A Word-Level Morphosyntactic Analyzer for Basque

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

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

* Universidad de Barcelona
Gran Vía de las Cortes Catalanas, 585
E-08007 Barcelona

*† UZEI, Aldapeta 20
E-20009 Donostia, Basque Country
jipgogak@si.ehu.es

Abstract

This work presents the development and implementation of a full morphological analyzer for Basque, an agglutinative language. Several problems (phrase structure inside word-forms, noun ellipsis, multiplicity of values for the same feature and the use of complex linguistic representations) have forced us to go beyond the morphological segmentation of words, and to include an extra module that performs a full morphosyntactic parsing of each word-form. A unification-based word-level grammar has been defined for that purpose. The system has been integrated into a general environment for the automatic processing of corpora, using TEI-conformant SGML feature structures.

1. Introduction

Morphological analysis of words is an indispensable basic tool when defining a general framework for the automatic processing of agglutinative languages like Basque (Aduriz et al., 98b). In this context, some applications 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, which need an additional global morphological parsing of the whole word. Several problems arise when trying to compose information from different morphemes. This fact reinforced our view on the need for deep morphosyntactic analysis (merging morphology and syntax) at word-level. This agrees with previous work by (Ritchie et al., 92).

As can be seen in Figure 1, the system for morphological processing takes as input text, lexical information (Aduriz et al., 98a) and two level rules (Alegria et al., 96) as other classical analyzers based on two-level morphology, but it also needs a description of the ways to compose morpheme information by means of a word-grammar.

The remainder of this paper is organized as follows. After an overview of Basque morphology, section 3 describes the main problems in the automatic treatment of its morphology. Section 4 specifies the framework for morphosyntactic analysis. Finally, the paper ends with some concluding remarks.

2. Overview of Basque morphology

The most important morphosyntactic features of Basque are the following (Alegria et al., 96):

a) It is an agglutinative language.
b) As prepositional functions are realized by case suffixes inside word-forms, Basque presents a relatively high power to generate word-forms. The number of simple word-forms covered by the 70,000 dictionary entries in our lexical database would not be less than 10 million. Taking into

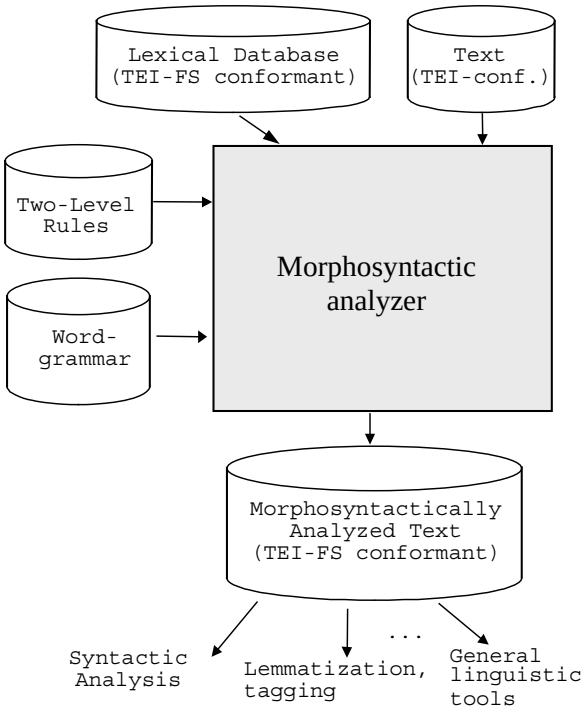

Figure 1. Morphosyntactic processing

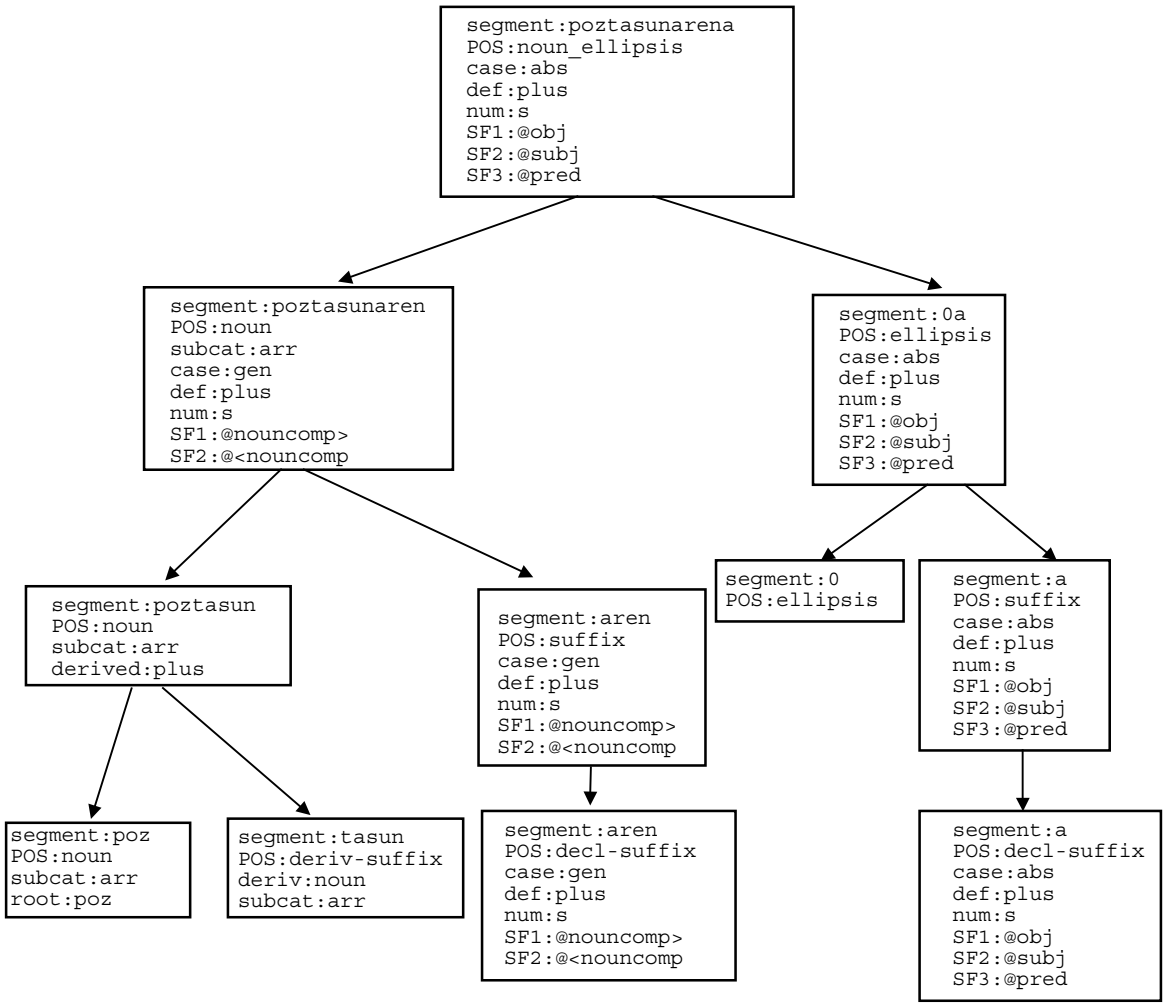

Figure 2. Analysis of *poztasunarena* (Eng.: *the one of the happiness*)

account a single level of noun ellipsis (actually, noun ellipsis may appear recursively inside a word-form), this number could be estimated over half a billion.

c) A verb form contains information about tense, aspect, and agreement (with the subject and the two objects).

d) Word-formation is very productive in Basque. It is very usual to create new compounds as well as derivatives.

3. The main problems in the automatic treatment of Basque morphology

There are four aspects that emphasized the need of morphosyntactic analysis in order to build up word-level information:

- The same feature appears in successive morphemes with different values. Here the problem is to determine which is the value for this feature at word-level. In the analysis of Figure 2 there are several values for the POS (noun, derivation-suffix, declension suffix and ellipsis), two for the case (genitive and absolutive), and many for the syntactic function (@nouncomp>, @<nouncomp, @subj, @obj and @pred).
- Words with phrase structure. In several cases the resulting structure for a single word is equivalent to the analysis of a whole phrase. The word analyzed in Figure 2 (*poztasunarena*) is equivalent to an entire noun phrase (*the one of the happiness*).
- 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. The leaves in the tree in Figure 2 show the five component morphemes (`poz+tasun+aren+0+a`) of *poztasunarena*, where the null morpheme ('`0`') reveals that a noun has been elided.
- Complex linguistic representation. The need to impose hierarchical structure upon sequences of morphemes and to build complex constructions from them forced us to choose a unification mechanism. Actually the feature structures in Figure 2 have been considerably simplified. Figure 3 shows the complete representation of the derived word *poztasun* situated at the bottom left of the tree in Figure 2.

4. The framework for morphosyntactic analysis

The framework for morphosyntactic analysis is shown in Figure 1. The morphological analyzer is composed of two complementary modules, that are applied sequentially to the input text (Aduriz et al., 2000).

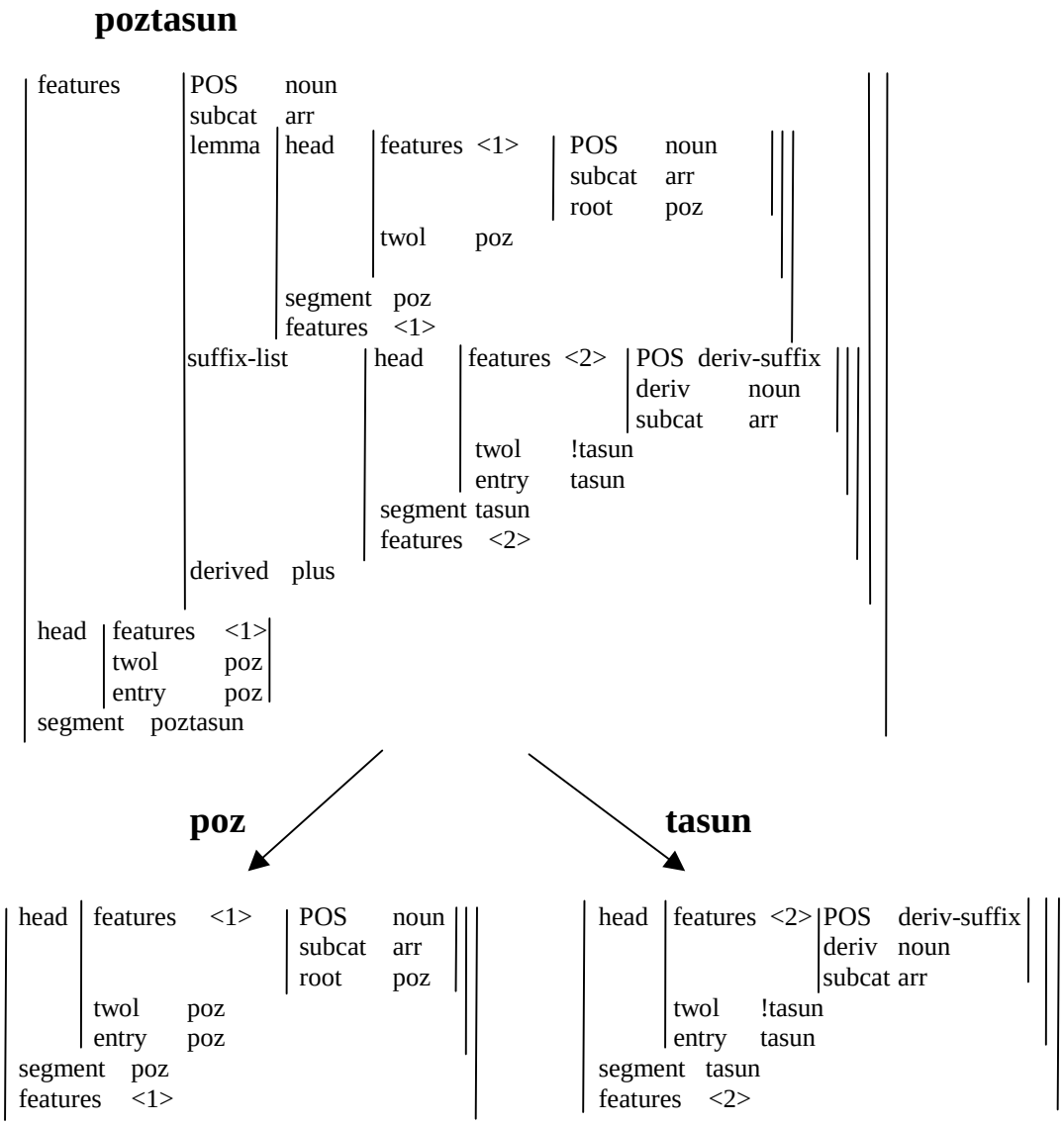

Figure 3: An example of complex morphological representation

The segmentation module, based on two-level morphology, produces the set of all the possible morphological segmentations of a word, where each morpheme is associated with its corresponding features in the lexicon: category, subcategory, declension case, number, definiteness, as well as syntactic function and some semantic features.

The PATR-II formalism was used for the definition of the morphosyntactic word-grammar. It offers adequacy for the treatment of complex phenomena, like agreement of constituents in case, number and definiteness. This is also useful for the manipulation of complex linguistic structures. Twenty-five 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.

As this morphological analyzer has to be integrated in a general environment for the automatic processing of the language (Artola et al., 2000), TEI-conformant SGML has been adopted for the communication among modules (Ide and Veronis, 95). Feature structures (FS) coded according TEI are used to represent linguistic information, including the input and output of the morphological analyzer. The use of SGML for encoding the I/O streams flowing between programs forces us to formally describe the mark-up, and provides software to check that these mark-up hold invariantly in an annotated corpus.

Figure 4 shows an example of the output of the analyzer. Four different files represent the result of the morphosyntactic analysis of an input text. It allows us to store different analysis sets (segmentations, complete morphosyntactic analyses, lemmatization results, and so on) linked to a tokenized piece of text, in which any particular analysis feature structure will not have to be repeated.

Having an SGML-tagged input text file (*.sgm*), the tokenizer takes this file and creates, as output, a *.w.sgm* file, which contains the list of the tokens recognized in the

input text. The tokenized text (*.w.sgm*) is of great importance in the rest of the analysis process, in the sense that it intervenes as input for different processes.

After the tokenization process, the morphosyntactic treatment module takes as input the tokenized text and the general lexicon issued from the lexical database, and produces two documents: the collection of morphosyntactic analyses (FSs) corresponding to the input text (*.morf.sgm*), and a link file (*.morflnk.sgm*) that contains the links between the tokens in the tokenized text file (*.w.sgm*) and their corresponding morphosyntactic analyses (one or more) in the *.morf.sgm* file.

5. Conclusion

The result of this research has been the design and implementation of a complete morphosyntactic analyzer for each word-form, without losing word-internal descriptions. 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. It processes 270 words per second on a Sun Ultra 10, suitable for our corpus processing needs.

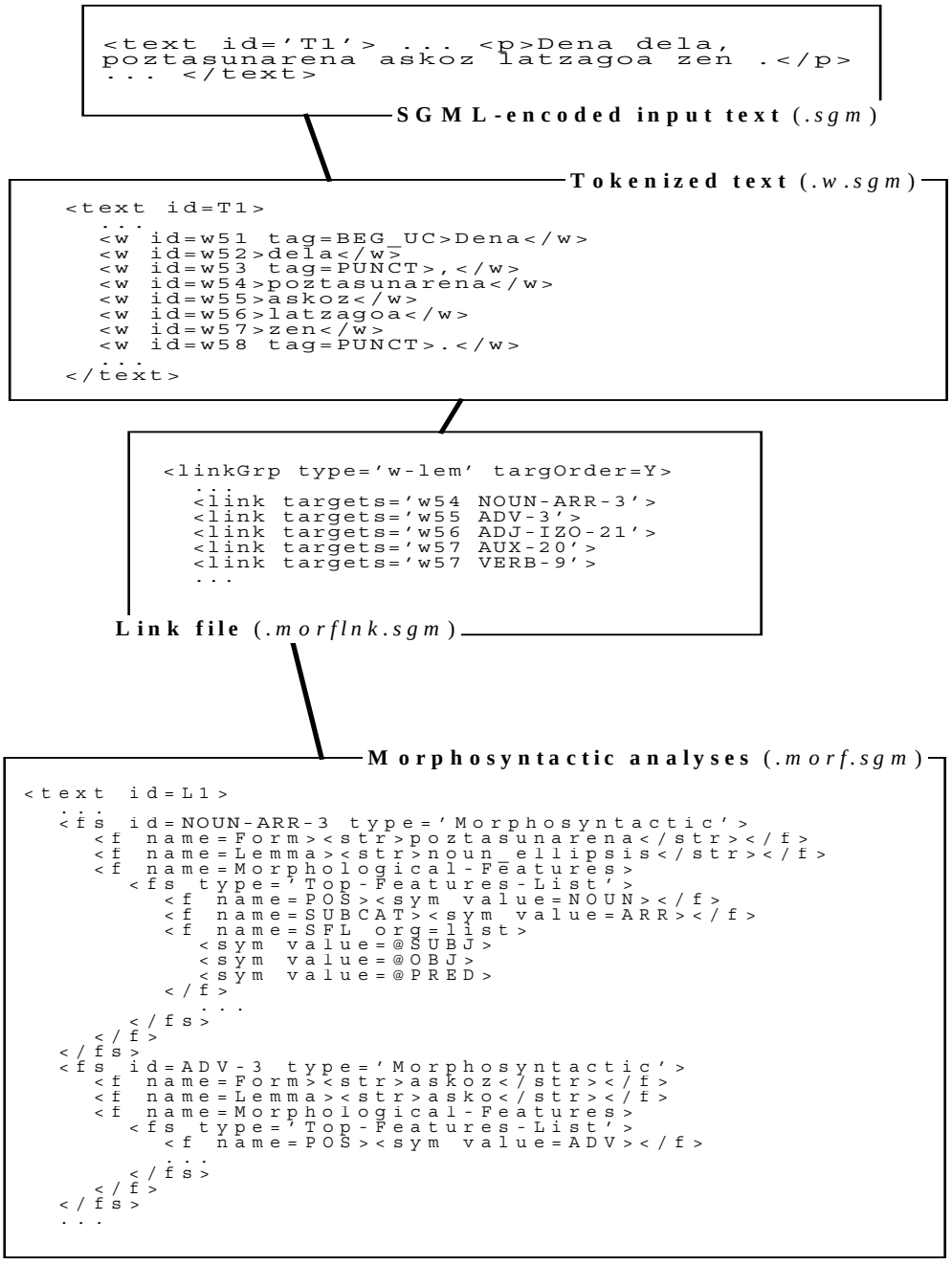

Figure 4. Output of the morphosyntactic analyzer: a sample of the document set

6. Acknowledgements

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

7. References

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.

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., Agirre E., Aldezabal I., Alegria I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar A., Sarasola K., Urkia M. 2000. A word-grammar based morphological analyzer for agglutinative languages. Proceedings of Coling 2000, Saarbrücken. (forthcoming)

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

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. In *Proc. of the Second Int. Conf. on Language Resources and Evaluation*. Athens (Greece).

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

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

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.

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, the MIT Press.

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