Building a Lexicon for an English-Basque Machine Translation System from heterogeneous wide coverage dictionaries

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

Abstract

We present a procedure for building the lexicon of a transfer-based machine translation system for NPs and PPs from English to Basque. The agglutinative nature of Basque implies the need for morphosyntactic information, so the lexicon was created by automatically extracting this necessary information from a bilingual dictionary and a wide coverage lexical database. The system translates with 83% precision, 18% better than the first approach which used the raw bilingual dictionary.

1 Introduction

In this paper we present a simple and effective procedure developed for building the bilingual lexicon of a transfer-based machine translation system from English to Basque, which is the first system that includes Basque language. This lexicon has been created almost automatically by merging information from two heterogeneous and wide coverage lexical resources.

We want to emphasize the importance of the reusability of linguistic resources for Basque, a minority language with an inevitably limited number of available resources and tools.

We employ two main resources: the EDBL lexical database for standard Basque and the electronic version of the Morris bilingual English-Basque dictionary

First we filtered the bilingual dictionary from the initial QuarkXPress format to a structured text version without any special adaptation for its use in a MT system. Experiments proved that most of the errors were due to the lack of adequacy of this bilingual lexicon. For this reason, we designed an appropriate lexicon and a procedure for automatically obtaining the required information from the bilingual dictionary and checking it according to the standard language contained in the EDBL.

The remainder of 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 resources we used, and section 5 the first version of the lexicon and the experiments performed to evaluate it. Section 6 explains the design of the second version of the MT bilingual lexicon, and finally section 7 specifies how to build it automatically reusing the resources developed for Basque and presents the final evaluation. 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. There are six Basque dialects. Since 1968 the Basque Academy of the Language has been involved in a standardization process. At present morphology is completely standardized, but the lexical standardization process is continuing.

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 are regularly applied whatever the previous elements are.
- Taking into account that prepositional functions are indicated by case suffixes inside word-forms, Basque presents a relatively high capacity to generate inflected word-forms. For instance, from one noun entry a minimum of 135 inflected forms can be generated. 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 (especially suffixes) intensively used in word-formation.

3 General architecture of the machine translation prototype

Our prototype translates NPs and PPs in real texts. The general architecture of the system (Diaz de Ilarraza *et al.* 1999) is represented in Figure 1.

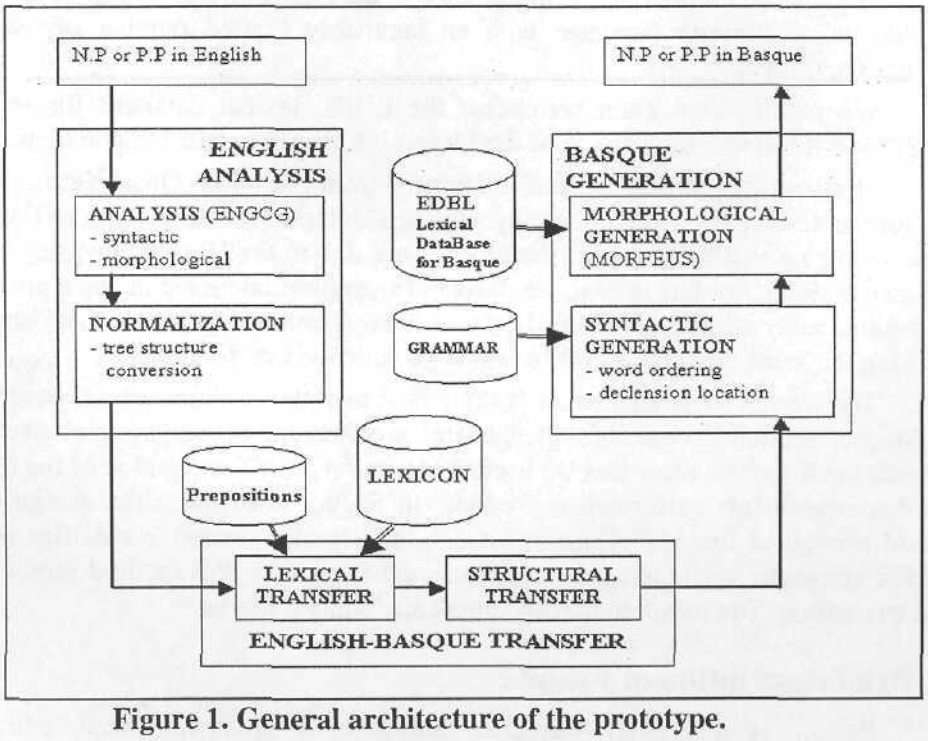

Figure 1. General architecture of the prototype.

A simple *chunker* recognizes NPs and PPs in real texts. The translation process of each phrase is performed in three phases:

- Analysis: the intermediate representation of the English source text is built using the information given by us the morphological analyzer by ENGCG (http://www.lingsoft.fi/cgi-pub/engcg/)
- Transfer: the bilingual lexicon and a special dictionary for prepositions are used to do the word-to-word lexical transfer. When there is more than one Basque equivalent, the system allows two possibilities: ask the user to select one of the alternatives or select the first because it is the most used one. The structural transfer is carried out by removing the nodes corresponding to English prepositions that do not have any lexical value; information about declension contained in these nodes is transferred to their daughters.
- Generation: once the word order is established using a context free grammar, a lexical transducer generates the inflected word forms. Nodes corresponding to determiners and adjectives need specific morphosyntactic information to decide their position with regard to the noun they modify (before or after the noun).

4 Linguistic resources

To build the lexicon we employed two main resources:

- The EDBL lexical database for Basque designed and implemented by the IXA group (http://ixa.si.ehu.es). This database (Aduriz *et al.* 1998) 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 electronic QuarqXPress version of the Morris bilingual English-Basque dictionary (Morris, 1999), where the Basque equivalents for English words are not completely standard.

5 First version of the lexicon

We converted the bilingual dictionary to HTML format, filtering it afterwards to extract the following attributes from each definition: English entry, category, sense and Basque equivalent. Due to the lack of structure and format homogeneity of the dictionary, similar problems to those described in Alshawi H. *et al.* (1989), we executed a post-process that took into account all the details to achieve a structured text version. This version has 20,042 entries, 36,121 meanings and 57,508 equivalents. As the bilingual dictionary does not provide the morphological information needed for translating prepositions, we manually built a small dictionary for the 10 most frequently used ones.

In order to make possible a fast search on the transfer phase we built an associative array in which the keys of the elements are the English entry and category, and the values of those elements contain the Basque equivalent. This structure was loaded as a hash file using PERL.

5.1 Evaluation

In order to test the prototype we used three texts each belonging to a different type: a literary text (by Oscar Wilde), a philosophical text (by Bertrand Russell) and a technical text about encryption.

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. We marked as good any possible and well-formed Basque translation of the English phrase. Table 1 shows some data about the precision of these experiments.

Number of NP/PPs	350	
OK	228	**65.1%**
Error	122	**35.9%**

Table 1. Precision in the experiments

We identified five error sources:

- **Prepositions**: The system only contained information about the 10 most frequently used.
- **Pre-nominal adjectives**: The dictionary does not give us enough information about the position of adjectives in the phrase, and the prototype places all of them after the noun. There are some exceptions that must be considered.
- **Non-lemmas**: The word supplied by the lexical transfer module is an inflected word and not a lemma; for that reason, it is not possible to use it in generation.
- **Multiword units in Basque**: When the translation of an English word is a multiword unit in Basque, generation is not possible (there are no multiword units in the lexical database).
- **Multiword units in English:** The ENGCG analyzer identifies multiword units in English texts, but multiword terms are not entries in the lexicon in use.
- **Unknown words**: Some English words are not included in the bilingual dictionary.

Table 2 shows the proportion of each kind of error in phrases that have been incorrectly translated:

Prep.	Prenominal adjectives	Non lemmas	Basque multiword	English multiwords	English unknown	Others
11%	21%	7%	5%	17 %	16%	21%

Table 2. The kinds of error and their proportion

We can see that the principal source of error (44%) is deficiencies in the bilingual lexicon. There are two main reasons:

- There is not enough morphosyntactic information in the dictionary, especially for adjectives and prepositions.
- Lack of coherence between the bilingual dictionary and the lexical database. After examining the 57,508 Basque equivalents, we saw that nearly 25% of them could not be correctly used in the generation phase because they were multiword terms or words that were not included in the lexical database.

Considering principles of efficiency and robustness, we rejected the possibility of resolving the problem during the translation process, and decided to build an appropriate lexicon, which will contain the information required by the translation system.

6 Design of a new bilingual lexicon

Our proposal for information to be contained in the bilingual lexicon for each English entry is presented in Figure 2. Five features have been distinguished:

English Entry	English POS	Basque Equivalent 1	Basque POS 1	Morphologic Segmentation 1
		Basque Equivalent 2	Basque POS 2	Morphologic Segmentation 2
		...	...	...

Figure 2. Design of the entries in the new MT bilingual lexicon.

- **English POS**: We adopt the set of categories used in the ENGCG system.
- **Basque equivalent**: This item represents the surface form of the target word. This information will be used only when morphological generation is not necessary for that word. For example, one of the meanings for "equipment" is "tresnak" which is always used in the plural, "tresna" being its lemma and "ak" the suffix related to the absolutive case and number plural. "tresnak" will be the value of *Basque equivalent* for "equipment".
- **Basque POS**: Its objective is to provide the required information for the syntactic generation phase. It contains the part of speech (POS) of the Basque equivalent of the English entry. The set of different POS is consistent with that of the EDBL lexical database. For example, as explained before, in the case of adjectives it is important to determine whether an adjective must be placed in a pre-nominal or post-nominal position. We use information associated with category and subcategory. Determiners have to be treated in the same way.
- **Morphological Segmentation**: Its objective is to provide the required information for the morphological generation phase. This information is obtained by analyzing the Basque equivalent[1] (Ezeiza *et al.*, 98) when it does not correspond with an entry in the EDBL lexical database but with an entry lemma plus one or more suffixes.

7 Construction of the new lexicon

The construction of the new lexicon we had designed was carried out in three steps:

- A procedure automatically extracted knowledge from the bilingual dictionary, adapting it to the information contained in the lexical database.
- Information associated with prepositions was manually coded.
- We built an associative array with the lexicon, loading it in a hash file as we did with the first lexicon.

Let us explain these steps and the final evaluation.

7.1 Automatic knowledge acquisition from the bilingual dictionary

The procedure implemented in PERL used the morphological analyzer for Basque (Aduriz *et al.* 2000) to analyze Basque words and created a structured text file that contains for every category of each English entries the following information: the English word form, its POS and the word form, POS and morphological segmentation of its Basque equivalents. The algorithm used can be summarized as follows:

```
foreach English_entry
  foreach English_meaning
    English_POS = Mapping_from_Morris_to_ENGCG_notation (English_Category)
    foreach Basque_equivalent
      Basque_analysis_list = Morphological_Analyzer (Basque_equivalent)
      if (not empty Basque_Analysis_list)
        if ( Basque_Analysis.lemma == Basque_equivalent and
             Basque_Analysis.POS == English_POS )
          print ( structured_text_lexicon,
                  English_entry & English_POS &
                  Basque_Analysis.( equivalent & POS & morphological_segmentation )
        elsif ( is_possible special_case )
          special_treatment
```

The most frequent situation is to find a simple term in which the Basque word-form contained in the bilingual dictionary is a lemma and the Basque word has the same category as the English entry. Nevertheless, there are some situations that need special treatment (see Figure 3):

- **Multiword terms** (about 8,700): In the cases when the Basque equivalent for an English form is a multiword unit, it is not necessary to analyze all the words contained in the multiword unit. Seeing that the morphological generation phase will only modify the last word, the analysis will be carried out on it only. For example, one equivalent of "*identification*" is "*nortasun agiri*".
- **The target-form has the same category as the entry but is not a lemma:** We store the target-form and the lemma obtained from the analysis process. In Figure 3 we show the information associated with the English form *equipment*.
- **Categories of the English word-form and the Basque equivalent are not the same** (and additionally sometimes the Basque word is not a lemma). Let us see some cases:
  - The English entry is a noun and its Basque equivalent is a nominalized verb. The morphosyntactic information item contains the noun category but information about the analysis is stored to be used, if necessary, in the generation phase (see "*adhesion / eranste*")
  - The English entry is an adjective and its Basque equivalent is:
    - a non-inflected noun: in this case the noun acts as pre-nominal adjective in compound words (see "*angular / angelu*").
    - a noun in the genitive case: the noun acts also as a pre-nominal adjective (see "*administrative / administrazioko*").
    - a verb with a modal suffix: The resulting word plays the role of an adverb (see "*desperate / etsita*") and it does not need any morphological information.
    - a nominalized verb with genitive declension: as with nouns inflected in the genitive its position will be before the noun; so the verb acts a pre-nominal adjective (see "*disappointed / galdutako*").
  - The English entry is an adverb: Morphological information is not needed because adverbs in Basque are never declined (see *"late / berandu"*).

English Entry	POS	Basque Equivalent	POS	Morphologic Segmentation
identification	N	nortasun agiri	Noun	agiri [Noun][Common]
equipment	N	tresnak	Noun	tresna [Noun][Common] [Plural]
adhesion	N	eranste	Noun	erantsi [Verb] + te [nominalization]
angular	A	angelu	Adj preN	angelu [Noun][Common]
administrative	A	administrazioko	Adj preN	administrazio [Noun] [Com] + ko[GEN]
desperate	A	etsita	Adverb	-
disappointed	A	galdutako	Adj preN	galdu [Verb] + ta [modal] + ko[GEN]
late	Adv	berandu	Adverb	-

Figure 3. Examples of special treatment in building the lexicon

Table 3 presents the results obtained in this process. 4,900 word-forms could not be analyzed and they were discarded because they could not be used in the generation phase. In the near future we will examine those entries manually in the way mentioned by Carroll *et al.* (1989). Actually, this is not an important problem because, at the end of the process, each meaning in the lexicon has an average of 1.6 equivalent-forms in Basque. From the 36,121 senses of the dictionary only 2,158 entries (under 6%) remain without any Basque equivalent.

	Basque outputs 57,508		
Directly extracted	42,972	52,608	Cases treated
Special treatment	9,636		
Cannot be analyzed	3,548	4,900	Cases discarded
Cannot be treated	1,352		

Table 3. Results obtained in the automatic process

7.2 Prepositions

The information associated with 267 English prepositions was coded manually because the bilingual dictionary does not give the information needed for declension, and it cannot be deduced from the information in the dictionary. The structure of these entries followed the general model.

7.3 Final evaluation

Table 5 presents the results obtained after running the previous test sets with the new lexicon. The results show that now 83% of the NPs and PPs have been properly translated by a *low-cost* system created by reusing previous resources. There has been an 18% improvement in precision.

Number of NP/PPs	350	
OK	289	**82.6%**
Error	61	**17.4%**

Table 4. Results obtained using the second version of the lexicon.

8 Conclusions and Future Work

We have presented the construction of a bilingual lexicon from existing lexical resources. This lexicon is used in a transfer-based machine translation prototype that

translates NPs and PPs from English to Basque. The process followed to build it has been shown to be effective and easy to implement. As a first step we adopted the raw version of a bilingual dictionary as a transfer lexicon obtaining 65% precision translating a set of 349 NPs and PPs. With the aim of improving these results, after examining the errors we concluded that 44% of them could easily be corrected. Two main tasks had to be carried out: morphosyntactic information must be considered for adjectives and determiners, and the lack of coherence between the bilingual dictionary and the lexical database used in generation had to be resolved. Once these tasks have been completed, the performance of the prototype improved: 83% of the NPs and PPs were properly translated. The whole system has been created reusing existing tools and resources and adapting them automatically.

Our immediate objectives are focused on studying the possibility of obtaining information on English multiword terms automatically to improve the lexicon.

9 Acknowledgements

We would like to thank Lingsoft and Xerox for letting us use their tools, and all the members of the IXA research team.

Notes

1 - The morphological analyzer for Basque uses the EDBL as its lexical source in the segmentation module.

References

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M. (1997)
"Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism". Proceedings of Recent Advances in NLP (RANLP97), 282-288. Tzigov Chark (Bulgary).

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A. and Insausti J. M. (1998)
"EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque". Proceedings of the First International Conference on Language Resources and Evaluation.Vol II. pp 821-826. Granada.

Aduriz I., Agiree 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 morphgological analyzer for agglutinative languages". COLING.

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.

Aldezabal I., Gojenola K., Oronoz M. (1999) "Combining Chart-Parsing and Finite State Parsing". Proceedings of the European Summer School in Logic, Language and Information (ESSLLI) Student Session 99. Utrecht, The Netherlands. 1999.
http://ixa.si.ehu.es/dokument/Artikulu/99acl-stu.ps

Alegria I., Artola X., Sarasola K. (1997)
"Improving a Robust Morphological Analyser using Lexical Transducers". Recent Advances in Natural Language Processing. Current Issues in Linguistic Theory (CILT) series. John Benjamins publisher company. Ruslan Mitkov and Nicolas Nicolov editors. Vol. 136. pp 97-110.

Alshawi H., Boguraev B., Carter D. (1989)
"Placing the dictionary on-line" Computational Lexicography for Natural Language Processing. Edited by Bran Boguraev & Ted Briscoe. pp 41-63.

Carroll J., Grover C. (1989)
"The derivation of a large computational lexicon for English from LDOCE". Computational Lexicography for Natural Language Processing. Edited by Bran Boguraev & Ted Briscoe. pp 117-133.

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 Language Processing. Cambridge University Press.
Díaz de Ilarraza A, Lersundi M., Mayor A., Sarasola K. (2000)
"Etiquetado semiautomático del rasgo semántico de animicidad para su uso en un sistema de traducción automática.". SEPLN 2000. Vigo (España).
Díaz de Ilarraza A, Mayor A., Sarasola K. (1999)
"Reusability of wide coverage linguistic resources in the construction of an English-Basque MT system". Hybrid Approaches to Machine Translation. Institute of Applied Information Sciences, Saarbruecken http://rockey.iis.sinica.edu.tw/oliver/jaiwp/
Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. (1998)
"Combining Stochastic and Rule-Based. Methods for Disambiguation in Agglutinative Languages". COLING-ACL'98, Montreal (Canada).
Hutchins W., Somers H. (1992)
"An Introduction to Machine 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.
"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, M. Morris Hiztegia (Basque-English/English-Basque). Klaudio Harlouxet Fundazioa. 1999.