Reusability of wide-coverage linguistic resources in the construction of a multilingual Machine Translation System

Arantxa Diaz de Ilarraza, Aingeru Mayor, Kepa Sarasola
(jipdisaa/jibmamaa/jipsagak@si.ehu.es)
IXA Group. Computer Science Faculty. University of the Basque Country

Abstract

We present a prototype of a transfer-based multilingual machine translation system from English and Spanish to Basque (a minority language). The system translates NPs and PPs in real texts solving interactively most of the ambiguities. One of our goals is to reuse and adapt wide-coverage tools and resources.

1 Introduction

In this paper we present a first prototype of a multilingual machine translation system (English-Basque and Spanish-Basque) based on transfer, which is the first system that includes Basque. The system translates NPs and PPs 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. Although some initial works are in progress (Aldezabal *et al.* 2000), we have not enough information about verbal subcategorization to face the translation of full sentences. The translation process is performed in three phases: analysis, transfer and generation.

The analysis phase for the English-Basque version uses ENGCG from Lingsoft Inc., and, for the Spanish-Basque version it uses MACO, TACAT and RELAX modules from the UPC (Polytechnic University of Catalonia).

The transfer modules (one per language pair) use adequate lexicons that we have built automatically using two bilingual dictionaries, English-Basque and Spanish-Basque, and EDBL (a monolingual lexical database for Basque).

Finally, the generation phase for Basque, at morphological level, is based on the morphological analyzer MORFEUS (Alegria *et al.* 1999; Aduriz *et al.* 1999). Generation at syntactic level is a simplified version of the syntactic analyzer developed by IXA group (Aldezabal *et al.*,1999).

This paper is organized as follows. After a brief description of Basque, section 3 describes the general architecture of our system. Section 4 presents the linguistic databases used and the grammatical rules. Section 5 explains the translation process, and then section 6 shows the final evaluation of the English-Basque prototype. The paper ends with some concluding remarks.

2 Brief Description of Basque

Basque is a Pre-Indo-European language of unknown origin and quite different from the surrounding European languages. Since 1968 the Basque Academy of the Language has been involved in a standardization process. These are some of the most important features of Basque:

- It is an agglutinative language; the determiner, the number and the declension case are appended to the last element of the phrase and always in this order. This

information is valid for all the elements of the phrase. For instance, *semeArEN etxeAN* (in the house of the son):

seme	A	r	EN	etxe	A	N
noun (son)	determiner	epenthetical	genitive	noun (house)	determiner	inessive case

- Basque has only one declension table, i.e., the 15 case suffixes regularly applied whatever the previous elements are.
- Prepositional functions are indicated by case suffixes inside word-forms. Basque presents a relatively high capacity to generate inflected word-forms. Regarding word-structure in Basque we prefer to use the term morphosyntax rather than morphology. For instance, the case morpheme adds syntactic information inside the word-form.
- Derivation and composition are productive in Basque. There are more than 80 derivation morphemes (mainly suffixes) intensively used in word-formation.

3 General Architecture

The general design of the system we have built is represented by the typical transfer schema shown in figure 1.

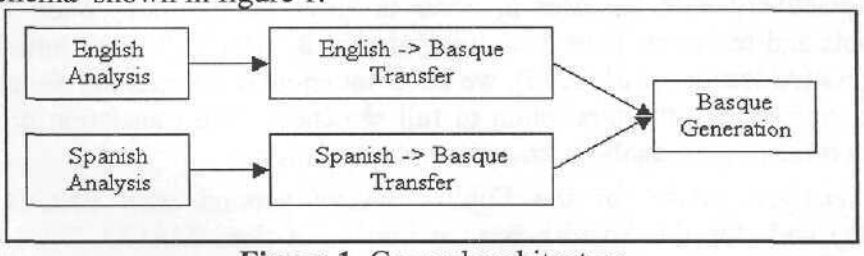

Figure 1. General architecture

We have designed and developed all these modules. The English-Basque part has been evaluated with good results as it will be explained in Section 6. Figure 2 shows the general architecture of the English-Basque prototype that will be explained in sections 4 and 5.

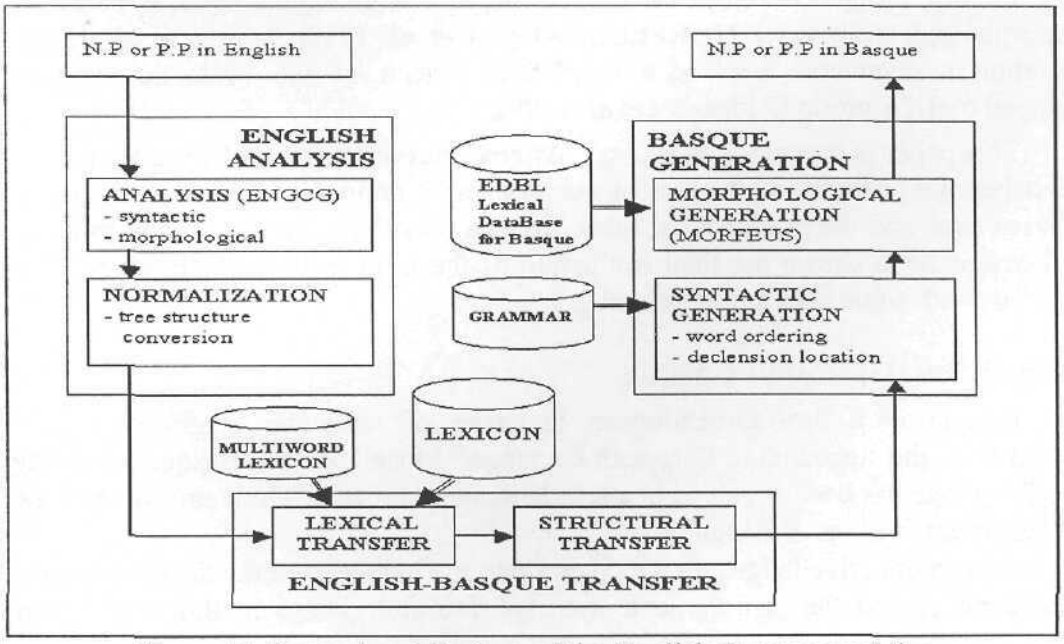

Figure 2. General architecture of the English-Basque modules

4 The Linguistic Databases

Our aim is to attain a clear separation between procedural and declarative aspects of translation. The system uses four types of data: the source monolingual dictionaries, the lexical database for Basque, the bilingual lexicons and the grammar rules.

4.1 The source monolingual dictionaries

The Spanish and English monolingual dictionaries are used by the analysis modules integrated in our prototype. These modules deal with the model of the structure and management of the monolingual dictionaries and lexicons they use.

4.2 EDBL: Lexical Data Base for Basque

The lexical database for Basque EDBL has been designed and implemented by the IXA group (http://ixa.si.ehu.es). This database (Aduriz et al. 1998) is the basis for the development of many linguistic tools; it contains 75,000 entries, corresponding to lemmas and affixes; each entry has its associated linguistic features (category, subcategory, case, number, etc.). This database has been developed under a commercial RDBMS (Oracle V7 manager) running on a UNIX machine. As the Basque language is being lexically standardized nowadays, linguists are permanently updating the EDBL database.

The morphological analyzer and generator use a transducer that has been built by compiling the lexical level exported from the lexical database and the two-level rules, as shown in figure 3

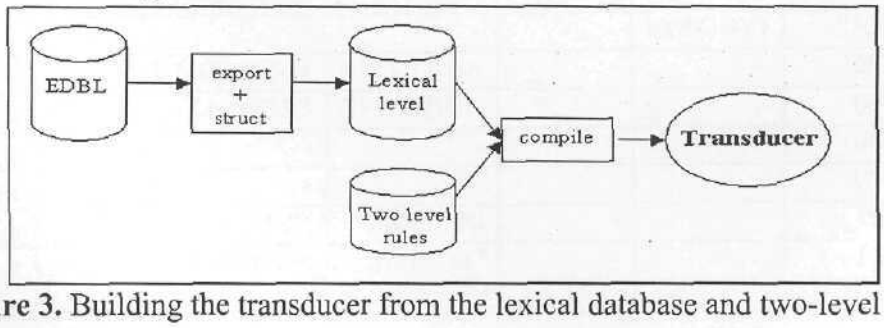

Figure 3. Building the transducer from the lexical database and two-level rules

4.3 The bilingual lexicons

Each entry in our bilingual lexicon is organized as shown in Figure 4. Five features have been distinguished:

Source Lang. Entry	Source Lang. POS	Basque Equivalent	Basque POS	Morphologic Segmentation

Figure 4. Design of the entries in the MT bilingual lexicons.

4.3.1 English-Basque bilingual lexicon

For its construction we designed an automatic process that analyzed the 57,508 Basque outputs contained in the Morris bilingual English-Basque dictionary (Morris, 1999) using the EDBL lexical database. In most of the cases the entry in the EDBL database and the word-form in the bilingual dictionary are identical and both have the same POS, therefore the information is taken from the EDBL database.

Nevertheless, there are some situations that had a special treatment. For example the information associated with the 267 English prepositions that appears in the Morris Dictionary has been coded manually, because the dictionary does not give the information needed for declension, and it cannot be automatically deduced from it.

4.3.2 Spanish-Basque bilingual lexicon

This has not been developed yet, and the system uses the raw electronic version of the Spanish-Basque Elhuyar dictionary. We are now working to build this lexicon in a similar way to the English-Basque one.

4.4 Grammatical rules

The system uses a set of 33 binary rules adapted from the grammar created in Aldeazabal *et al.* (2000). These rules use POS information and, in some cases, subcategory and lemma information to decide the correct order of the words in nominal and prepositional phrases. The order of the elements in a NP/PP in Basque is the following:

```
(PP)* (DET)* (ADJ) NOUN  NOUN) (ADJ) (DET)* DECLENSION
```

These binary rules are executed during the syntactic generation phase. Following are some examples of these rules [x_0 -> x_1 x_2]:

x_0	x_1			x_2			order
	POS	**subcat.**	**lemma**	**POS**	**subcat.**	**lemma**	
NP0	NOUN	COMMON	-	NOUN	COMMON	-	x_2 x_1
NP0	NOUN	COMMON	-	-	-	-	x_1
NP0	NP0	-	-	ADJECTIVE	POST-NOM	-	x_1 x_2
NP0	NP0	-	-	ADJECTIVE	PRE-NOM	-	x_2 x_1
NP1	NP0	-	-	-	-	-	x_1
NP1	ADJ	-	-	-	-	-	x_1
NP1	NP1	-	-	DETERM.	NUMERAL	bat	x_1 x_2
NP1	NP1	-	-	DETERM.	NUMERAL	-	x_2 x_1
...							

5 Translation Processes

As a transfer-based system, the prototype has three phases: analysis, transfer and generation. We present here the English analysis, the English-Basque transfer, and the Basque generation modules.

5.1 Analysis phase

The analysis phase for English is performed in two steps explained in sections 5.1.1 and 5.1.2.

5.1.1 Analysis of the NP or PP

We use the morphological analyzer by ENGCG available on the web (http://www.lingsoft.fi/cgi-pub/engcg/). As an example, let us show the analysis of the

following PP: "apart from these consequences of good machine translation".

WORD	LEMMA	MORPHOLOGICAL INFORMATION	SYNTACTIC INF.
apart=from	apart=from	<CompPP> PREP	ADVL
these	this	DET CENTRAL DEM PL	DN>
consequences	consequence	N NOM PL	<P
of	of	PREP	<NOM-OF
good	good	A ABS	AN>
machine	machine	N NOM SG	NN>
translation	translation	N NOM SG	<P

5.1.2 Tree-structure conversion

Based on this information we built the tree-structure (Figure 5). It will be the intermediate representation of the English phrase, and it will have the following information: lexical value, morphological and syntactic information, and the index of its mother and daughter nodes.

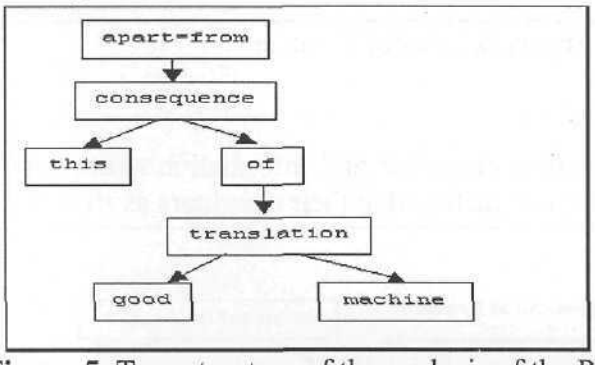

Figure 5. Tree-structure of the analysis of the PP

5.2 Transfer module

The transfer module that converts the intermediate representation of the source language into a target language intermediate representation is subdivided into two sub-modules: Lexical Transfer and Structural Transfer.

5.2.1 Lexical transfer.

First we see if there are any multiword terms in the phrase. For this task we use the multiword lexicon previously mentioned. For the moment only multiwords composed of two words have been considered. The strategy for detecting multiword terms looks for two consecutive words in the phrase that constitute a subtree of the analysis tree in this phase. When a multiword term is detected, the system gathers all its nodes in one and includes its lexical and morphosyntactic information in it. For example, "machine translation" will be translate to "itzulpen automatiko" avoiding the literal but incorrect translation *"makina itzulpena".

Secondly, the system carries out single-word lexical transfer using the bilingual lexicon. Due to the nature of Basque, we would like to focus on certain aspects of the lexical transfer process (Figure 6):

- The nodes corresponding to English prepositions have information about declension ("of" -> "[GEN]"). When they are translated into Basque postpositions, they will have a lexical value associated. Example: "apart=from" will be translated as " -- [INSTRUMENTAL] + gain".

- In other cases, the lexical, morphological, and syntactic information will be taken from the bilingual lexicon.

When there is more than one alternative in the lexical transfer phase, we can choose between selecting the first one (the most used) or asking the user to select an alternative. In the future, we plan to use a parallel corpus for this task.

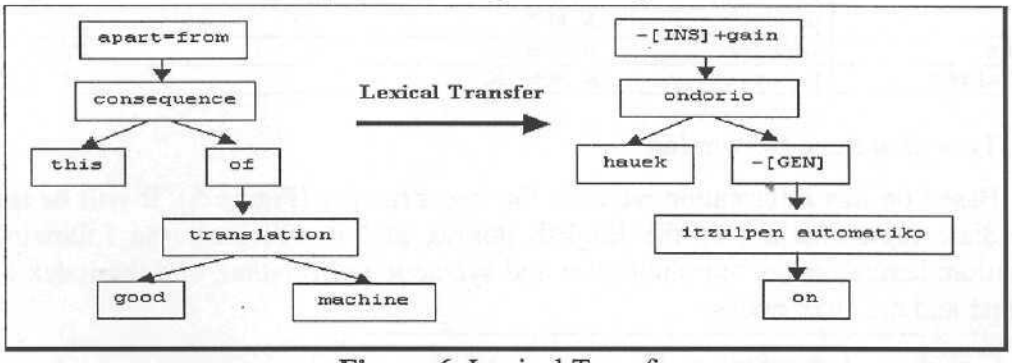

Figure 6. Lexical Transfer

5.2.2 Structural transfer

Nodes without lexical value disappear and information about declension and postposition associated to nodes is transferred to their daughters as shown in Figure 7.

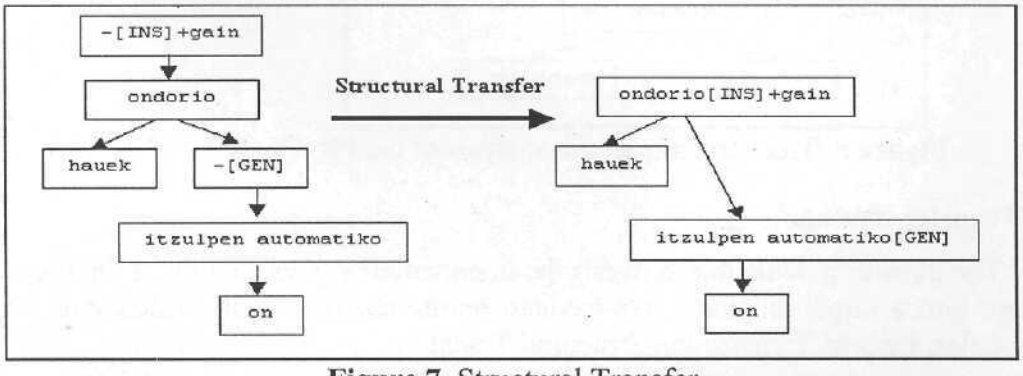

Figure 7. Structural Transfer

5.3 Generation phase

The input for the generation phase is the Basque intermediate representation. This phase works in two steps.

5.3.1 Syntactic generation

Using the Free Context Grammar the system establishes the word order in nominal and prepositional phrases. The system tries recursively grouping any node x_1 with one of its daughter x_2 using rule [x_0 -> x_1 x_2] and collapsing the tree as Figure 8 shows. The new node x_0 collects the words presents in x_1 and x_2 following the order associated to the rule. This process continues until the tree will be reduced to one node that contains all the words correctly ordered.

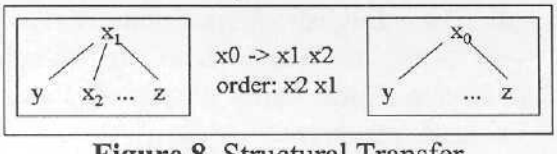

Figure 8. Structural Transfer

Finally information about declension is transferred to the last word of the phrase. Figure 9 shows these steps.

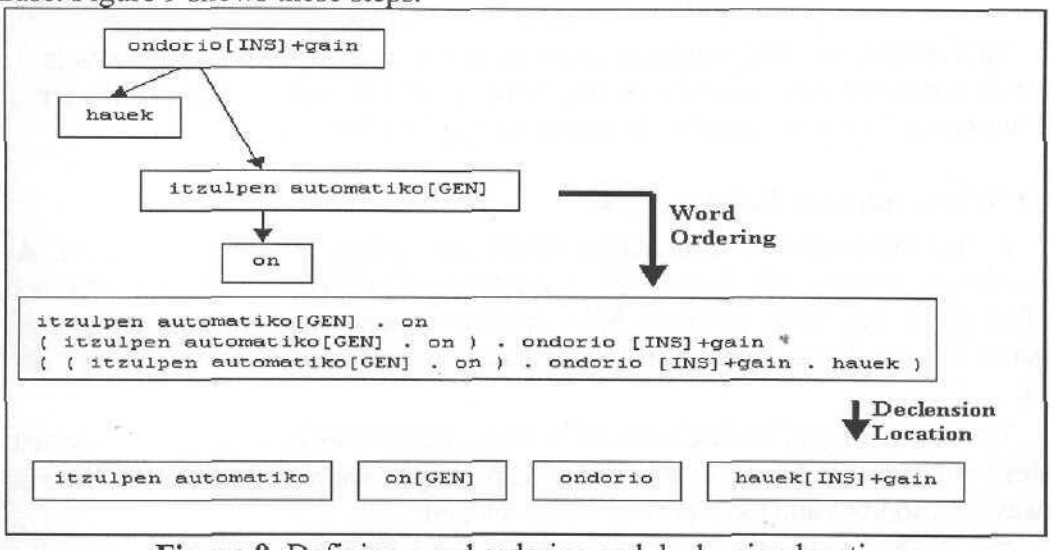

Figure 9. Defining word ordering and declension location

5.3.2 Morphological generation.

When the word order has been established, then we use a lexical transducer to generate the inflected word-form of the words that have some declension information; the word-form in other cases is just the lemma (see Figure 10). The transducer was developed using finite state technology from Xerox (Alegria et al., 97).

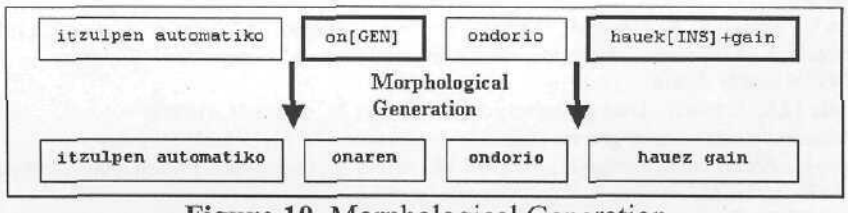

Figure 10. Morphological Generation

6 Evaluation of the prototype

In order to test the English-Basque part of our prototype we built a simple *chunker* that recognizes NPs and PPs in real texts. We used three texts each belonging to a different type: a tale by Oscar Wilde (T1, literary text), an essay by Bertrand Russell (T2, philosophical text), a text about encryption (T3, technical text).

We translated the NPs and PPs of these three texts without taking into account the context, so that whenever there was more than one alternative for the lexical transfer, the system selected the first meaning in the bilingual dictionary. Table 1 shows the precision we obtained in these experiments: 83% of the total set of phrases were correctly translated

Texts	Number of NP/PPs		
		OK	Error
T1	145	123	22
T2	109	95	14
T3	95	70	25

%	
OK	Error
85%	15%
87%	13%
74%	26%

T1+T2+T3	349	288	61	**83%**	17%

Table 1. Precision in the experiments

This table shows that results in technical texts are the poorest. The reason is that we do not use any technical dictionary. Many technical terms and acronyms are not in our general lexicon (*cracker, plaintext, encryption, http...*)

7 Conclusions and future work

A first prototype for a machine translation system has been presented. At present, the prototype works from English and Spanish to Basque, translating NP and PPs. The result has been positive: 83% precision translating NPs and PPs from English to Basque. In our opinion, it is a substantial 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 focused on the treatment of more complex NPs and PPs and enhancing the prototype for translating sentences. The treatment of multiword terms should be generalized and extensively tested. We also plan to consider the resolution of ambiguity in translation, using different strategies based on Corpus, such as example-based (Sumita et al. 1991) and statistical approaches (Brown et al. 1990).

References

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

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Maritxalar M., Urkia M. (1996). "Euskararako murriztapen-gramatika: Lehen urratsak". 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. (1999) . "MORFEUS: Euskararako analizatzaile morfosintaktikoa". UPV-EHU/ LSI/ TR 1-99

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

Aldezabal I., Gojenola K., Sarasola K. (2000) "A Bootstrapping Approach to Parser Development" International Workshop on Parsing Technologies (IWPT2000). Trento.

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

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

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

Hutchins W., Somers H. (1992). *"An Introduction to Mahcine Translation"*. Academic Press Ltd.

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

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

Voutilainen A. & Silvonen M. (2000). "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.