A word-grammar based morphological analyzer for agglutinative languages

Aduriz I.[‡], Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J. M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M.[†]

Dept. of Computer Languages and Systems, University of the Basque Country, 649 P. K., E-20080 Donostia, Basque Country

[†]UZEI, Aldapeta 20, E-20009 Donostia, Basque Country

[‡]Universidad de Barcelona, Gran Vía de las Cortes Catalanas, 585, E-08007 Barcelona

jipgogak@si.ehu.es.

Abstract

Agglutinative languages present rich morphology and for some applications they need deep analysis at word level. The work here presented proposes a model for designing a full morphological analyzer.

The model integrates the two-level formalism and a unification-based formalism. In contrast to other works, we propose to separate the treatment of *sequential* and *non-sequential morphotactic constraints*. Sequential constraints are applied in the segmentation phase, and non-sequential ones in the final feature-combination phase. Early application of sequential morphotactic constraints during the segmentation process makes feasible an efficient implementation of the full morphological analyzer.

The result of this research has been the design and implementation of a full morphosyntactic analysis procedure for each word in unrestricted Basque texts.

Introduction

Morphological analysis of words is a basic tool for automatic language processing, and indispensable when dealing with highly agglutinative languages like Basque (Aduriz *et al*., 98b). In this context, some applications, like spelling correction, do not need more than the segmentation of each word into its different component morphemes along with their morphological information. However, there are other applications such as lemmatization, tagging, phrase recognition, and determination of clause boundaries (Aduriz *et al*., 95), which need an additional *global morphological parsing*[1] of the whole word.

Such a complete morphological analyzer has to consider three main aspects (Ritchie *et al*., 92; Sproat, 92):

1) Morphographemics (also called morphophonology). This term covers orthographic variations that occur when linking morphemes.
2) Morphotactics. Specification of which morphemes can or cannot combine with each other to form valid words.
3) Feature-combination. Specification of how these morphemes can be grouped and how their morphosyntactic features can be combined.

The system here presented adopts, on the one hand, the two-level formalism to deal with morphographemics and sequential morphotactics (Alegria *et al*., 96) and, on the other hand, a unification-based word-grammar[2] to combine the grammatical information defined in morphemes and to tackle complex morphotactics. This design allowed us to develop a full coverage analyzer that processes efficiently unrestricted texts in Basque.

The remainder of this paper is organized as follows. After a brief description of Basque morphology, section 2 describes the architecture for morphological processing, where the morphosyntactic component is included. Section 3 specifies the phenomena covered by the analyzer, explains its design criteria, and presents implementation and evaluation details. Section 4 compares the

[1] This has also been called *morphosyntactic parsing*. When we use the term *morphosyntax* we will always refer to the hierarchical structure at word level, combining morphology and syntax.

[2] The term *word-grammar* should not be confused with the syntactic theory presented in (Hudson, 84).

system with previous works. Finally, the paper ends with some concluding remarks.

1 Brief description of Basque morphology

These are the most important features of Basque morphology (Alegria *et al*., 96):

- 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 a single noun a minimum of 135 inflected forms can be generated. Therefore, the number of simple word-forms covered by the current 70,000 dictionary entries would not be less than 10 million.
- 77 of the inflected forms are simple combinations of number, determination, and case marks, not capable of further inflection, but the other 58 word-forms ending in one of the two possible genitives (possessive and locative) can be further inflected with the 135 morphemes. This kind of recursive construction reveals a noun ellipsis inside a 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. Taking into account a single level of noun ellipsis, the number of word-forms could be estimated over half a billion.
- Verbs offer a lot of grammatical information. A verb form conveys information about the subject, the two objects, as well as the tense and aspect. For example: *diotsut* (Eng.: *I am telling you something*).
- Word-formation is very productive in Basque. It is very usual to create new compounds as well as derivatives.

As a result of this wealth of information contained within word-forms, complex structures have to be built to represent complete morphological information at word level.

2 An architecture for the full morphological analyzer

The framework we propose for the morphological treatment is shown in Figure 1. The morphological analyzer is the front-end to all present applications for the processing of Basque texts. It is composed of two modules: the segmentation module and the morphosyntactic analyzer.

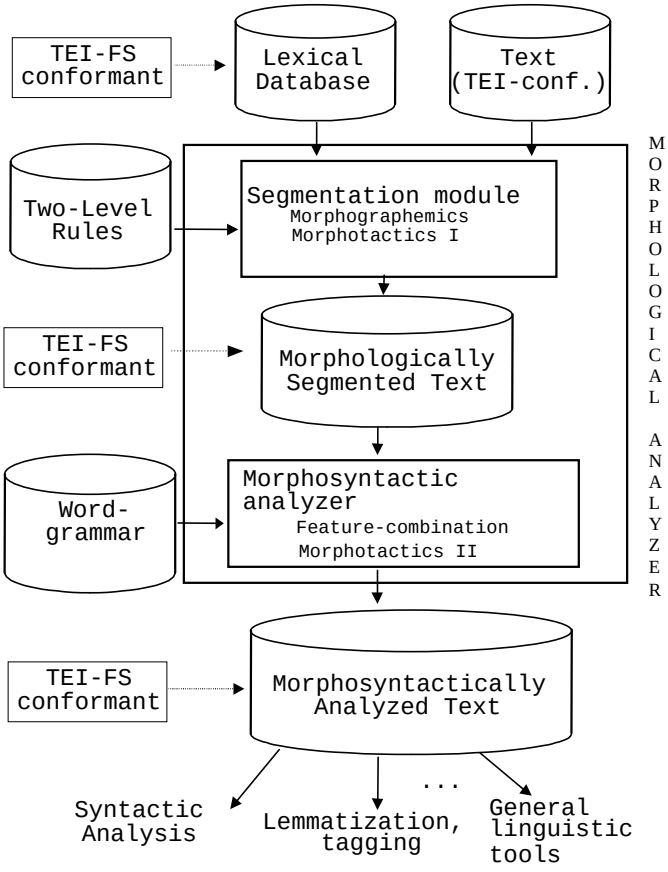

Figure 1. Architecture for morphological processing.

The segmentation module was previously implemented in (Alegria *et al*., 96). This system applies two-level morphology (Koskenniemi, 83) for the morphological description and obtains, for each word, its possible segmentations (one or many) into component morphemes. The two-level system has the following components:

- A set of 24 morphographemic rules, compiled into transducers (Karttunen, 94).
- A lexicon made up of around 70,000 items, grouped into 120 sublexicons and stored in a general lexical database (Aduriz *et al*., 98a).

This module has full coverage of free-running texts in Basque, giving an average number of 2.63 different analyses per word. The result is the set of possible morphological segmentations of a word, where each morpheme is associated with its corresponding features in the lexicon: part of speech (POS), subcategory, declension case, number, definiteness, as well as syntactic function and some semantic features. Therefore, the output of the segmentation phase is very rich, as shown in Figure 2 with the word *amarengan* (Eng.: *on the mother*).

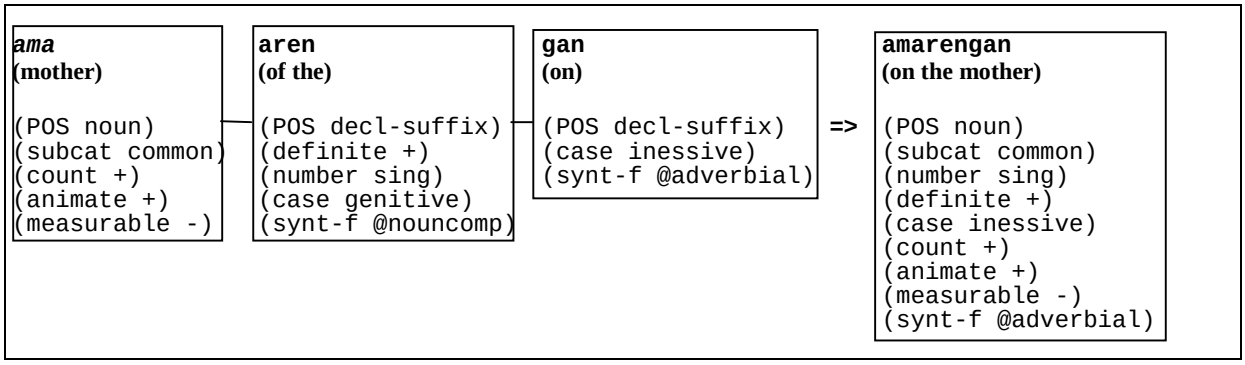

Figure 2. Morphosyntactic analysis[3] of *amarengan* (Eng.: *on the mother*)

The architecture is a modular environment that allows different types of output depending on the desired level of analysis. The foundation of the architecture lies in the fact that TEI-conformant SGML has been adopted for the communication among modules (Ide and Veronis, 95). Feature structures coded according TEI are used to represent linguistic information, including the input and output of the morphological analyzer. This representation enables the use of SGML-aware parsers and tools, and can be easily filtered into different formats (Artola et al., 00).

3 Word level morphosyntactic analysis

This section first presents the phenomena that must be covered by the morphosyntactic analyzer, then explains its design criteria, and finally shows implementation and evaluation details.

3.1 Phenomena covered by the analyzer

There are several features that emphasized the need of morphosyntactic analysis in order to build up word level information:

1) Multiplicity of values for the same feature in successive morphemes. In the analysis of Figure 2 there are two different values for the POS (noun and declension suffix), two for the case (genitive and inessive), and two for the syntactic function (@nouncomp and @adverbial). Multiple values at morpheme-level will have to be merged to obtain the word level information.
2) Words with phrase structure. Although the segmentation is done for isolated words, independently of context, in several cases the resulting structure is equivalent to the analysis of a phrase, as can be seen in Figure 2. In this case, although there are two different cases (genitive and inessive), the case of the full word-form is simply inessive.
3) Noun ellipsis inside word-forms. A noun ellipsis can occur within the word (occasionally more than once). This information must be made explicit in the resulting analysis. For example, Figure 3 shows the analysis of a single word-form like *diotsudanarekin* (Eng.: *with what I am telling you*). The first line shows its segmentation into four morphemes (`diotsut+en+0+arekin`). The feature `comp1` in the final analysis conveys the information for the verb (*I am telling you*), that carries information about person, number and case of subject, object and indirect object. The feature `comp2` represents an elided noun and its declension suffix (*with*).
4) Derivation and composition are productive in Basque. There are more than 80 derivation morphemes (especially suffixes) intensively used in word-formation.

3.2 Design of the word-grammar

The need to impose hierarchical structure upon sequences of morphemes and to build complex constructions from them forced us to choose a unification mechanism. This task is currently unsolvable using finite-state techniques, due to the growth in size of the resulting network (Beesley, 98). We have developed a unification based word-grammar, where each rule combines information from different morphemes giving as a result a feature structure for each interpretation of a word-form, treating the previously mentioned cases.

[3] Feature values starting with the "@" character correspond to syntactic functions, like @nouncomp (noun complement) or @adverbial.

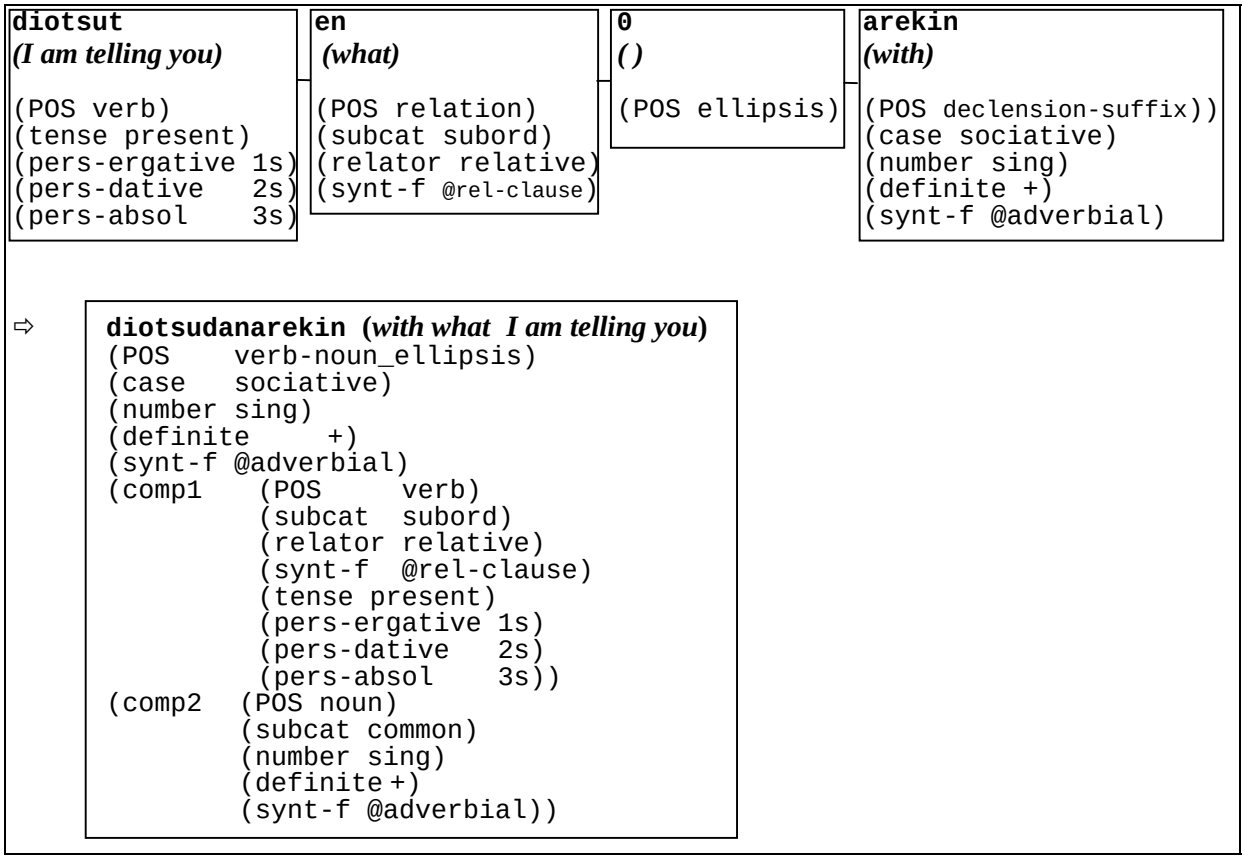

Figure 3. Morphosyntactic analysis of *diotsudanarekin* (Eng.: *with what I am telling you*)

As a consequence of the rich morphology of Basque we decided to control morphotactic phenomena, as much as possible, in the morphological segmentation phase. Alternatively, a model with minimal morphotactic treatment (Ritchie *et al*., 92) would produce too many possible analyses after segmentation, which should be rejected in a second phase. Therefore, we propose to separate sequential morphotactics (i.e., which sequences of morphemes can or cannot combine with each other to form valid words), which will be recognized by the two-level system by means of continuation classes, and non-sequential morphotactics like *long-distance dependencies* that will be controlled by the word-grammar.

The general linguistic principles used to define unification equations in the word-grammar rules are the following:

1) Information risen from the lemma. The POS and semantic features are risen from the lemma. This principle is applied to common nouns, adjectives and adverbs. The lemma also gives the number in proper nouns, pronouns and determiners (see Figure 2).
2) Information risen from case suffixes. Simple case suffixes provide information on declension case, number and syntactic function. For example, the singular genitive case is given by the suffix *-ren* in *ama+ren* (Eng.: *of the mother*). For compound case suffixes the number and determination are taken from the first suffix and the case from the second one. First, both suffixes are joined and after that they are attached to the lemma.
3) Noun ellipsis. When an ellipsis occurs, the POS of the whole word-form is expressed by a compound, which indicates both the presence of the ellipsis (always a noun) and the main POS of the word. For instance, the resulting POS is `verb-noun_ellipsis` when a noun-ellipsis occurs after a verb. All the information corresponding to both units, the explicit lemma and the elided one, is stored (see Figure 3).
4) Subordination morphemes. When a subordination morpheme is attached to a verb, the verb POS and its features are risen as well as the subordinate relation and the syntactic function conveyed by the morpheme.
5) Degree morphemes attached to adjectives, past participles and adverbs. The POS and

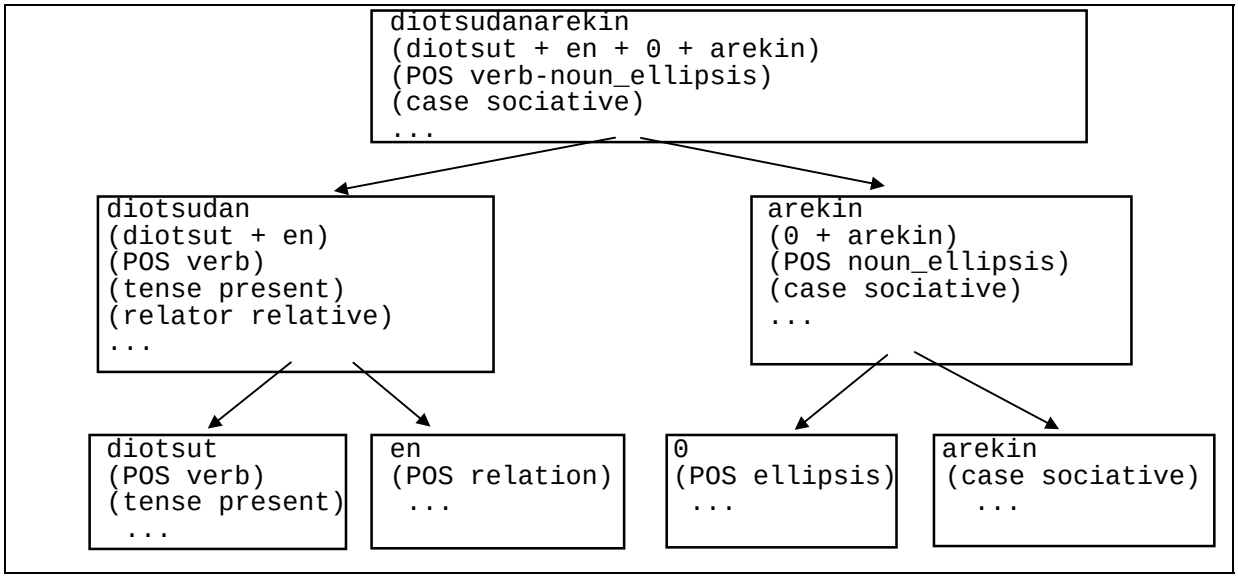

Figure 4. Parse tree for *diotsudanarekin* (Eng.: *with what I am telling you*)

main features are taken from the lemma and the features corresponding to the degrees of comparison (comparative, superlative) are taken from the degree morphemes.

6) Derivation. Derivation suffixes select the POS of the base-form to create the derivative and in most cases to change its POS. For instance, the suffix *-garri* (Eng.: *-able*) is applied to verbs and the derived word is an adjective. When the derived form is obtained by means of a prefix, it does not change the POS of the base-form. In both cases the morphosyntactic rules add a new feature representing the structure of the word as a derivative (root and affixes).
7) Composition. At the moment, we only treat the most frequent kind of composition (noun-noun). Since Basque is syntactically characterized as a right-head language, the main information of the compound is taken from the second element.
8) Order of application of the morphosyntactic phenomena. When several morphosyntactic phenomena are applied to the same lemma, so as to eliminate nonsensical readings, the natural order to consider them in Basque is the following: lemmas, derivation prefixes, derivation suffixes, composition and inflection (see Figure 4).
9) Morphotactic constraints. Elimination of illegal sequences of morphemes, such as those due to long-distance dependencies, which are difficult to restrict by means of continuation classes.

The first and second principles are defined to combine information of previously recognized morphemes, but all the other principles are related to both feature-combination and non-sequential morphotactics.

3.3 Implementation

We have chosen the PATR formalism (Shieber, 86) for the definition of the morpho-syntactic rules. There were two main reasons for this choice:

- The formalism is based on unification. Unification is adequate for the treatment of complex phenomena (e.g., agreement of constituents in case, number and definiteness) and complex linguistic structures.
- Simplicity. The grammar is not linked to a linguistic theory, e.g. GPSG in (Ritchie *et al*., 92). The fact that PATR is simpler than more sophisticated formalisms will allow that in the future the grammar could be adapted to any of them.

25 rules have been defined, distributed in the following way:

- 11 rules for the merging of declension morphemes and their combination with the main categories,
- 9 rules for the description of verbal subordination morphemes,
- 2 general rules for derivation,
- 1 rule for each of the following phenomena: ellipsis, degree of comparison of adjectives (comparative and superlative) and noun composition.

3.4 Evaluation

As a consequence of the size of the lexical database and the extensive treatment of morphosyntax, the resulting analyzer offers full coverage when applied to real texts, capable of treating unknown words and non-standard forms (dialectal variants and typical errors).

We performed four experiments to evaluate the efficiency of the implemented analyzer (see Table 1). A 10,832-word text was randomly selected from newspapers. We measured the number of words per second analyzed by the morphosyntactic analyzer and also by the whole morphological analyzer (results taken on a Sun Ultra 10). In the first experiment all the word-forms were analyzed one-by-one; while in the other three experiments words with more than one occurrence were analyzed only once. In the last two experiments a memory with the analysis of the most frequent word-forms (MFW) in Basque was used, so that only word-forms not found in the MFW were analyzed.

Test description	# analyzed words	words/sec Morphosynt. analyzer	words/sec Full morphological analyzer
All word forms	10,832	15,13	13,5
Different word forms	3,692	44	40
MFW 10,000 words (15 Mb)	1,483	111	95
MFW 50,000 words (75 Mb)	533	308	270

Table 1. Evaluation results.

Even when our language is agglutinative, and its morphological phenomena need more computational resources to build complex and deep structures, the results prove the feasibility of implementing efficiently a full morphological analyzer, although efficiency was not the main concern of our implementation. The system is currently being applied to unrestricted texts in real-time applications.

4 Related work

(Koskeniemmi, 83) defined the formalism named two-level morphology. Its main contribution was the treatment of morphographemics and morphotactics. The formalism has been successfully applied to a wide variety of languages.

(Karttunen, 94) speeds the two-level model compiling two-level rules into lexical transducers, also increasing the expressiveness of the model

The morphological analyzer created by (Ritchie *et al*., 92) does not adopt finite state mechanisms to control morphotactic phenomena. Their two-level implementation incorporates a straightforward morphotactics, reducing the number of sublexicons to the indispensable (prefixes, lemmas and suffixes). This approximation would be highly inefficient for agglutinative languages, as it would create many nonsensical interpretations that should be rejected by the unification phase. They use the word-grammar for both morphotactics and feature-combination.

In a similar way, (Trost, 90) make a proposal to combine two-level morphology and non-sequential morphotactics.

The PC-Kimmo-V2 system (Antworth, 94) presents an architecture similar to ours applied to English, using a finite-state segmentation phase before applying a unification-based grammar.

(Prószéky and Kis, 99) describe a morpho-syntactic analyzer for Hungarian, an agglutinative language. The system does not use the two-level model for segmentation, precompiling suffix-sequences to improve efficiency. They claim the need of a word-grammar, giving a first outline of its design, although they do not describe it in detail.

(Oflazer, 99) presents a different approach for the treatment of Turkish, an agglutinative language, applying directly a dependency parsing scheme to morpheme groups, that is, merging morphosyntax and syntax. Although we are currently using a similar model to Basque, there are several applications that are word-based and need full morphological parsing of each word-form, like the word-oriented Constraint Grammar formalism for disambiguation (Karlsson *et al*., 95).

Conclusion

We propose a model for full morphological analysis integrating two different components. On the one hand, the two-level formalism deals with morphographemics and sequential morphotactics and, on the other hand, a

unification-based word-grammar combines the grammatical information defined in morphemes and also handles complex morphotactics.

Early application of sequential morphotactic constraints during the segmentation process avoids an excessive number of meaningless segmentation possibilities before the computationally more expensive unification process. Unification permits the resolution of a wide variety of morphological phenomena, like ellipsis, that force the definition of complex and deep structures to represent the output of the analyzer.

This design allowed us to develop a full coverage analyzer that processes efficiently unrestricted texts in Basque, a strongly agglutinative language.

The analyzer has been integrated in a general framework for the processing of Basque, with all the linguistic modules communicating by means of feature structures in accord to the principles of the Text Encoding Initiative.

Acknowledgements

This research was partially supported by the Basque Government, the University of the Basque Country and the CICYT (Comisión Interministerial de Ciencia y Tecnología).

References

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A., Insausti J. M. (1998a) *EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation, Granada.

Aduriz I., Agirre E., Aldezabal I., Alegria I., Ansa O., Arregi X., Arriola J.M., ArtolaX., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Maritxalar M., Oronoz M., Sarasola K., Soroa A., Urizar R., Urkia M. (1998b) *A Framework for the Automatic Processing of Basque*. Proceedings of the First International Conference on Language Resources and Evaluation, Granada.

Aduriz I., Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Eceiza N., Gojenola K., Maritxalar M. (1995) *Different Issues in the Design of a lemmatizer/Tagger for Basque*. From Texts to Tags: Issues in Multilingual Language Analysis. ACL SIGDAT Workshop, Dublin.

Alegria I., Artola X., Sarasola K., Urkia M. (1996) *Automatic morphological analysis of Basque*. Literary and Linguistic Computing, 11 (4): 193-203. Oxford University.

Antworth E. L. (1994) *Morphological Parsing with a Unification-based Word Grammar*. North Texas Natural Language Processing Workshop, Texas.

Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K., Maritxalar A., Soroa A. (2000) *A proposal for the integration of NLP tools using SGML-tagged documents.* Proceedings of the Second Conference on Language Resources and Evaluation (LREC 2000). Athens, Greece 2000.

Beesley K. (1998) *Arabic Morphological Analysis on the Internet*. Proceedings of the International Conference on Multi-Lingual Computing (Arabic & English), Cambridge.

Hudson R. (1990) *English Word Grammmar*. Oxford: Basil Blackwell.

Ide N., Veronis J. K. (1995) *Text-Encoding Initiative, Background and Context*. Kluwer Academic Publishers.

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

Karttunen L. (1994) *Constructing Lexical Transducers*. Proc. of COLING´94, 406-411.

Koskenniemi, K. (1983) *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*, University of Helsinki, Department of General Linguistics. Publications nº 11.

Oflazer K (1999) *Dependency Parsing with an Extended Finite State Approach*. ACL'99, Maryland.

Prószéky G., Kis B (1999) *A Unification-based Approach to Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages*. ACL'99, Maryland.

Ritchie G., Pullman S. G., Black A. W., Russel G. J. (1992) *Computational Morphology: Practical Mechanisms for the English Lexicon*. ACL-MIT Series on Natural Language Processing, MIT Press.

Shieber S. M. (1986) A*n Introduction to Unification-Based Approaches to Grammar*. CSLI, Stanford.

Sproat R. (1992) *Morphology and Computation*. ACL-MIT Press series in Natural Language Processing.

Trost H. (1990) *The application of two-level morphology to non-concatenative German morphology*. COLING'90, Helsinki.