The design of a digital resource to store the knowledge of linguistic errors

Itziar Aduriz , Izaskun Aldezabal, Maxux Aranzabe, Bertol Arrieta, Jose Mari Arriola, Aitziber Atutxa, Arantza Díaz de Ilarraza, Koldo Gojenola, Maite Oronoz (1), Kepa Sarasola, Ruben Urizar

Affiliation: IXA Group (http://ixa.si.ehu.es)
University of the Basque Country (UPV/EHU)
Postal address: Faculty of Computer Science
649 p.k., 20080 Donostia (The Basque Country)
Tel.: +34 943 01 5298
Fax: +34 943 219306
Email: jiporanm@si.ehu.es (1)

0. Introduction

In this paper we present the design of a digital resource which will be used as a repository of information of linguistic errors. As a first step in the design of this database, we made a classification of possible errors. This classification is based on information contained in Basque grammars (Alberdi et al., 2001; Zubiri, 1994) and our previous experience on knowledge representation of language students during their learning process (Díaz de Ilarraza et al. 1997). Besides, it has been carried out in collaboration with linguists of our group (http:ixa.si.ehu.es). With the purpose of validating this classification, a questionnaire was presented to experienced Basque teachers and proofreaders from newspapers or publishing houses. With their advice we completed a classification of possible errors. We designed a Zope interface (Zope is a framework for building web applications that lets you connect to external databases (Latteier et al., 2001)) so that linguists and experts in the subject will be able to introduce, through Internet, any error found in a corpus (along with its corresponding information).

The importance of this database relies on the fact that the information contained in it will be used in two different, but in some sense, complementary, projects: i) a robust Basque grammar corrector that would get added to the already existing spelling corrector (Agirre et al., 1992) developed by our research group and integrated in Microsoft tools and, ii) a Basque tutor for syntax correction that would improve the one existing now (Díaz de Ilarraza et al., 1998) which includes different facilities related to giving adapted advise about morphological questions.

1. Classifying the errors

As mentioned, in order to make a thorough classification of the mistakes users might make, we used as a basis a set of Basque grammars, our previous experience in error classification (Maritxalar 1999) and followed the advice of some linguists in our group. Besides, we contrasted our classification with other works on error typology (Becker et al., 1999).
This way, we obtained a classification in which all errors were divided into five main categories:

- Spelling errors
- Morphological, syntactical or morphosyntactic errors
- Semantic errors
- Punctuation errors
- Errors due to the lack of standardisation of Basque

Each category was subcategorised so as to make a classification as detailed as possible.

2. The questionnaire

Our error classification is focused on Basque. It must be taken into account that, as the standardisation of Basque started in the late sixties, it has not been yet completed. The Basque Language Academy (Euskaltzaindia) publishes periodically rules for the standardisation of the language but they do not cover all its aspects. So, sometimes it is difficult to decide whether a given structure may be considered standard or not.
All these characteristics made more difficult to create a proper error classification in Basque. Therefore, as we assumed that learners of Basque and language professionals do not make the same mistakes and with the same frequency, we prepared a questionnaire in which experienced Basque teachers and proofreaders were asked about two aspects. We wanted to know if all the errors we considered were actually errors and, if this was the case, which was their frequency of occurrence in the kind of texts they usually work with. In the near future, we intend to continue implementing rules (Gojenola et al., 2000) for the detection of errors starting with those ranked with the highest frequency in the questionnaire.

3. The error database

We carried out the design of the database with two objectives in mind: to be open and flexible enough (this will allow the addition of new information), and user friendly (the interface for human users has to be designed as an easy-to-use tool). We designed a simple, standard database to collect errors of the different types mentioned in point 1. The database will allow anybody with permission to do so, to update the database using network technology.
The database is composed of four entities: error, category, text and correction. In the entity named ‘error’, we store, among other things, the following technical information: whether the error is automatically detectable/rectifiable, and in such case, which is the most appropriate tool to detect/correct it. Some

psycholinguistic information is stored too: the level of language knowledge where errors occur and the level of language knowledge where errors are supposed to be corrected. This psycholinguistic information will be used to improve a computer-assisted language-learning environment developed in our group. Depending on the students' language knowledge level, the system will give advice and correct or not the error.
We also specify the origin of the error (e.g. influence of Spanish) and the possible cause of it. For each occurrence of the error in each sentence, an attribute with a value ranging from 0 to 5 indicates to which extent we are sure that it is really an error in this context. A given word or structure might be always considered an error or it might be considered an error just in some given contexts (e.g. "The bread ate John" might be correct in poetry).

4. Conclusions and future work

This work presents how experts' knowledge can be used to improve a classification of errors and how these errors could be encoded in a database. For the future, we intend to fill this database using network technologies. With this information we will continue implementing syntactical patterns for the detection of the most frequent errors. We also intend to use the information encoded in the database to construct a grammar checker. This grammar checker will be adapted to complete an intelligent computer-assisted language-learning environment for Basque (Díaz de Ilarraza et al., 1998).

Bibliography

[Agirre E., Alegria I., Arregi X., Artola X., Díaz de Ilarraza A., Maritxalar M., Sarasola K., Urkia M.]
"Xuxen: 1 Spelling Checker/Corrector for Basque based in Two-Level Morphology"
Proceedings of ANLP'92, 119-125, Povo Trento, 1992.

[Alberdi X., Sarasola I.]
"Euskal estilo libururantz. Gramatika, estiloa eta hiztegia"
Euskal Herriko Unibertsitateko argitalpen zerbitzua. Bilbo. 2001.

[Becker M., Bredenkamp A., Crysmann B., Klein J.]
"Annotation of Error Types for German News Corpus" *Proceedings of the ATALA workshop on Treebanks*, Paris. 1999.

[Díaz de Ilarraza A., Maritxalar A., Maritxalar M., Oronoz M.]
"Integration of NLP Tools in an Intelligent Computer Assisted Language Learning Environment for Basque: IDAZKIDE" *Proceedings of Natural Language Processing and Industrial Applications* Moncton, Canada. 1998.

[Díaz de Ilarraza A., Maritxalar M., Oronoz M.]
"Reusability of NLP tools for detecting rules and contexts when modelling language learners' knowledge"
Proceedings of Recent Advances in NLP (RANLP97), 342-348. Tzigov Chark (Bulgary). 1997.

[Gojenola K., Oronoz M.]
"Corpus-Based Syntactic Error Detection Using Syntactic Patterns"
NAACL-ANLP00,Student Research Workshop . Seattle. April 30, 2000

[Latteier A., Pelleitier M.]
“The Zope Book”.
New Riders. ISBN: 0735711372. July, 2001.

[Maritxalar M.]
"Mugarri: Bigarren Hizkuntzako ikasleen hizkuntza ezagutza eskuratzeko sistema anitzeko ingurunea"
Computer Science Faculty, UPV-EHU, Donostia, 1999.

[Zubiri I.]
“Gramática didáctica del euskera”
Didaktiker, S.A. Bilbo. 1994.