Methodology and steps towards the construction of EPEC, a corpus of written Basque tagged at morphological and syntactic levels for the automatic processing

Aduriz I.*, Aranzabe M.J., Arriola J.M., Atutxa A.,
Díaz de Ilarraza A., Ezeiza N., Gojenola K., Oronoz M., Soroa A., & Urizar R.

Department of Computer Languages and Systems
Computer Science Faculty
University of the Basque Country
P.O. box 649, E-20080 Donostia
jibaregj@si.ehu.es

*Department of Linguistics
Faculty of Philology
University of Barcelona.
E-08007

EPEC, the Reference Corpus for the Processing of Basque, is a corpus of standard written Basque that has been manually tagged at different levels (morphology, surface syntax, phrases) and is currently being hand tagged at deep syntax level. It is aimed to be a "reference" corpus for the development and improvement of several NLP tools for Basque.

Although small (50,000 words), EPEC is a strategic resource for the processing of Basque and has already been used for the development and improvement of some tools. Half of this collection was obtained from the *Statistical Corpus of 20th Century Basque*, a reference corpus of Basque including 4.7 million word-forms. The other half was extracted from *Euskaldunon Egunkaria*, the only daily newspaper written entirely in standard Basque.

When defining a general framework for the automatic processing of agglutinative languages like Basque, a morphological analyser of words is an indispensable basic tool. However, previous to the completion of the morphological analyser MORFEUS, the design of the tagset and a lexical database had to be accomplished.

Choosing an appropriate tagset is a crucial task since the usefulness and ambiguity-rate of the analyser depend on it. For the morphosyntactic treatment of Basque texts, the tag system we developed is a four level system, ranging from the simplest part-of-speech tagging scheme up to the full morphosyntactic information. In addition to these four levels, further tags are added to mark verb chains, noun phrases, and postpositional phrases. Nowadays, we are involved in the syntactic tagging of the corpus, following the Dependency Structure-based Scheme in order to build a treebank.

The Lexical Database for Basque (EDBL) is a general-purpose lexical database used in several text-processing tools for Basque. This large repository of lexical knowledge is the basis in many different NLP tasks, and provides lexical information for several language tools including, obviously, the morphological analyser. At present, it consists of nearly 80,000 entries.

Morfeus is a robust morphological analyser for Basque. It is a basic tool for current and future work on NLP of Basque. It is based on the two-level formalism proposed by Koskenniemi (1983). Morfeus consists of three main modules: (i) the standard analyser, capable of analysing and generating standard word-forms, (ii) the analyser of linguistic variants (dialect uses and competence errors), and (iii) the guesser or analyser of words without lemmas in the lexicon.

The manual disambiguation of the corpus was performed on the output of Morfeus. Thus, the whole corpus was morphosyntactically analysed giving to each word-form every possible analysis, without taking into account the context in which it appeared. Once each word-form in the corpus was analysed, we carried out the manual disambiguation process. Two linguists marked independently the correct syntactic tag to each word in the corpus, applying the “double blind” method described in Voutilainen & Järvinen (1995). Both linguists' answers were compared and, when differences occurred, they agreed a single tag.

This manually disambiguated corpus was used both to improve a Constraint Grammar disambiguator and to develop a stochastic tagger.

After disambiguating the morphological tags in the corpus, the next step was to assign the corresponding syntactic tag to each word-form. Syntactic function tags follow the philosophy of the Constraint Grammar (CG). By adopting the CG formalism, we express the syntactic functions of words and the interdependencies that exist among them rather than deep structural relations. So, the syntactic tags at this level refer to shallow syntactic functions, i.e. they may provide information about the surface structure of verb chains, noun phrases, or postpositional phrases.

Once each word-form in the corpus was given at least one syntactic tag, we carried out the manual disambiguation process again. The method used was similar to the one used for the morphological disambiguation in the previous step.

At this stage we have the corpus manually tagged with surface syntactic tags following the CG syntax. No phrase units are marked yet, although based on this representation, the identification of various kinds of phrase units, such as verb chains, noun phrases, and postpositional phrases is reasonably straightforward.

In order to detect verb, noun, and postpositional phrases, we use different function tags as well as some particles (such as negative or modal particles). At present, a linguist is checking the tags that the first set of mapping rules marked up in the corpus. Whenever necessary, she adds, removes, or changes the tags automatically assigned. Once this work is finished, the first set of mapping rules developed will be tested on the corpus and the results will be used to improve the rules iteratively as well as to develop new ones.

Nowadays, we are also involved in the syntactic tagging of the corpus following the Dependency Structure-based Scheme to tag syntactically the corpus in order to build a treebank.

During the last three years, a great effort has been done in our research group (Artola *et al.*, 2002) to integrate the NLP tools for Basque described in previous sections. Due to the complexity of the information to be exchanged among the tools, Feature Structures (FSs) are used to represent it. Feature structures are coded following the TEI's DTD for FSs, and Feature Structure Definition descriptions (FSD) have been thoroughly defined. The documents used as input and output of the different tools, contain TEI-P3-conformant feature structures (FS) coded in SGML.

In the future, we also intend to extend the corpus annotation to word sense tagging and anaphora annotation.