Methodology and steps towards the construction of a Corpus of written Basque tagged in morphological, syntactic, and semantic levels for the automatic processing (IXA Corpus of Basque, ICB)

Authors

Aduriz I., Aranzabe M. J., Arriola J.M., Ezeiza N., Gojenola K., Oronoz M., Soroa A., Urizar R.

Department of Computer Languages and Systems

University of the Basque Country

Computer Science Faculty, P.O. Box. 649, E-20080 Donostia

jibaregj@si.ehu.es

The IXA Corpus of Basque (ICB) is a 50, 000 words sample collection of written standard Basque[1]. Half of the mentioned collection has been obtained from the EEBS[2] project, and the other from *Euskaldunon Egunkaria* a newspaper written in standard Basque. This corpus is being used for Natural Language Processing and although it is small, it is a strategic resource for a minority language like Basque. Each step in building the corpus has involved an extra work for the following reasons: (1) the difficulty in obtaining the corpus; (2) the smallness of the obtained corpora (3) the lack of linguistic systematisation of Basque. The first basic work was the design of the tagset. The main problem we found while defining the tagset was the absence of an exhaustive one for automatic use. Moreover, Basque printed dictionaries also lacked systematisation of categories. Besides, as the linguistic description provided by the morphological analyser is too rich and it was difficult to obtain an applicable tagset from it. Bearing all these considerations in mind, the tagset has been structured in four levels, ranging from the simplest part-of-speech tagging scheme up to the full morphosyntactic information. Complex tags are also present since they are vital for derivation, as well as for multiword terms, acronyms etc. After designing the tagset, the corpus was morphosyntactically analysed by means of EUSLEM the lemmatiser/tagger for Basque (Aduriz *et al.*, 1996). It has two main modules: (1) MORFEUS, a robust morphosyntactic analyser which is performed in two main phases: the treatment of single word units and the treatment of multiword lexical units. (2) The morphosyntactic disambiguation module, which is accomplished in two steps: first it applies a constraint grammar and, then, a HMM-based disambiguator. As Basque is an agglutinative and highly inflected language, the tagging task becomes more complex than for other languages such as English. After tagging morphosyntactically the corpus, it was manually disambiguated by two different linguists and the results were compared, applying the "double blind" method described in (Voutilainen & Järvinen, 1995a). This manually disambiguated text serves two purposes: (1) the construction of a common definition of the tagging scheme (a grammatical representation, that is, a source that can be consulted in case of disagreement); (2) as a test for evaluating the results obtained with automatic taggers. The morphological analyser gives all the possible analyses of each token in the text.

For the automatic disambiguation process we use the Constraint Grammar (CG) formalism (Karlsson *et al.* 1995) to disambiguate and analyse the corpus morphosyntactically. At this stage we have the corpus syntactically analysed following the CG syntax which stamps each word in the input sentence with a surface syntactic tag. In this syntactic representation there are not phrase units. But on the basis of this representation, the identification of various kinds of phrase units such as verb chains and noun phrases is reasonably straightforward. Nowadays, we are involved in the syntactic tagging of the corpus following the Dependency Structure-based Scheme to tag syntactically the corpus in order to build the treebank. In the future we plan to encode all the different processes that have been done over the corpus according to TEI-conformant feature structures (FS) coded in SGML. These FSs describe the linguistic information that is exchanged among the integrated analysis tools. Moreover, we are studying how to build a flexible visualization tool designed to be used by data intensive linguists working in the syntactic annotation of the corpora. Finally we are involved in the collection of more texts in order to get a more extent corpora, at this stage this task will be easier than at the beginning because the knowledge acquired. We also intend to extend the corpus annotation to word sense tagging and anaphora annotation.

In this paper we focus on the methodology currently followed for the design of the Basque corpus for NLP applications. First, the process of building the corpus has been carried out in different steps: (1) compilation of texts and their classification; (2) design of the tagset ; (3) morphosyntactic tagging of the corpus; (4) manual disambiguation of the corpus; (5) disambiguation process; (6) delimiting the chunks; (7) syntactic tagging following the dependency style (treebank); (8) formalization of the all the input text used in the previous steps by means of the SGML. Finally, we will discuss some possible improvements and future research.

---

1 The standarisation of the corpus with respect to the language is very important because we are involved so far in automatic processing of written Basque which is still under normalization

2 The EEBS project was carried out by the Language Academy in collaboration with UZEI (Centre for the Lexical Standardisation of Basque). Its aim was to record and lemmatise a five million-words corpus for the elaboration of a unified dictionary.