Language Technology is an effective tool to promote use of Basque

Aldezabal I., Alegria I., Arriola J.M., Diaz de Ilarraza A., Lersundi M., Sarasola K.
(josemaria.arriola@ehu.es)

Summary

We present an open proposal for making progress in Human Language Technology in the case of a minority language like Basque. Our main objective is to promote basic research in language engineering, orienting this investigation towards the requirements of the globalized environment of the present day. The spell-checker and the lemmatizer have proven to be particularly active tools in the ongoing standardization of Basque.

Proposal Description:

The IXA Group (ixa.si.ehu.es) was created in 1988 with the aim of promoting the modernization of Basque by means of developing basic computational resources for it. As a result of our research, four applications are currently available for common users: a spelling checker, a lemmatization based web-crawler; a lemmatization based on-line bilingual dictionary, and an open source translation system working from Spanish to Basque.

Human Language Technologies will make an essential contribution to the success of the information society, but most of the working applications are only available in English. For those working with minority languages, a great effort is needed to face this challenge.

There are 700,000 Basque speakers, and these comprise about 25% of the total population of the Basque Country although not evenly distributed. There are six dialects, but since 1968 the Academy of the Language has been involved in a standardization process. At present, morphology, which is very rich, is completely standardised, but the lexical standardization process is still going on. Our spell-checker and the lemmatizer have proven to be particularly active tools in the ongoing standardization of Basque.

From our seventeen years' experience we know that language foundations and research are essential for the creation of any tool or application; but in the same way, tools and applications will be very helpful in the research and improvement of language foundations. Therefore, these three levels (language foundations, tools, and applications) need to be developed incrementally, in a parallel and coordinated way, in order to get the best benefit from them. We propose five phases as a general strategy to follow in the processing of a language.

- **Initial phase**: Laying foundations (corpus, lexical database, morphological description, and speech data-base).
- **Second phase**: Basic tools (morphological analyzer and lemmatizer/tagger).
- **Third phase**: Tools of medium complexity (environment for tool integration, spell-checker, web crawler, surface syntax, structured versions of dictionaries, and enriched version of previous resources).
- **Fourth phase**: Advanced tools (treebank, grammar checker, lexical-semantic knowledge base, word-sense disambiguation, and language learning systems).
- **Fifth phase**: Multilinguality and general applications (information retrieval, information extraction, translation aids, and dialogue systems).