Compilation and Structuring of a Spanish-Basque Parallel Corpus

A. Casillas *, A. Díaz de Illarraza *, J. Igartua *, R. Martínez †, K. Sarasola *

* Dpto. Electricidad y Electrónica & IXA Taldea
Euskal Herriko Unibertsitatea, Universidad del País Vasco (UPV-EHU)
{arantza.casillas, jipdisaa, webigigj, ksarasola}@ehu.es
† Dpto. Lenguajes y Sistemas Informticos
UNED - E.T.S.I. Informtica
raquel@lsi.uned.es

Abstract

In this paper we explain how we have compiled a Spanish-Basque parallel corpus. We also propose a corpus structure for containing: (1) translation units and the linguistic information for each unit, and (2) the whole documents with their linguistic information. The proposed corpus structure may be seen as composed of several XML documents and is based on stand off annotation model. This structure permits to work with the corpus from two points of view: as a annotated corpus with linguistic information, as well as a translation memory.

1. Introduction

There are two official languages in the Spanish side of the Basque Country: Basque (or Euskara) and Spanish. The latter is the third most spoken language of the world, and the former, Basque, is a minority language spoken in northern Spain and south-western France. There are 700,000 Basque speakers, and these comprise about 25% of the total population of the Basque Country - but they are not evenly distributed. There are six dialects, but since 1968 the Academy of the Basque Language has been involved in a standardization process. At present, morphology, which is very rich, is completely standardized, but the lexical standardization is still in progress. Most of the main public institutions such as Basque Government or universities try to publish official documents in the two official languages. In most of the cases, these type of documents are first written in Spanish and then translated manually into Basque.

A bilingual compiled corpus can be a helpful tool for different purposes: serving as training datasets for inductive programs; it can be used to learn models for machine translation, cross-lingual information retrieval; it could also be useful for automatic descriptor assignment, document classification, cross-lingual document similarity and other linguistic applications. This means that a Spanish-Basque corpus would be a very valuable resource for the research community. Once the corpus is compiled, it is possible to find different language resources inside it; for example translation memories or groups of classified documents. But nowadays the compiled Spanish-Basque corpus is poor, so there is not enough reference corpus to consult.

In this paper we explain how we have collected the bilingual corpus. We also propose a bilingual corpus structure that contains two types of informations: (1) translation units with their corresponding linguistic information, and (2) the whole documents with their linguistic information. We propose so a rich structure because our corpus resources are poor and we want them to be general and useful for different tasks in language technology research.

Similar works on compiling and representing bilingual corpus are: (Erjavec 2002), (Erjavec et. al. 2005) and (Tadie 2000). In all these three works one of the involved languages, at least, is a minority language. In (Tadie 2000) are presented procedures and formats used in building a newspaper bilingual corpus for Croatian-English. The author compares the two different ways to encode parallel corpus using XML: alignment by storing pointers in separate documents and translation memory (TMX) inspired encoding. One of the paper conclusions is to use the former due to the DTD's simplicity, because the original document keeps more unchanged, and because even with the stand-off way there is no problem to keep aligned sentences together in the same element while retaining upper levels of text encoding. In (Erjavec 2002) is used also a stand-off representation for bilingual corpus, so that linguistic information is in other separate files, that is, it is not included within the text. The authors in (Erjavec et. al. 2005) explain the compilation of massively multilingual corpora, the EU ACQUIS corpus, and the corpus annotation tool "totale". The EU ACQUIS corpus contents 8 to 82 million running words depending on the language. It contains EU law texts in all the languages of the current EU, and more, i.e. parallel texts in over twenty different languages. Unfortunately, we can not use Europarl (Koehn 2006) for Basque, the most useful corpus nowadays for research in MT.

Next section explains the characteristics of the bilingual corpus collected and the steps carried out to compile it. In Section 3 the structure of the bilingual corpus, which includes translation units and linguistic information, is explained. Finally, conclusions and future work are included.

2. Corpus Compilation

We have compiled a bilingual parallel corpus of 3 million words. This corpus is composed of two types of documents: official (about 2 million words) and not official documents (about 1 million words). The official documents are from local governments and from the University of the Basque Country. Mainly, they are edits, bulletins, letters or an-

nouncements. We have also collected some books, not official documents, that have been translated into Basque by this public university and they are about various subjects: fossils, music, education, etc.
Starting from the original plain text we are successively enriching the information contained in this corpus. The process consists of the following steps:

1. Obtaining the texts: we have downloaded the government official publications from (EHAA), in addition we have collected the available documents from the university. Actually, we continue with this collecting work and every day we download official publications from different websites. We also have got in touch with the editors of the public university to get more publications of this type.

2. Normalization of the texts into a common format: we have processed manually all the official publications because the documents were incomplete or there were some mistakes. On the contrary, there was no need of pre-processing the books. In both cases we have converted and saved all the documents into ASCII format.

3. Tokenization: involves linguistic analysis for the isolation of words.

4. Segmentation: to determine the boundaries of different types of units such as: paragraphs, sentences and entities (person, location, organization). Due to the differences between Spanish and Basque it was necessary to execute particular algorithms for each language in the detection process.

5. Alignment: the units detected in both languages were aligned. With the alignment process we have related the Spanish and Basque units of the same type that have the same meaning. Nevertheless, the alignment algorithms are independent of the language pair. The algorithms that we have executed to detect and align the different units are explained in more detail in (Martínez et al. 1998a) and (Martínez et al. 1998b).

6. Lematization and morpho-syntactic analysis: to know the lemma, number, gender and case of each word. FreeLing package (FreeLing) has been used for generating Spanish linguistic information. In the case of Basque, we have used a set of different linguistic processing tools. The parsing process starts with the outcome of the morphosyntactic analyzer MORFEUS (Aduriz et al., 2001). It deals with all the lexical units of a text, both simple words and multiword units, using the lexical database for Basque EDBL (Aldezabal et. al. 2001). This morphosyntactic analysis is an important step in our analysis process due to the agglutinative character of Basque. From the obtained results, grammatical categories and lemmas are disambiguated. The disambiguation process is carried out by means of linguistic rules (CG grammar) and stochastic rules based on Markovian models (Ezeiza et. al. 1998) with the aim of reduce the set of parsing tags for each word taking into account its context. Once morphosyntactic disambiguation has been performed, we have morphosyntactically fully disambiguated text. By the moment this is the deepest level we use to represent linguistic information in bilingual corpus, but we preview the inclusion of information about chunks, phrases and syntactic functions, in the same way we are doing for Basque monolingual corpora.

3. Bilingual Corpus Structure

The two main features that characterize the corpus structure are: (1) the richness of the linguistic information represented, and (2) the inclusion of relationships between units of the two languages which have the same meaning. The corpus structure proposed is based on the data model presented in (Artola et. al. 2005), which represents and manages monolingual corpus with linguistic annotations based on a stand off annotation and a typed feature structure. This representation may be seen as composed of several XML documents. Figure 1 shows the currently implemented document model for the bilingual corpus which includes: linguistic information, translation units (paragraphs, sentences and entities) and alignment relations. Next in this section, we will present the XML documents that constitute the proposed data model indicating their content.
With the corpus we have carried out two different processes: (1) detecting and aligning translation units, and (2) adding linguistic information to each subcorpus. With the proposed corpus structure, we have merged the output information of both processes. The final structure of the corpus is composed of the manuscript texts and of several files to define stand off annotations; these annotations contain the linguistic information and the delimitation of the units detected and aligned. The information to be exchanged among the different tools to manage this corpus is complex and diverse. Because of this complexity, we decided to use Feature Structures (FSs) to represent this information (Artola et. al. 2005). Feature structures are coded following the TEIs DTD for FSs (Sperberg-McQueen et al. 1994), and Feature Structure Definition descriptions (FSD) have been thoroughly defined for each document created. The documents created as input and output of the different tools are coded in XML. The use of XML for encoding the information flowing between programs forces us to describe each document in a formal way, with the advantages it offers to keep coherence, reliability and maintenance. This structure avoids unnecessary redundancies in the representation of linguistic features of repeated units.
The annotations which contain the linguistic information are saved into four XML documents:

- *eus.w.xml* and *cas.w.xml*: they contain single-word tokens in Basque and Spanish respectively.

- *eus.lem.xml* and *cas.lem.xml*: they keep for each single-word token of the two languages: its lemma, its syntactic function and some significant features of the morphological analysis. Words can be ambiguous and correspond to more than one lemma or syntactic function.

In order to represent the annotations that delimit translation units we have created six XML documents:

- *eus.par.xml* and *cas.par.xml*: these two documents are used to delimit the paragraphs detected in the bitext. Paragraphs are delimited with references to their first single-word token and their last single-word token.

- *eus.sen.xml* and *cas.sen.xml*: they contain the sentences of the parallel corpus by means of references to their first and last single-word token.

- *eus.nen.xml* and *cas.nen.xml*: they keep the name entities.

We have also created XML documents that relate units of the two languages with the same meaning:

- *alpar.xml*: this document is used to relate the paragraphs delimited in the files *cas.par.xml* and *eus.par.xml*. Each paragraph in one language is related with its corresponding paragraph (or paragraphs) in the other language, using the paragraph identifiers.

- *alsen.xml*: in this document are saved the relations between corresponding sentences from both languages. It is possible to set up 1-1 or N-M alignments.

- *alnen.xml*: name entities are aligned by means of this document. Relations of 1-1 and N-M are contemplated.

While translation memories take translation units as their primary "corpus", the corpus structure proposed contains the whole documents and the translations units detected and aligned. In the case of pure translation memories, only the units are saved, that is, the source text, the context from which the units come from, does not exist.

4. Conclusion and Future Work

In this paper we have explained how we have compiled a Spanish-Basque parallel corpus, the resultant language resources and its structure. The proposed structure supports linguistic information of the texts, as well as information of the alignment of the detected translation units.
The information contained in the resultant XML files is: (1) the whole document, (2) the linguistic information for each word, and (3) relations between translation units of both languages. This means that we have obtained mainly two resources: a translation memory and a morpho-syntactic tagged parallel corpus.
The main disadvantage of our proposal is that it needs more space than a translation memory or than a tagged corpus. Nevertheless, we think this representation will ensure the use of this "small" corpus in different tasks in language technology research. The compiled corpus, taking into account its structure, can be used as a translation memory for the automatic translation process or can be employed as a tagged parallel corpus for research in corpora based machine translation, machine learning, document clustering, cross-lingual information retrieval and other language applications.
Instead of repeating the same processing of the texts once and again for so different research lines, our representation makes easier and more efficient the use of parallel corpus, adding to the corpus structure to keep coherence, reliability and maintenance. Indeed, the work done so far confirms the scalability of our approach.
In the future we preview the inclusion of a new level of alignment at phrase or chunk level. We also plan to extend the graphical web interface EULIA (Artola et. al. 2004) for creating, browsing and editing also parallel corpora.

5. References

Aldezabal I., Ansa O., Arrieta B., Artola X., Ezeiza A., Hernndez G. and Lersundi M. "EDBL: a general lexical basis for the automatic processing of Basque". *IRCS Workshop on linguistic databases*, 2001.

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

Artola X., Díaz de Illarraza A., Ezeiza N.,Gojenola K., Labaka G., Salogaistoa A., Soroa A. "A framework for representing and managing linguistic annotations based on typed feature structures". *RANLP*, 2005.

Artola X., Daz de Ilarraza A., Ezeiza N., Gojenola K., Sologaistoa A., Soroa A. "EULIA: a graphical web interface for creating, browsing and editing linguistically annotated corpora". *LREC 2004, Lisbon*,2004.

"Euskal Herriko Agintaritzaren Ofiziala (EHAA)". *http://www.euskadi.net.*

Erjavec Tomaz: "Compiling and using the IJS-ELAN Parallel Corpus". *Informatica, 26* , 299-307,2002.

Erjavec T., Pouliquen C., Steinverger B., "Massive multilingual corpus compilation: Acquis Communautaire and totale". *Proceedings of the 2nd Language & Technology Conference*, 32-36, 2005.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. "Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. *Proceedings of COLING-ACL'98*, 1998.

"FreeLing 1.2 An Open Source Suite of Language Analyzers" *http://garraf.epsevg.upc.es/freeling/.*

Koehn Philipp. "Europarl: A Multilingual Corpus for Evaluation of Machine Translation" *http://people.csail.mit.edu/koehn/publications/europarl/.*

Martínez R., Abaitua J., Casillas A. "Bitext Correspondences through Rich Mark-up". *Proceedings of the* 17^{th} *International Conference on Computational Linguistics (COLING'98) and 36th Annual Meeting of the Association for Computational Linguistics (ACL'98)*, 812-818, 1997.

Martínez R., Abaitua J., Casillas A. "Aligning tagged bitext". *Proceedings of the Sixth Workshop on Very Large Corpora*, 102-109, 1998.

MarSperberg-McQueen C.M., Burnard L. "Guidelines for Electronic Text Encoding and Interchange". *TEI P3 Text Encoding Initiative*, 1994.

Tadié Marko. "Building the Croatian-English Parallel Corpus". *LREC*, 2000.

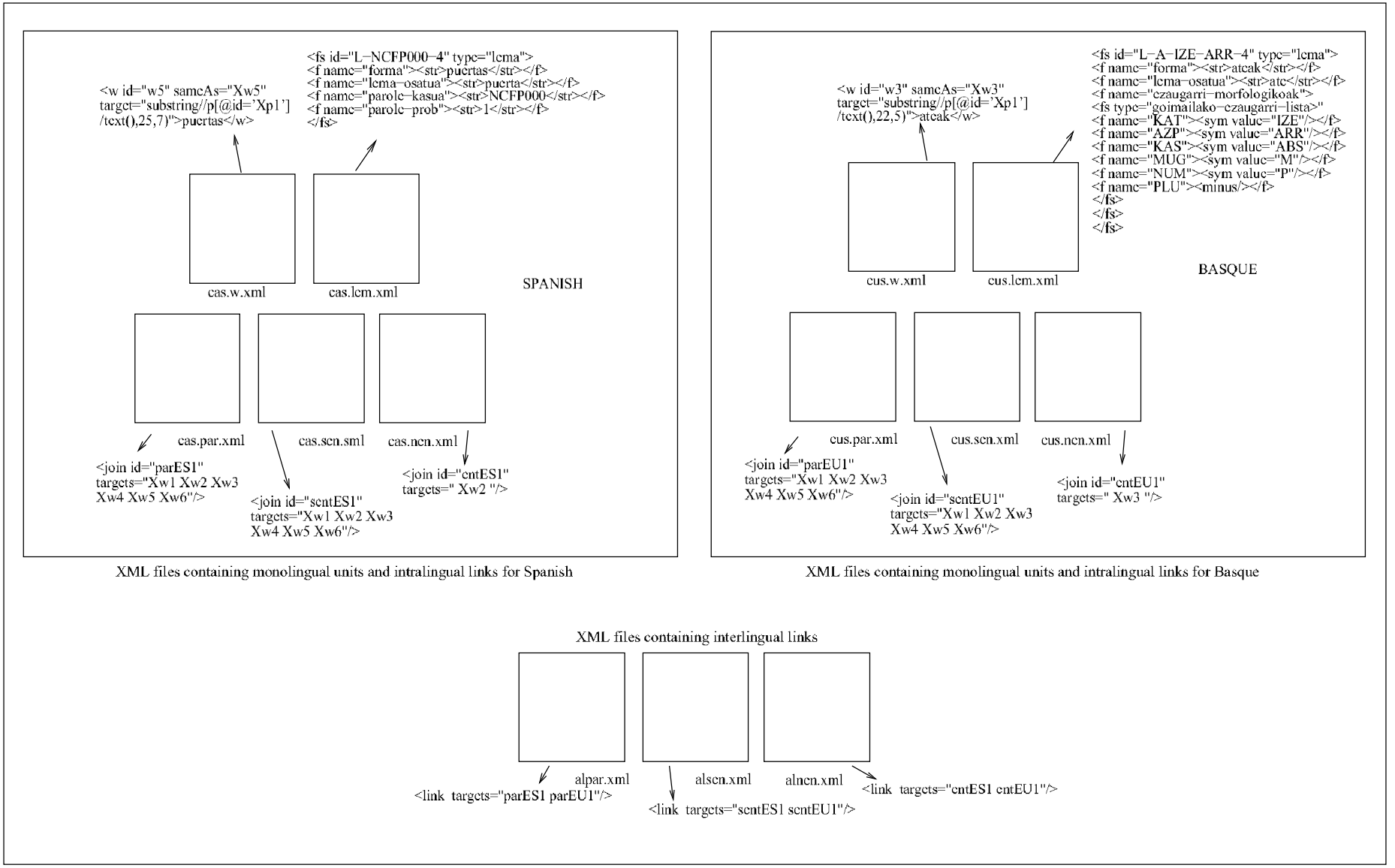

Figure 1: Corpus Structure: XML documents and their contents