# Basque error corpora: a framework to classify and store it

Itziar Aldabe, Bertol Arrieta, Arantza Diaz de Ilarraza, Ianire Niebla, Montse Maritxalar, Maite Oronoz and Larraitz Uria[1]

**Abstract**

In the IXA research group, we compile error corpora with two different aims: automatic error treatment and intelligent computer-assisted language learning/teaching. As errors are used for these two research fields, we group the compiled texts according to their writers: Basque learners, native speakers and learners of LSP (Language for Specific Purposes).

We have already collected a 485.514 words error corpus. Corpora are stored in different formats: error examples and whole texts containing errors. In the case of Basque learners' texts, we use a specific code system indicating the source, level, student and exercise type corresponding to each file.

As far as error annotation is concerned, errors are classified according to the tags specified in the error classification developed for Basque. The information regarding to the manually detected errors is stored in two databases: *errors* and *deviations*. In the first one, computational linguists complete errors' technical information for their automatic treatment. In the second one, language teachers store learners' data as well as the possible causes of the performed errors. Therefore, based on the data collected in the *errors* database, computational linguists create rules for automatic error detection and diagnosis. And the *deviations* database stores learners' psycholinguistic information for providing feedback to learners. However, the information stored in both databases is complementary. The databases have been evaluated by means of an experiment: language teachers were asked to fill the databases with error examples and at present they contain 770 error instances.

In addition, we have just developed an Error Editor Tool which makes easier the manual annotation of errors as well as their possible corresponding corrections. Its output is a tagged corpus composed of XML documents which are used to enrich our databases with new examples.

[1] *e-mail*: larraitz.uria@ehu.es