IZE ARR @KM>)
/<aurreratzen>/
    ("aurrera" ADI SIN + ASP EZBU @-JADNAG)
    ("aurrera" ADI SIN + LOT MEN KONP @-JADNAG_MP)
/<duenean>/
    ("*edun" ADL A1 NR_HU NK_HU + LOT MEN DENB @+JADLAG_MP)
    ("*edun" ADL A1 NR_HU NK_HU + LOT MEN ERLT + DEK NUMS MUGM + DEK INE  @+JADLAG_MP)
    ("*edun" ADT A1 NR_HU NK_HU + LOT MEN DENB @+JADNAG_MP)
    ("*edun" ADT A1 NR_HU NK_HU + LOT MEN ERLT + DEK NUMS MUGM + DEK INE  @+JADNAG_MP)
```

Figure 3.- Different levels of tagging[1]

[1] The following list contains the equivalences in English of some tags:
LOT = CONJUNCTION IZE = NOUN ADB = ADVERB
ADI = MAIN VERB ADT = FINITE MAIN VERB ADL = FINITE AUXILIARY VERB

Comparing the levels in our system with those defined in the PAROLE project (ITEM 97), they are “similar” with 19 and 16 main categories at the first level respectively. They cover the same spectrum, the differences being mainly due to particularities in the language (like, for example, the lack of a category corresponding to *article* in Basque).

3. Morphosyntactic ambiguity

We define morphosyntactic ambiguity to denote all ambiguities in the word-form domain excluding syntactic functions[2]. As the ambiguity rate depends on the granularity of the linguistic description, we can expect the input to the morphosyntactic disambiguator (that is, the last level of tagging, see Figure 3) to be highly ambiguous.

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

a) Categorial ambiguity, like Noun/Verb, Verb/Adjective/Adverb, etc. As Table 1 shows, the ambiguity when the annotation is reduced to the basic 19 categories in the lexicon, there is an average of 1.55 interpretations for each word-form (ambiguous and non-ambiguous).

	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 1. Ambiguity with respect to the main POS categories.

b) Morphosyntactic ambiguity. There are several possible morphosyntactic interpretations attached to each input word-form, due to declension and other morphosyntactic features.

Table 2 contains data taken from a text consisting of 10,000 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. The number of morphosyntactic interpretations for these non-standard words is higher than for standard words. 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 2. General ambiguity in the output of the morphological analyser.

Morphosyntactic ambiguity is one of the most pervasive problems to deal with. As an example, the word *gizonak* has two readings, one as NOUN-ABSOLUTIVE-PLURAL (the men), acting as object or subject, and the other as NOUN-ERGATIVE-SINGULAR (the man), acting as subject. For instance:

gizon + ak	NOUN + ABS(olutive) P(lural) ‘the men’
	NOUN + ERG(ative) S(ingular) ‘the man’

Example 1. Case ambiguity in *gizonak*

Another example is the case of subordinative morphemes. Firstly, we have to decide which kind of subordination is established by the morpheme and, secondly, which is the syntactic function of the subordinative clause. For instance:

Subordinative morpheme	Subordination relationship
-la	CC (Clausal Complement)
-la	AC (Adverbial Clause)

Example 2. Syntactic ambiguity in *-la* morpheme

The resolution of these ambiguities 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.

These examples show how in agglutinative languages morphology and syntax are tightly

[2] In the CG formalism, the syntactic functions are preceded by the @ character.

related to each other, as the resolution of these types of ambiguity implies the determination of syntactic function. This is the reason why we use the term morphosyntax in our description. As a consequence, we have opted for solving these ambiguities as soon as possible, instead of postponing the hard work to the syntactic disambiguation level.

c) Syntactic ambiguity. We must also consider that there are some cases in which the ambiguity concerns only to syntax. In the previous example, the Clausal Complement reading has two possible syntactic functions, as subject (@SUBJ) or object (@OBJ). This kind of ambiguities has not been taken into account in the present paper.

4. The design of a disambiguation grammar

In this section we describe the steps followed in the design of the rules and the results obtained. The main basis in the design of the grammar are the Lexical Database for Basque (LDB, Agirre et al. 95) and the morphological analyser.

4.1. Methodology

In this section we 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) Study of the phenomenon of morphosyntactic ambiguity. We examined firstly the categorial ambiguity in the entries of the lexical database, taking into account the percentage of lexical entries for each type (e.g. Adjective/Adverb, Verb/Adjective/Noun, etc), and secondly we studied the output of the morphological analyser in order to identify different types of morphosyntactic ambiguity.

2) Manual disambiguation of a corpus. Part of the corpus (about 24,000 words) has been disambiguated by hand. The given morphosyntactic description had its effect in the process of manual text disambiguation, which has been performed on the output of the analyser. The corpus has been disambiguated by two different linguists and the results were compared, applying the “double blind” method described in (Voutilainen & Järvinen, 95a). This manually disambiguated text serves two purposes:

- the obtention of a common definition of the tagging scheme (a grammatical representation, that is, a source that can be consulted in case of disagreement).

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

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 lack of standardisation of the language.

3) Design of rules adequate for disambiguating the cases established before. These rules were formulated, implemented, and tested using a part of the disambiguated corpus (14,000 words) and the corpus recorded in the EEBS[3] project (Urkia & Sagarna, 91). The rest of the manually disambiguated one (10,000 words) was used for testing. The detection of differences produced the reformulation of the rules and the addition of new ones. This process continued until the treatment of the types of morphosyntactic ambiguity considered was successful.

4) Test of the rules designed in the fourth step. In case the disambiguation rate is not satisfactory, the design of constraint rules goes back to the third step in order to implement a new version of the last designed rules.

At the moment, there are 547 morphosyntactic disambiguation rules: 49 of them use unbounded context conditions, 486 are limited to one or two words around the ambiguous word, 298 treat specific word-forms, and 91 are for syntactic disambiguation. Most of the work on general disambiguation rules has been completed, and now we are focusing on the design of rules associated to particular word-forms and on the assignment of syntactic functions.

4.2. Results

Table 3 gives an overview of the results of the disambiguation applied to the full output of the morphological analyser. These results are

[3] The EEBS project was carried out by the Language Academy in collaboration with UZEI (Center for the Lexical Standardisation of Basque). Its aim is to record and lemmatise a three million-words corpus for the elaboration of a unified dictionary.

taken from a 10,000 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 total word-forms	% of ambiguous word-forms
General input	2.65	64.88%	100%
Output	1.63	25.85%	97.51%

Table 3. Results of morphosyntactic disambiguation.

The table presents the disambiguation performed on general texts. The number of interpretations is reduced to about a half, maintaining more than 97.51% 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 progress, and also that the original 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 ambiguity rate is higher for unknown words, with 3.8 analyses per word left in the case of disambiguating the full morphological output. This also adds 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 total word-forms	% of ambiguous word-forms
General input	1.55	36.54%	100%
Output	1.10	7.57%	99.12%

Table 4. Results of disambiguation with respect to the main categories.

When only the 19 main categories are considered, we get 1.10 interpretations for each word-form, on the same input texts. 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 been improved for the remaining correct interpretations, reaching to 99.12%.

(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.

5. Evaluation of the formalism and pending problems

Our experience with the application of the CG formalism shows that, far from the rigidity imposed by other formalisms, it is satisfactory for languages like Basque, with some degree of free order of sentence constituents and rich morphology. However, as (Voutilainen 94) points out, the difficulty of determining intra-sentential clause boundaries is a core problem that makes the resolution of morphosyntactic ambiguities extremely laborious. Anyway, the disambiguation practice shows us that this is not so important at categorial level.

Another question related to this task is the difficulty when referring to phrase-like units (like noun phrases), a problem inherent to the CG formalism due to the fact that the basic unit is the word-form. Therefore, these units are referenced in an indirect way, by means of tags corresponding to individual words.

There is also another problem due to Basque morphology, to establish the linking between the output of the morphological analyser and the CG parser. The fact that the morphosyntactic information given for each word-form is very detailed poses some difficulties: a) multiplicity of values for some features, b) words with noun phrase structure, and c) noun ellipsis inside word-forms.

For this reason, we are developing a unification based word-grammar to combine these morphosyntactic features in order to give a more adequate description. Each grammar rule combines information from different morphemes giving as a result a feature structure for each interpretation of the word-forms. This word-grammar will give a more optimized output, in the sense that it will facilitate the design of the rules. This is similar to the approach taken by (Ritchie et al. 87); (Armstrong et al. 95).

Apart from that, there are many cases of unresolved ambiguity. Some of them are treated at the moment, as with subordinative clauses (see Example 2), while others would require semantic or pragmatic information for their resolution, such as *eskola garaia* (‘high school’ / ‘school time’).

6. Conclusions

This paper presents the application of the Constraint Grammar formalism to Basque, currently under development. We have presented the design, implementation and test of rules for morphosyntactic disambiguation. The results are satisfactory in the case of disambiguating the full morphosyntactic description, where a 97.5% accuracy is obtained at the cost of maintaining part of the ambiguity in the result. The results improve considerably, reaching 99% accuracy, when only categorial disambiguation is performed. The work also shows us the tight relationship between morphosyntactic 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 and their invaluable help.

This research was supported by the Basque Government, the University of the Basque Country and the Interministerial Commision for Science and Technology (CICYT).

References

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

(Agirre et al. 92) 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 et al. 95) 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 et al. 94) 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 et al. 95) Alegria I. *Euskal morfologiaren tratamendu automatikorako tresnak* Ph.D. thesis University of the Basque Country, 1995

(Alegria et al. 95) Alegria I., Artola X., Sarasola K. *Improving a robust morphological analyzer using lexical transducers* RANLP, Bulgaria. 1995

(Alegria et al. 96a) Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Aduriz I. *A Corpus-Based Morphological Disambiguation Tool for Basque,* Actas del XII. Congreso de la SEPLN Sevilla, 1996.

(Alegria et al. 96b) Alegria I., Sarasola K., Urkia M. *Automatic morphological analysis of Basque,* Literary and Linguistic Computing, Vol. 11, Nº. 4, 1996.

(Armstrong et al. 95) Armstrong S., Russel G., Petipierre D., Robert G. *An open architecture for multilingual text processing,* Proceedings. of the 7th Conf. of the EACL, 1995

(Black et al. 91) 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 95) Elworthy D. *Tagset Design and Inflected Languages* From Texts to Tags: Issues in Multilingual Text Analysis. ACL SIGDAT Workshop, Dublin, 1995

(Gojenola & Sarasola 94) Gojenola K., Sarasola K. *Aplicaciones de la relajación gradual de restricciones para la detección y corrección de errores sintácticos* SEPLN, Córdoba, 1994

(ITEM 97) *Estándares de codificación morfológica en ITEM* ITEM 001-1/97 Technical Report 1997

(Karlsson et al. 95) Karlsson F., Voutilainen A., Heikkila J., Anttila A. *Constraint Grammar: Language-independent System for Parsing Unrestricted Text* Mouton de Gruyter. 1995

(Koskenniemi 83) Koskenniemi K. *Two-level Morphology: A general Computational Model for*

Word-Form Recognition and Production Ph D. thesis, University of Helsinki. 1983

(Ritchie et al. 87) Ritchie G., Pullman S..G., Black A.W., Russel G.J. *A computational framework for lexical description* Computational Linguistics, Vol. 13, 1987

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

(Tapanainen 96) Tapanainen P. *The Constraint Grammar Parser CG-2* . University of Helsinki. Publications nº 27, 1996

(Urkia & Sagarna 91) Urkia M., Sagarna A. *Terminología y Lexicografía Asistida por Ordenador. La experiencia de UZEI* Actas del VII Congreso SEPLN, 1991

(Voutilainen 94a) Voutilainen, A. *Three studies of grammar-based surface parsing of unrestricted English text* Ph.D. thesis. University of Helsinki. Publications nº 24, 1994

(Voutilainen 94b) Voutilainen, A. *Designing a parsing grammar* University of Helsinki. Publications nº 22, 1994

(Voutilainen & Järvinen 95a) Voutilainen A, Järvinen T. *Specifying a shallow grammatical representation for grammatical purposes* Proceedings of the 7th Conference of EACL, Dublin, 1995

(Voutilainen 95b) Voutilainen A. *A syntax-based part-of-speech analyser* Proceedings of the 7th Conference of the EACL, Dublin, 1995