Reusability of wide-coverage linguistic resources in the construction of an English-Basque Machine Translation System

Díaz de Ilarraza, A. Mayor and K. Sarasola
IXA Group http://ixa.si.ehu.es/
Informatika Fakultatea (Computer Science Faculty)
University of the Basque Country
DONOSTIA (Basque Country - Spain)
jibmamaa@si.ehu.es

Abstract

In this paper we present a prototype of a machine translation system (English-Basque) based on transfer. The system operates interactively with the user in order to solve some ambiguities. Our aim is to solve these ambiguities by means of different and combined methods (those based on ontologies, examples, Conceptual Density, statistic measures, etc.).

The prototype translates noun and prepositional phrases from English to Basque. It is important to emphasise that the prototype operates with real texts. The treatment of Basque implies to reuse and to adapt wide-coverage linguistic tools and resources for the language developed by our group (IXA group, http://ixa.si.ehu.es); on the other hand, we will take advantage of other tools and resources developed for English and Spanish.

First we will introduce the system, in section 2 we will present the general architecture; section 3 is devoted to explain how the system works by means of an example and, finally, conclusions and future work.

1. Introduction.

In this paper we present a first prototype of a machine translation system (English-Basque) based on transfer. The system translates NP and PP's in real texts and it operates interactively with the user in order to solve ambiguities. Wide-coverage linguistic tools and resources have been integrated in it. The translation process is performed in three phases: analysis, transfer and generation.

The generation phase for Basque, at morphological level, is based on the morphological analyser, MORFEUS (Alegria *et al.* 99; Aduriz *et al.* 99). Generation at syntactic level syntactic is a simplified version of the syntactic

analyser developed by our group (Aldezabal *et al., 99)*.

The analysis phase for the English-Basque version uses ENGCG from Lingsoft Inc.

The transfer modules (one per each language and sense) are built using bilingual dictionaries English-Basque and Spanish-Basque, and the EDBL, a monolingual lexical data base for Basque.

2. General Architecture.

The general architecture of the system is represented in Figure 1.

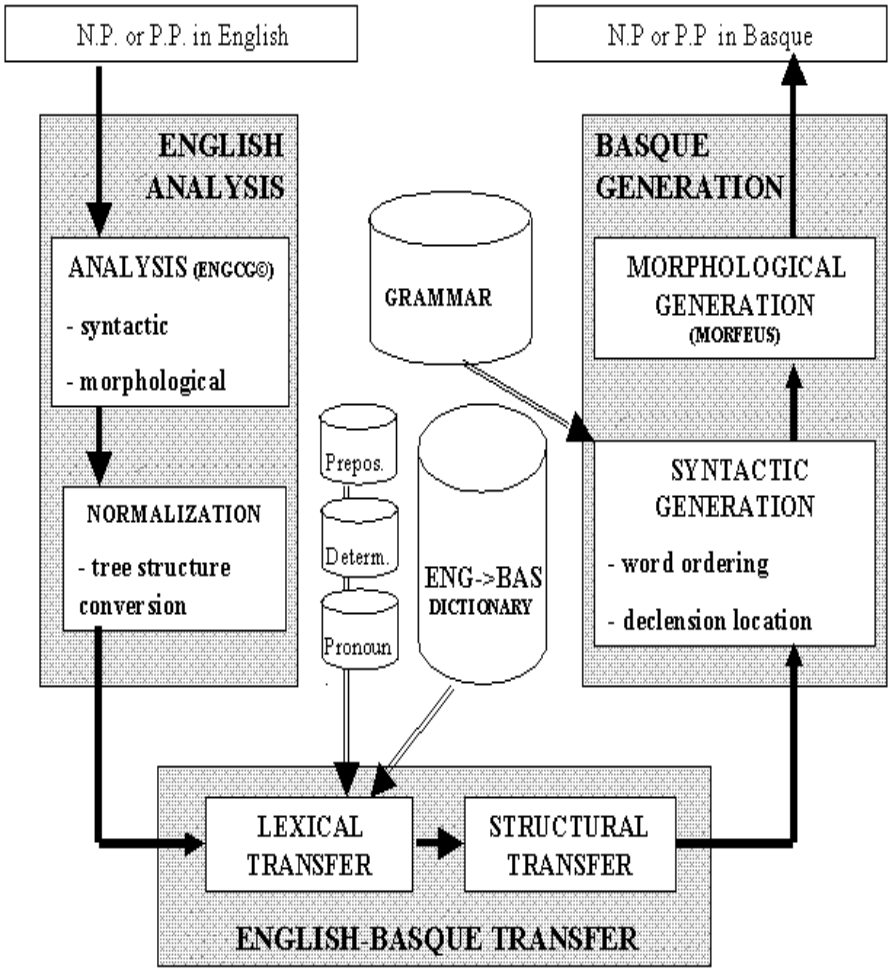

Figure 1. General architecture

Let us explain the different phases:

- The **analysis phase** for English is performed in two steps:

1. Analysis of the NP or PP. We use the morphological analyser by ENGCG (http://www.lingsoft.fi/cgi-pub/engcg/). As an example, let us show the analysis of the following PP: "on the house of these big mountains"

LEMA	MORPHOLOGICAL INF.	SYNTACTIC INF.
on	PREP	ADVL
the	<Def DET CENTRAL ART SG/PL	DN
house	N NOM SG	<P
of	PREP	<NOM-OF
this	DET CENTRAL DEM PL	DN
big	A ABS	AN
mountain	N NOM PL	<P

2. Based on this information we built the tree-structure (Figure 2).

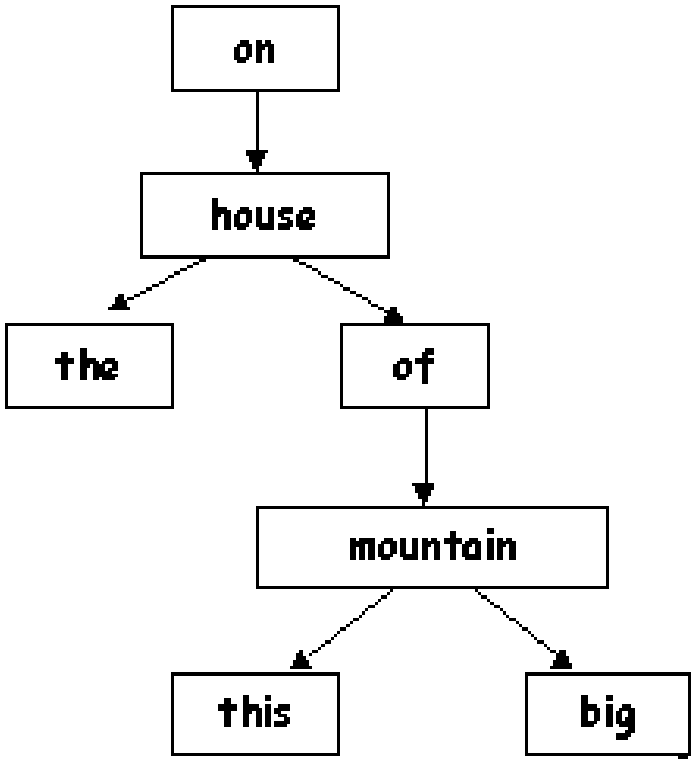

Figure 2. Tree-structure of the analysis of the PP

The **transfer module** is subdivided in two sub-modules:

1. Lexical transfer. Here the electronic version of the bilingual dictionary Morris (English-Basque/Basque-English) is used. The dictionary, analysed and tagged by our group, contains more than 80.000 entries. If there is more than one alternative for the lexical transfer we can chose one of these two possibilities:
2.

- Select the first one because it is the most used one.
- Ask the user to select one alternative.

We plan to use parallel corpus for this task. Figure 3 shows the lexical transfer process. Due to the nature of Basque, we would like to focus on the fact that the nodes corresponding with prepositions and determiners in Basque do not have any lexical value but information about declension.

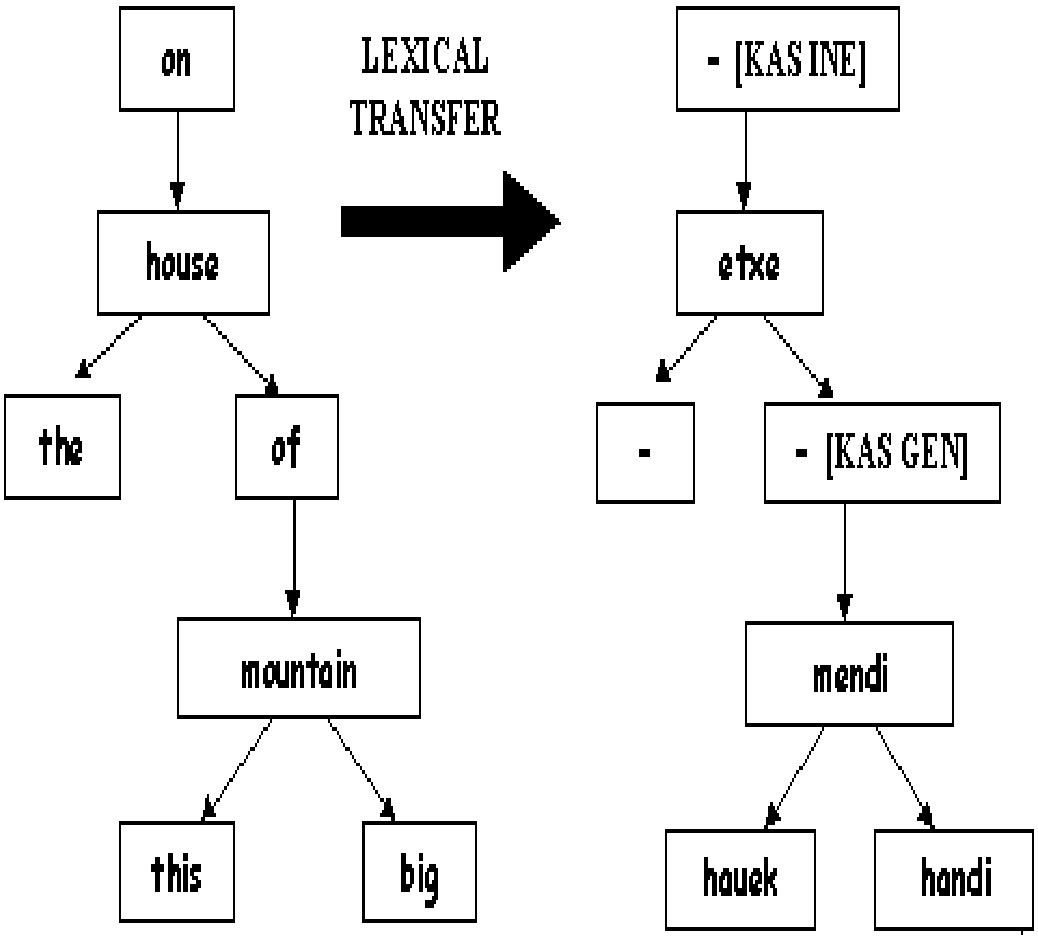

Figure 3. Lexical Transfer

- Structural transfer. Nodes without lexical value disappear and the information associated to them is transferred to their sons as shown in Figure 4.

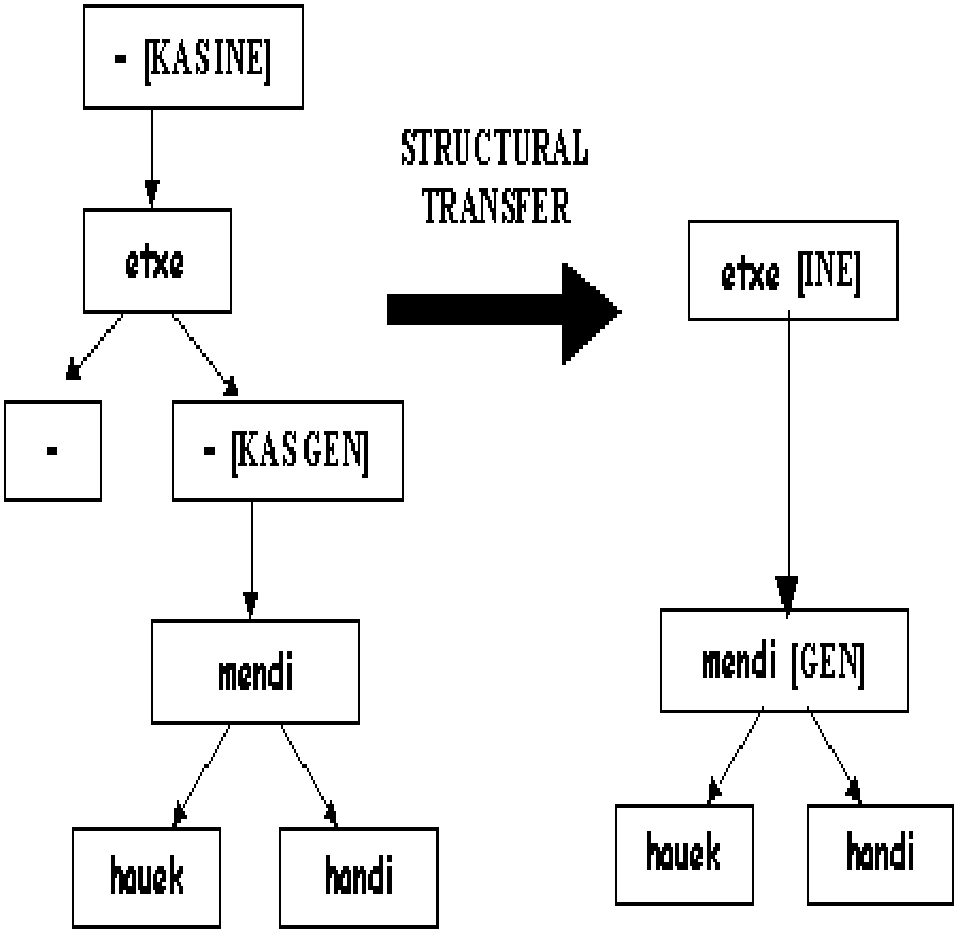

Figure 4. Structural Transfer

The **generation phase** works on two steps also.

■

1. Syntactic generation. It uses a unification grammar that generates nominal and prepositional phrases. The objective of the grammar is to establish the order of words in the phrase. Information about declension is passed and transferred to the last word of the phrase. Figure 5 shows this step.

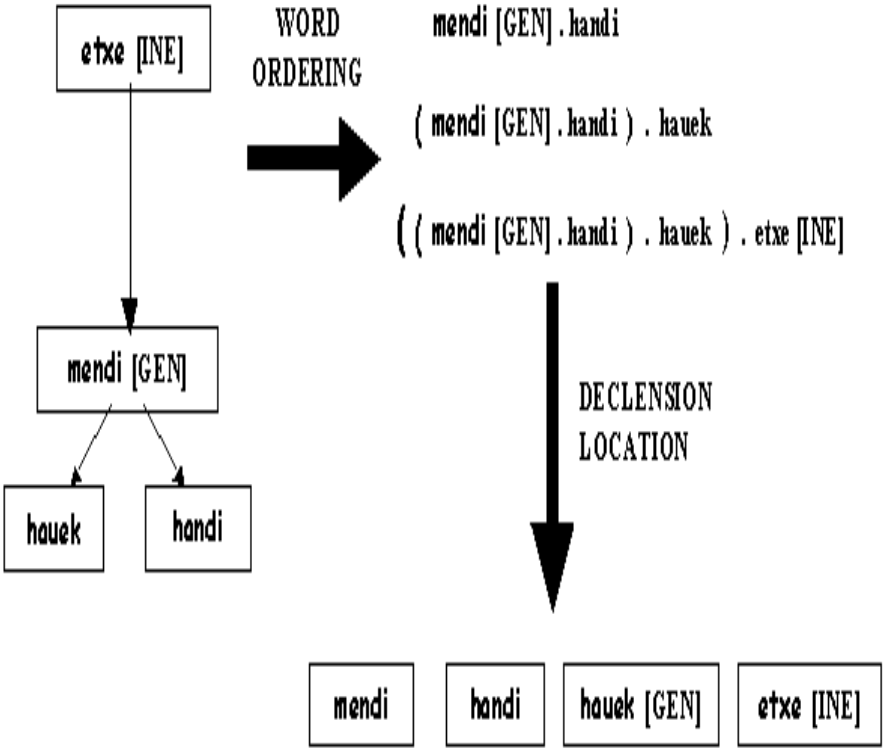

Figure 5. Defining word ordering and declension location

2. Morphological generation. Once we have established the word order we use MORFEUS to generate the declined word-form. Here, we find some ambiguity problems because in Basque, apart from other grammatical features, semantic information (living/not living) about the core of the phrase can be needed in order to generate the appropriate word-form (Figure 6).

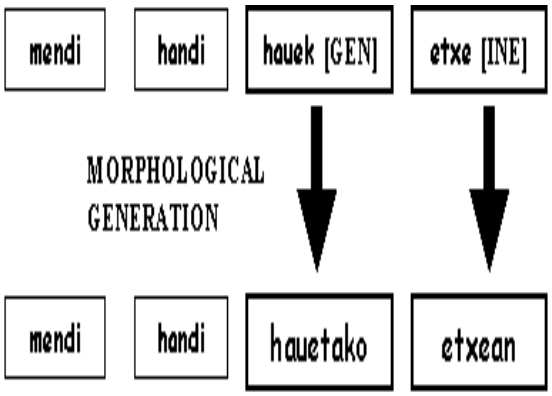

Figure 6. Morphological Generation

3. Conclusions and future work.

A first prototype for a machine translation system has been presented. At present, the prototype works from English to Basque. As a first step, it translates NP and PPs. The result has been positive. In our opinion, it is an interesting step for a minority language like Basque.

The modularity of the architecture presented allows us to introduce easily new modules to improve its functionality. With this prototype we prove the reusability and adequacy of resources and tools previously developed.

Our near objectives are focussed on the treatment of more complex NP and PP and enhance the prototype for translating sentences. On the other hand, the study of methods for disambiguation will constitute and important advance.

4. References

[Agirre E., Ansa O., Arregi X., Arriola J.M., Díaz de Ilarraza A., Lersundi M., Soroa A., Urizar R.]
"Extracción de relaciones semánticas mediante gramáticas de restricciones"
Congreso SEPLN98.Alicante. Spain. 1998

[Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Maritxalar M., Urkia M.]
"Euskararako murriztapen-gramatika: Lehen urratsak". 1996.
UPV-EHU/ LSI/ TR 2-96

[Aduriz I., Agirre E., Aldezabal I., Arregi X., Arriola J.M., Artola X., Gojenola K., Maritxalar A., Maritxalar M., Sarasola K., Urkia M.]
"MORFEUS: Euskararako analizatzaile morfosintaktikoa"
UPV-EHU/ LSI/ TR 1-99

[Aldezabal I., Gojenola K., Oronoz M.]
"Combining Chart-Parsing and Finite State Parsing". 1999.
http://ixa.si.ehu.es/dokument/Artikulu/99acl-stu.ps

[Alegria I., Artola X., Sarasola K., Urkia M.]
"Automatic morphological analysis of Basque"
Literary & Linguistic Computing Vol. 11, No. 4, 193-203. Oxford University Press. 1996.

[P. Brown, J. Cocke, S.A. Della Pietra, V.J. Della Pietra, F. Jelinek, J.D. Lafferty, R.L. Mercer, P.S. Roossin]
"A Statistical Approach to Machine Translation"
Computational Linguistics 16, 79-85. 1990.

[Cole R., Mariani J., Uszkoreit H., Varile G.B., Zaenen A., Zampolli A., Zue V.]
"Survey of the State of the Art in Human Language Technology"
Studies in Natural Language Processing. Cambridge University Press. 1997

[Hutchins W., Somers H.]
"An Introduction to Mahcine Translation"

Academic Press Ltd. 1992

[Somers H.L.]
"Current Research in Machine Translation"
Machine Translation 7, 231-246. 1993.

[Sumita E., Iida H.]
"Experiments and Prospects of Example-Based Machine Translation"
Proceedings of the Association for Computational Linguistics, 185-192. Berkeley 1991.

[Voutilainen A. & Silvonen M.]
"A Short Introduction to ENGCG"
http://www.lingsoft.fi/doc/engcg/intro/.

Dictionaries

Sarasola, I. Hauta-lanerako Euskal Hiztegia. Donostia: KUTXA, 1991.

Elhuyar Hiztegia (Basque-Spanish/Spanish-Basque). Donostia: Elhuyar, 1996.

Morris Histegia (Basque-English/English-Basque). Klaudio Harlouxet Fundazioa. 1999.