A database system for storing second language learner corpora

Bertol Arrieta(1), Arantza Díaz de Ilarraza, Koldo Gojenola, Montse Maritxalar, Maite Oronoz

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.: +34 943 015 061
Fax: +34 943 219 306
E-mail (1): bertol@si.ehu.es

Abstract

With the aim of storing learner corpora as well as information about the Basque language students who wrote the texts, two different but complementary databases were created: ERREUS and IRAKAZI. Linguistic and technical information (error description, error category, tools for detection/correction…) will be stored in ERREUS, while IRAKAZI will be filled in with psycholinguistic information (error diagnosis, characteristics of the writer, grammatical competence…). These two databases will be the basis for constructing i) a robust Basque grammar corrector and, ii) a computer-assisted language-learning environment for advising on the use of Basque syntax.

1. Introduction

The IXA research group has been working in Natural Language Processing during the last 14 years. At the same time, we have worked on Intelligent Computer Assistant Language Learning (ICALL) environments. The work we present in this paper has a wide background in these fields: NLP tools, error detection and ICALL environments using adapted NLP tools.

Background in NLP tools

In order to work on error detection, a very important background in NLP tools is needed. In this sense, these are the tools implemented in our group:

a. *EDBL*, a lexical database, which at the moment contains more than 80,000 entries (Aduriz *et al.*, 1998).

b. A tokeniser that identifies tokens from the input text.

c. *Morpheus*, a wide-coverage morphosyntactic analyser for Basque (Alegria *et al.*, 2002) that includes a segmentiser, a morphosyntactic analyser and a recogniser of multiword lexical units (MWLUs).

d. *EusLem*, a general-purpose tagger/lemmatiser. (Ezeiza *et al.*, 1998).

e. A shallow syntactic analyser that identifies noun phrases and verbal chains.

Background in error detection

As we have developed most of the tools in the linguistic analysis chain (morphology, morphosyntax, surface syntax, phrases, etc.), we started working on error detection. Thus, a robust spelling corrector, called Xuxen (Aduriz *et al*., 1997), was developed some years ago. With the aim of following with this work, a syntactic approach was planned. This way, some work in syntax error detection has been done in the last years, using different approaches:

a. We have combined a robust partial parser which obtains the main components of the sentence (implemented in PATR-II), and a finite-state parser used for the description of syntactic error patterns (Xerox Finite State Tool, XFST, (Karttunen *et al*., 1997)) to detect errors in dates (Gojenola K. & Oronoz M., 2000). We defined six different types of errors and its combinations.

b. The Constraint Grammar (Karlsson, 1995) formalism has been used to analyse 25 types of errors about postpositions and other 10 different types of grammar errors.

c. The relaxation of syntactic constraints (Douglas & Dale, 1992) has been used for the detection of agreement errors between the verb and the subject, object or indirect object (Gojenola, 2000). This grammar-based method allows the analysis of sentences that do not fulfil some of the constraints of the language by identifying a rule that might have been violated, determining whether its relaxation might lead to a successful parse.

Background in ICALL environments

The main work in this field done in our group is an environment for studying the learning process of language learners, called MUGARRI (Maritxalar, 1999). In this environment, we find three systems: IRAKAZI, IDAZKIDE and HITES. IRAKAZI helps the teacher in gathering psycholinguistic information about the students and the texts they write; IDAZKIDE is a student oriented ICALL environment for second language learning; and HITES is a system for modelling the interlanguage of particular learners and the common interlanguage of learners at the same language level. IRAKAZI interacts with the teacher, IDAZKIDE with the student, and HITES with the psycholinguist.

ERREUS and its connection with IRAKAZI

With this background, we realised that gathering error corpora is a very important task in order to i) have a basis for deciding which type of linguistic phenomena are important to treat, ii) have a corpora for tool-testing and evaluating. That is why we began thinking about a system that would store information about the errors of the corpora. The ERREUS database was born with this aim. ERREUS has the purpose of storing technical and linguistic information about any type of error, and it was designed for being, in some sense, a repository of error corpora. On the other hand, IRAKAZI is used to store all the information about the student (mainly, relative to his/her learning process) and the deviant structures he/she has used.

Working in the design of the ERREUS database, we realized that ERREUS is complementary with IRAKAZI. In ERREUS we are going to store any kind of error made by language learners and native speakers, and in IRAKAZI, only the deviations made by language learners. It must be pointed out that in the ICALL environment, we will speak about deviant structures instead of errors. The word "error" is directly joined to correction, and it has a negative sense. That is why we have decided to use the word deviation when speaking about the learning process, following some psycholinguistic trends (Maritxalar *et al*, 1996). So, all deviations made by students are going to be referenced in both the IRAKAZI and the ERREUS databases, while the errors found in corpora that were not written by students are only going to be stored in the ERREUS database, as we can see in Figure 1.

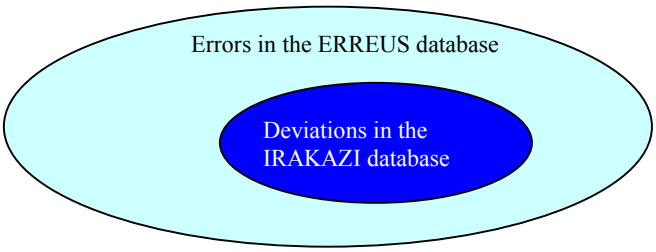

Figure 1: Errors *vs*. deviations

Due to the fact that both databases provide different points of view about the same matter, we saw the need of joining the two databases. Thus, for each error-containing-text, we would have its technical-linguistic information in ERREUS, as well as its corresponding psycholinguistic information in IRAKAZI.

Taking into account the information stored in each database (see example in figure 2), we note that the information about the text that contains the error and its category appears in both. However, there is a difference when representing the linguistic category. In the case of IRAKAZI, we store the concrete category of the deviation (AGREEMENT_SUBJ_VERB), while in ERREUS we use a hierarchical classification of linguistic errors (Morphosyntactic → Agreement → Agreement between subject and

verb). Being the case that the final category in ERREUS matches with the category in IRAKAZI, it is viable to join both databases.

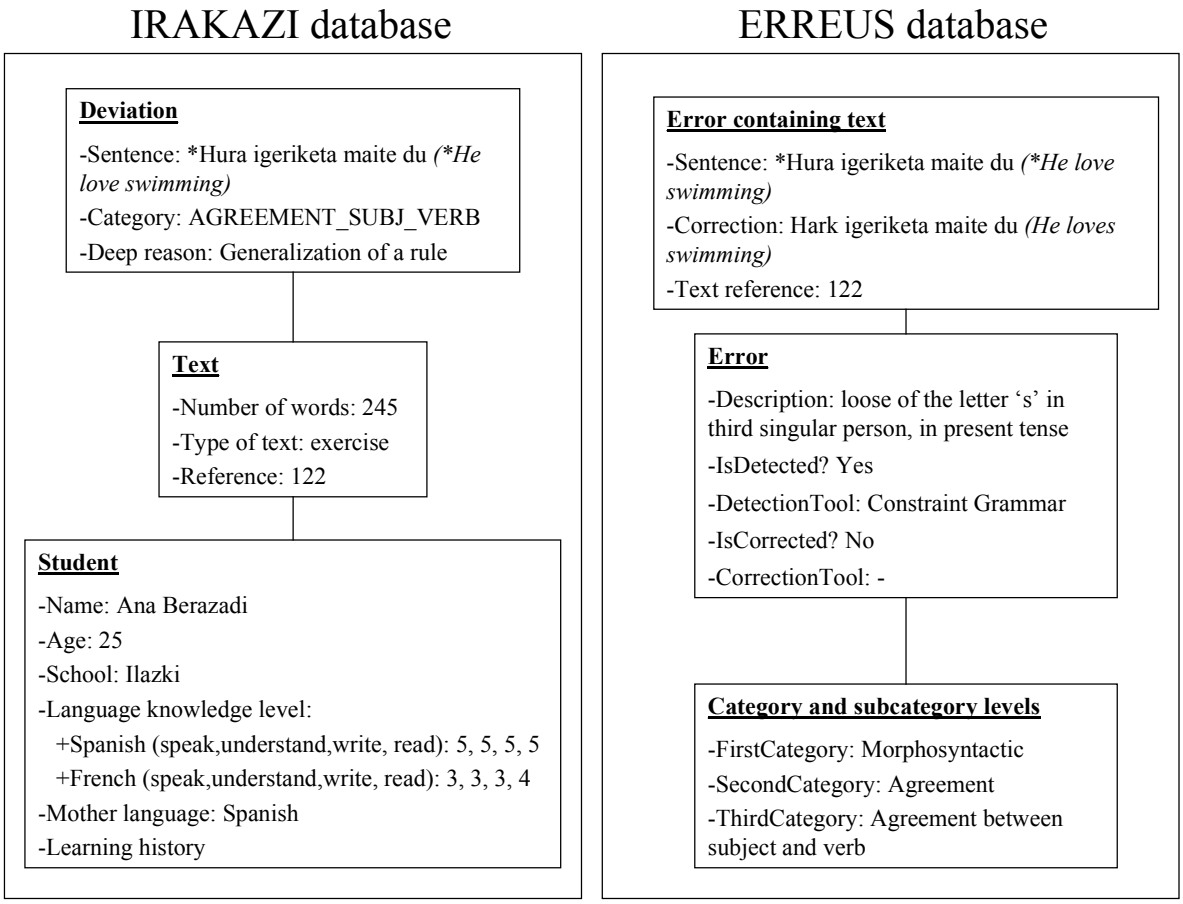

Figure 2: A view of the main information stored in ERREUS and in IRAKAZI (these boxes are not the entities of neither ERREUS nor IRAKAZI)

2. The ERREUS database

The ERREUS database will be the basis for constructing a robust Basque grammar corrector. That is the reason why we have designed a database that stores linguistic and technical information of errors found in the corpora.

Designing the database, we have followed several steps for assuring a good design and development, taking into account that a) it is important to access the database via Internet, b) many non-specialized users would access it, c) we want to store a very large range of linguistic errors, and d) we want to link ERREUS to IRAKAZI.

Next, we will briefly explain the steps we followed to build the database. Firstly, we made a complete classification of errors based on bibliographic research and hand-made studies of real corpora (step 1). Secondly, this classification was complemented with the results of a questionnaire made to some proofreaders and Basque language teachers (step 2). And, finally, this classification was used as a basis for designing and constructing the ERREUS database (step 3) and its corresponding ZOPE based interface (step 4).

Step 1: Classifying the errors

As mentioned before, in order to make a thorough classification of the errors we could find in any corpora, we used as a basis a set of Basque grammars, our previous experience in error classification (Maritxalar, 1999), and the advice of the linguists in our research group. Besides, we contrasted our classification with other works on error typology (Becker *et al.*, 1999) made in other languages.

This way, we obtained a classification in which all errors were divided into five main categories:

- Spelling errors
- Morphological, syntactic or morphosyntactic errors
- Semantic errors
- Punctuation errors and style suggestions
- Errors due to the standardisation process of Basque

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

The relevance of this error classification relies on guiding the user through the interface into the appropriate category/subcategory. This procedure will let us organise in the database all the error occurrences according to the mentioned classification.

Step 2: The questionnaire and its results

The standardisation of Basque has not been yet completed. The Basque Language Academy (http://www.euskaltzaindia.net) publishes periodically rules for the standardisation of the language, but they do not cover all its aspects. For this reason, 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, we prepared a questionnaire in order to contrast our classification. As we assumed that learners of Basque and natives do not make the same kind of errors and with the same frequency, we asked both, experienced Basque teachers and proofreaders, about two different aspects. We gave them our first draft of the error classification, and asked whether they knew any error category that was not included in such classification and whether all the errors we considered were actually errors. If this was the case, we also wanted to know, which was the frequency of occurrence of each error in the kind of texts they usually work with. Using this data, we completed our error classification. In the near future, we are going to intend to continue implementing rules for the detection of errors, starting with those ranked with the highest frequency in the questionnaire. Our objective is to use these rules to detect automatically such errors in real corpora.

Step 3: The design of the database

We carried out the design of the database with the objectives of being open and flexible enough to allow the addition of new information. We designed a simple, standard database to collect errors of the different types mentioned in the classification. The system will also allow restricted users to update the database via Internet.

The database is composed of these main entities: error, linguistic categories, 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 NLP tool to detect/correct it. We also specify the origin of the error (e.g. influence of Spanish) and the possible cause of it.

We have used four tables, each one for each level of the hierarchy in the classification. For example, in the first table (FirstLevelCategory), we have the general category of the error (orthographic, morphosyntactic, semantic, punctuation, style and errors due to the lack of standardisation of the language). Besides, each general category is divided into second level subcategories using the table (SecondLevelCategory), and so on (see figure 3).

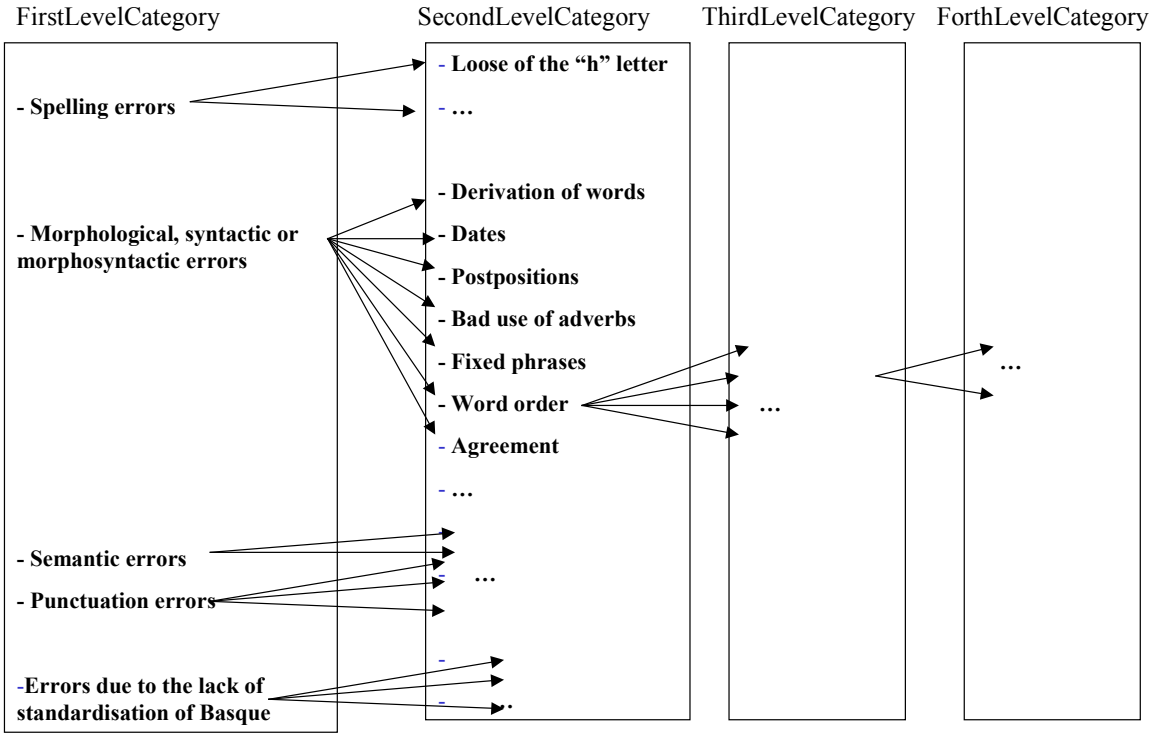

Figure 3: Classification hierarchy

The entity named 'text' stores, for each error occurrence, the sentence that contains the error. Besides, we have an attribute (with a value ranging from 0 to 5) to indicate to which extent we are sure that it is really an error in the context where it appears. 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 a literary context).

In the 'correction' entity, we store the correction of each error occurrence. In this sense, it is important to remark that if we have more than one error in a concrete sentence, we will have one different text occurrence for each error, in order to i) have the proper reference to each kind of error, and ii) have one correction for each kind of error.

Step 4: The ZOPE based interface

As the interface will be used by people that are not specialised in computers, and, therefore, it has to be an easy-to-use tool, we designed a simple and user-friendly interface based in ZOPE technology (Latteier & Pelleitier M., 2001).

This way, we built an interface to guide the user into the error classification in order to choose one concrete category/subcategory. The user has the possibility of making different operations:

a. Consulting operations as to find real examples of errors in the corpora for the chosen category. For example, if the category/subcategory "Morphosyntactic / Agreement / AgreementBetweenSubjectVerb" is chosen, the system will show all the texts with this error, e.g. "Hura igeriketa maite du" ("He love swimming"), "Bera ingelesa daki" ("She cans speak English") and so on.

b. Inserting operations as to insert an error-containing-text into the chosen category, with its own correction. For example, let us suppose that the linguists find the next sentence "That's not very *a**p**propriate*". The steps to follow should be:
   i. Find the proper category/subcategory for that error (Spelling error)
   ii. Check if the error has already been inserted (one "p" instead of the double "p")
   iii. If not, then, insert the error and its technical characteristics.
   iv. Check if the sentence has already been inserted.
   v. If not, insert the sentence and its correction.

c. Updating operations, related to error information as well as text information.

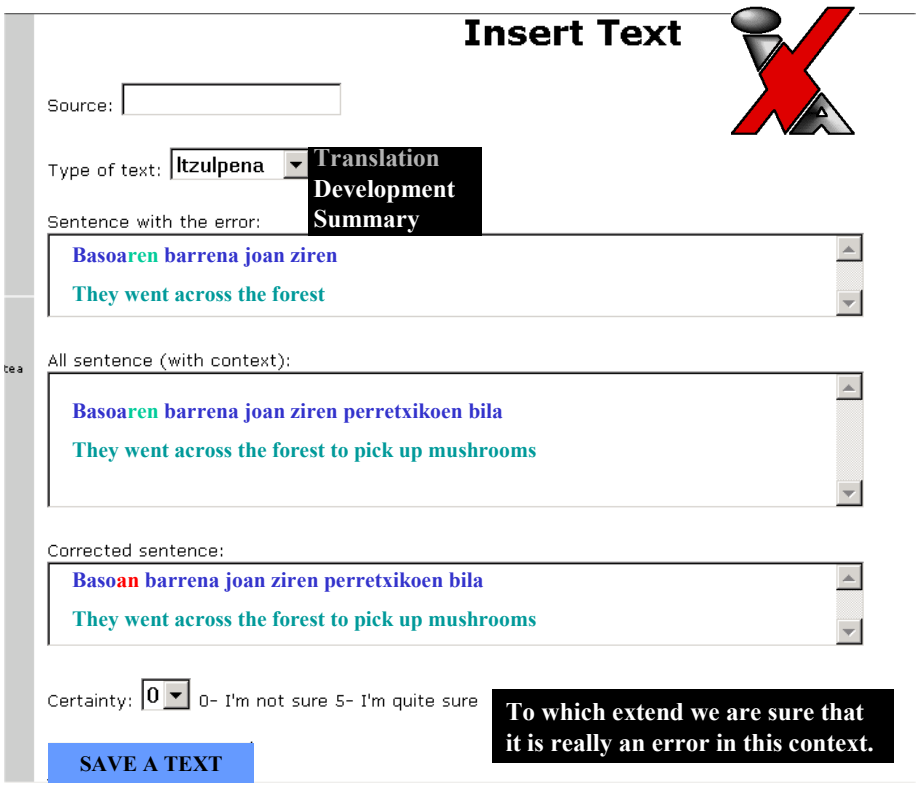

Figure 4: Inserting a text in ERREUS

3. The IRAKAZI database

In the introduction, we have done a distinction between the information stored in ERREUS (linguistic/technical) and the information stored in IRAKAZI (psycholinguistic). In the same way, as we mentioned before, we distinguish errors (ERREUS) from deviant structures (IRAKAZI). So, when speaking about IRAKAZI, we will refer to students' deviations.

IRAKAZI is responsible of storing the knowledge about the student, given by the teacher. The main goal is twofold:

- To get information about the student, relative to his/her learning process.

- To work on the diagnosis of the deviant structures of the learner.

In the future, all this information will be used in the development of the diagnosis module of IDAZKIDE (Diaz de Ilarraza *et al*, 1999), a student oriented ICALL system for second language learning.

IRAKAZI is composed of an interface to interact with the teacher and a knowledge base (field work). This knowledge base contains information about the learners, the type of exercises they do and the deviations found in texts written by them (see figures 5, 6 and 7).

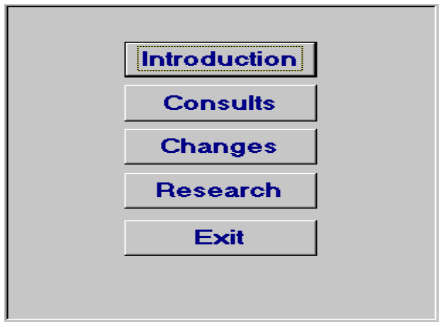

Figure 5 First screen in IRAKAZI.

The interface of IRAKAZI helps the teacher to provide the necessary information to keep in the knowledge base. Such information is composed of:

- *Specific features of the student* such as age, language level, mother tongue, other languages, studies, frequency of use of Basque, environment of use (home, business, tourism...) and so on (see figure 6).

- *Information about texts* written by the student (including a list of deviations written in the texts). Each deviation will be classified from three different points of view: i) a superficial point of view (e.g. omission of a letter), ii) a linguistic/metalinguistic point of view (e.g. agreement between subject and verb, creation of new words by means of loans…), and iii) a deep perspective (e.g. transfer from mother tongue…). The last one includes the reasons why the deviations were committed (Maritxalar & Díaz de Ilarraza, 1993).

- *Information about the exercises proposed* and their objective. For example, some texts can be the result of guessing a story or talking about something heard before, etc.

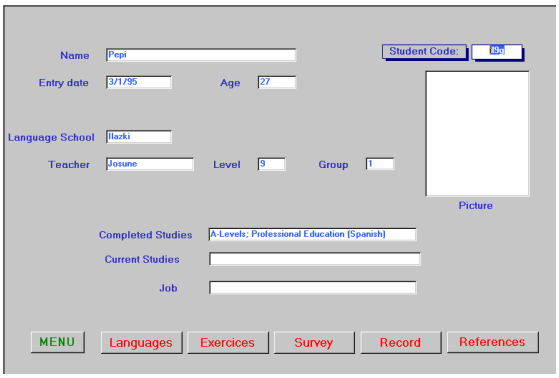

Figure 6: Specific features of the student

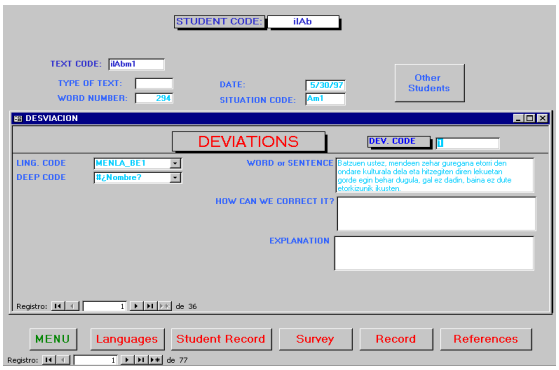

Figure 7: Texts and deviations in the texts

At this moment, the implementation of IRAKAZI is done in Access, but a new improved version implemented in Zope will be available in Internet in few months. In this new version where we are working on, we will add information about the necessary strategies that should be followed when helping the student in improving the knowledge related to his/her deviant structures. In order to do that, we will collect information about the most adequate types of exercises for the treatment of the deviant structures in each case, that is, in the case of each particular learner (see figure 7).

Some years ago, when IRAKAZI and MUGARRI were designed and implemented, a field work was done collecting Basque students texts from some schools specialised in the teaching of the language (Diaz de Ilarraza *et al.*, 1998). These text corpora are a very interesting source of data, and they will be described in the next section.

4. Text corpora

Annotated corpora of errors for Basque is a very important resource, not only for deriving an empirically based error classification, but also as a basis for the development of error detecting tools. In our case, ERREUS and IRAKAZI will be used as repositories of errors that will be annotated from linguistic/technical and psycholinguistic points of view, respectively. Text corpora provide the necessary information to both databases.

In text corpora, each kind of error occurs with very low frequency and, therefore, big corpora are needed for testing. The task of collecting corpora is not easy and it turns very difficult when error corpora have to be collected. Even if such corpora were available, the task of recognising error instances for evaluation is a hard task, as there are no syntactically annotated treebanks in Basque with error marks. So, if we want to obtain naturally occurring test data, hundreds of texts have to be automatically and manually examined and marked.

This work of collecting Basque students texts was done following some criteria: we collected written material from different language schools (IRALE[1], ILAZKI, AEK) and grouped this material depending on some features of the texts as i) the kind of exercise proposed by the teacher (e.g. abstract,

[1] IRALE, ILAZKI and AEK: schools specialised in the teaching of Basque

article about a subject, letter…) and ii) the student who wrote the text. These were students who attended classes regularly, and with different characteristics and motivations for learning Basque (e.g. different learning rates, different knowledge about other languages, mother tongue…). The corpus is made up of 350 texts written from 1990 to 1995. We codified the texts of the corpora following a prefixed notation (e.g. il10as) showing the language school (e.g. "il", **IL**AZKI), the language level (e.g. "10", Level **10**), the learner's code (e.g. "a", first letter of the name **A**inhoa), and the type of exercise proposed (e.g. "s", **s**ummary). Information related to this corpus is stored in ERREUS and IRAKAZI.

In addition to these texts, an archive of 1600 e-mail messages from the mailing list "EuskaraZ", the first workgroup in Basque, has been collected. This list was created in 1996 with the purpose of exchanging information about everything related to Basque. This corpus has the advantage of being easily accessible, electronically available and contains linguistic errors. On the other hand, it has the disadvantage of being a corpus written in an informal language, sometimes with incomplete words and abbreviations, so it is not easy to analyse it.

Apart from this text corpus, we will use grammars with error examples (Zubiri, 1994) as a source of errors and texts for filling in ERREUS.

5. Conclusions and Future Work

We have implemented two databases that gather information on linguistic errors analysed from different points of view. An interdisciplinary approach has been followed when analysing written errors in Basque texts. IRAKAZI contains the information related to the learning process of the student (diagnosis, writer characteristics, grammatical competence…), and ERREUS is provided with a vast linguistic and technical description of the errors (classification of the error, description, occurrences in texts, possible tools used for detection/correction…). Both databases have a reference to previously built and encoded learner corpora. So, we can link the two databases and create a complete system that will take into account different points of view about errors/deviations.

In the near future, we have two projects in mind:

a. A robust grammar corrector of Basque.

b. A system for syntax teaching that will improve IDAZKIDE.

The information contained in ERREUS is essential in the development of both projects, while IRAKAZI is a very important source of psycholinguistic information that will be used in IDAZKIDE.

HITES is a system for modelling the interlanguage of particular learners and the common interlanguage of learners at the same language level. Using the information obtained from HITES, IDAZKIDE will be able of giving linguistic advice to the students considering their level. For that purpose, the tools previously constructed in our NLP research group (the spelling corrector, the electronic dictionaries, the tool for shallow parsing…) could be adapted taking into account the knowledge level of the student.

In the future, we will construct mechanisms in the form of linguistic rules, grammars or statistical methods for the detection of, basically, grammar errors and deviations.

Acknowledgements

This research is supported by the University of the Basque Country (9/UPV00141.226-14601/2002) and, the Ministry of Industry of the Basque Government (XUXENG project, OD02UN52)). Thanks to Eli Pociello for her help writing the final version of the paper.

References

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A., Insausti J. M. 1998 EDBL: a Multi-Purposed Lexical Support for the Treatment of Basque. In *Proceedings of the First Int. Conf. on Language Resources and Evaluation*, vol II, 821-826. Granada (Spain).

Aduriz I., Alegria I., Artola X., Ezeiza N., Sarasola K., Urkia M. 1997 A spelling corrector for Basque based on morphology. *Literary & Linguistic Computing,* Vol. 12, No. 1. Oxford University Press. Oxford. 1997.

Alegria I., Aranzabe M., Ezeiza A., Ezeiza N., Urizar R. 2002 Robustness and customisation in an analyser/lemmatiser for Basque. In proceedings of the *LREC-2002 Customizing knowledge in NLP applications* workshop.

Becker M., Bredenkamp A., Crysmann B., Klein J. 1999 Annotation of Error Types for German News Corpus. In Proceedings *of theATALA workshop on Treebanks*, Paris.

Díaz de Ilarraza A., Maritxalar A., Maritxalar M., Oronoz M. 1999 IDAZKIDE: an intelligent CALL environment for second language acquisition. In *Proceedings of a one-day conference "Natural Language Processing in Computer-Assisted Language Learning" organised by the Centre for Computational Linguistics , UMIST, in association with EUROCALL, a special ReCALL publication,* 12-19. UK.

Díaz de Ilarraza A., Maritxalar M. Integration of natural language techniques in the ICALL systems field: the treatment of incorrect knowledge *UPV/EHU-LSI TR 9-93*.

Díaz de Ilarraza A., Maritxalar M., Oronoz M. 1998 An Implemented Interlanguage Model for Learners of Basque. *Language Teaching and Language Technology. Swets and Zeitlinger (Publisher). Sake Jager, John Nerbonne and Arthur van Essen editors.* Lisse. *pp 149-166.*

Douglas, S., Dale R. 1992. Towards Robust *PATR*. In *proceedings of COLING'92*, Nantes.

Ezeiza N., Aduriz I., Alegria I., Arriola J.M., Urizar R. 1998 Combining Stochastic and Rule-Based Methods for Disambiguation in Agglutinative Languages. In *Proc. COLING-ACL'98*, 10-14. Montreal (Canada).

Gojenola K. & Oronoz M. 2000. Corpus-Based Syntactic Error Detection Using Syntactic Patterns. In *proceedings of NAACL*-ANLP00,*Student Research Workshop* . Seattle.

Gojenola, K. 2000 *EUSKARAREN SINTAXI KONPUTAZIONALERANTZ. Oinarrizko baliabideak eta beren aplikazioa aditzen azpikategorizazio-informazioaren erauzketan eta erroreen tratamenduan*. Unpublished PhD thesis, University of the Basque Country.

Karlsson F., Voutilainen A., Heikkilä J., Anttila A. 1995 Constraint Grammar: A Language-independent System for Parsing Unrestricted Text. Mouton de Gruyter.

Karttunen L., Chanod J-P., Grefenstette G., Schiller A. 1997 Regular Expressions For Language Engineering. *Journal of Natural Language Engineering*.

Latteier A. & Pelleitier M. 2001. *The Zope Book*. New Riders.

Maritxalar M., Díaz de Ilarraza A., Alegria I., Ezeiza N. 1996 Modelización de la competencia gramatical en la interlingua basada en el análisis de corpus. *Procesamiento del Lenguaje Natural (SEPLN),* 19: 166-178.

Maritxalar, M. 1999 *Mugarri: Bigarren Hizkuntzako ikasleen hizkuntza ezagutza eskuratzeko sistema anitzeko ingurunea*. Unpublished PhD thesis, University of the Basque Country.

Zubiri I. 1994 *Gramática didáctica del euskera* Didaktiker, S.A.