IDAZKIDE: an intelligent computer-assisted language learning environment for second language acquisition

Arantza Díaz de Ilarraza, Aitor Maritxalar, Montse Maritxalar and Maite Oronoz
Dept. of Computer Languages and Systems.
University of the Basque Country. 649 Postakutxa. 20080 Donostia. Spain.

{jipdisaa, jibmagoa, jipmaanm, jiboranm}@si.ehu.es
URL: http://ixa.si.ehu.es

Abstract

This paper presents IDAZKIDE, a prototype of an intelligent language learning environment (ICALL) for learners of Basque. The philosophy of the system is to make different Natural Language Processing tools simultaneously accessible to students to help them (mainly at the morphological level) to write in Basque, as well as to give advice, taking into account some characteristics of the student gathered in a student model.

1. Introduction

The main goal of IDAZKIDE is to improve the Basque language knowledge of the students who use it. The environment provides a graphical user interface that allows the learner, among other operations, to look up a dictionary, to analyse morphologically a word-form, to check whether it is correct and to check for alternative proposals for a word. All these operations are performed by means of NLP tools. In our case, it is possible to reuse NLP tools previously developed by our group because they are independent of the language.

In the near future, when the student model which has already been developed is integrated into the interface, the system will correct texts written by learners according to the level of their knowledge of Basque. Based on empirical work previously performed (Andueza et al. 95) we concluded that it was necessary to have a standard set of user models, commonly in use, which model the process of learning Basque. Therefore, we represent the language level of users by the interlanguage level models built by means of HITES (a system for modelling interlanguage of language learners). The system interacts with a psycholinguist who gives hints to the system in the construction of the learner models. The first versions of the interlanguage (Maritxalar et al., 97: 50-59) have been built from the analysis of a corpus consisting of about 300 texts from two different schools specialised in the teaching of Basque. In the future we will integrate both systems, that is, IDAZKIDE and HITES. In this way interactions with the user will be possible and we will be able to adapt the standard model of the language level to a personal model. In this article we only explain IDAZKIDE. Information about HITES can be found in Maritxalar et al. 97.

The interest of IDAZKIDE is two-fold: firstly, Spanish speaking learners of Basque usually find several problems in learning the morphology of Basque due to the fact that it is an agglutinative language. Secondly, most learners do not have the necessary knowledge about the rules for normalisation of the language periodically published by the Basque Language Academy (Euskaltzaindia). These developments must be taken into account, as the normalisation process of Basque language started at the end of the sixties, and there are native speakers of Basque who are in an alphabetisation process, for whom IDAZKIDE is also useful.

For a better understanding of the usefulness of our tool, first we will give a brief description of Basque morphology. In the second section we will explain the architecture of IDAZKIDE. The NLP tools module and the interface will be described in sections three and four. In the fourth section we will also present a demonstration of how IDAZKIDE works. And finally, we will present some conclusions and ideas about future work.

2. A brief description of Basque morphology

Basque is an agglutinative language; that is, in which the formation of a word takes the dictionary entry and independently has appended to it each of the elements necessary for the different functions (syntactic case included).

One of the principal characteristics of Basque is its declension system with numerous cases, which differentiates it from the languages of surrounding countries. The inflections of determination, number and case appear only after the last element in the noun phrase. This last element may be the noun, but also typically an adjective or a determiner. For example:

etxe zaharreAN (etxe zaharrean: in the old house)
etxe: noun (house)
zahar: adjective (old)
r and e: epenthetical elements
A: determinate, singular
N: inessive case

So, these inflectional elements are not repeated in each individual word of a noun phrase as in the Romance languages.

Basque declension is unique; that is, there exists a single declension table for all declined entries, compared to Latin for instance, which has five declension paradigms.

As prepositional functions are realised by case suffixes inside word-forms, Basque has a powerful capability for generating inflected word-forms. For instance, from one noun entry a minimum of 135 inflected forms can be generated. Moreover, while 77 of them are simple combinations of number, determination, and case marks, not capable of further inflection, the other 58 are word-forms ending with one of the two possible genitives or with a sequence composed of a case mark and a genitive mark. If the latter is the case, then by adding again the same set of morpheme combinations (135) to each one of those 58 forms a new, complete set of forms could be recursively generated. This kind of construction reveals a noun ellipsis inside a complex noun phrase and could be theoretically extended ad infinitum; in practice, it is not usual to find more than two levels of this kind of recursion in a word-form but, in turn, some quite frequent forms contain even three or more levels. This means that a morphological analyser for Basque should be able to recognise the amount of 77 + 58 (77 + 58 (77 + 58)) = 458683 inflected forms for each noun taking into account only these two levels of recursion. For example:

semeA (the son)
semeArI (to the son)
semeArEN (of the son)
semeArEN etxeA (the house of the son)
semeArENA (the one (house) of the son)
semeArENArI (to the one (house) of the son)

This generation capability is similar for all parts of speech. In the case of adjectives, due to the possibility of graduation, this capability is four times greater.

Although grammatical gender does not exist in Basque, that is, there are no masculine and feminine cases, the verb system uses the difference sometimes, depending on the receiver and the grade of familiarity which appears as the allocutive verb forms.

Verb forms are composed of a main verb and an auxiliary finite form. The verb system in Basque is a rich one and also it is often found in single finite verb form morphemes corresponding to ergative, absolutive and dative cases.

Derivation and composition are quite productive and they are widely used in the formation of neologisms.

All the characteristics of Basque morphology have been taken into account in IDAZKIDE making it robust and rich because of its use of NLP tools previously developed by the authors' research group.

3. The architecture of IDAZKIDE

IDAZKIDE is not an intelligent language teaching environment, but an intelligent language learning environment. The main aim is to help the students during their learning process, giving them hints in order to improve their Basque language knowledge. To do that, IDAZKIDE includes information about the learner's characteristics and language knowledge gathered in the student model. Moreover, we include in the system, i.e. in the linguistic tools, information about the language knowledge of students of the same language level so that we can detect the errors according to their level..

Note that the system does not have teaching material previously stored as most CALL systems have; on the contrary, the success of IDAZKIDE lies in the independence of the systems integrated in it and the possibilities for the integration of new tools in the future. For example, the linguistic tools could be used without the interface. Separating the linguistic tools from the user interface in this way makes it possible to experiment with different user interfaces. So, IDAZKIDE is composed of the following modules: a set of wide-coverage linguistic tools, an adaptative user interface (Downton, 91), a student model, an advisor and a scheduler. The language student works with the system only through the interface.

We will explain more extensively the NLP tools module and the interface module in the next sections; here we will give a brief description of the other modules.

3.1. The student model

As a result of previous experiences in the area of Tutoring Systems, we decided to include in the description of our Student Model general information about him/her and information relating their current session (Diaz de Ilarraza et al., 98), (Maritxalar et al., 97). For the information relating to the current session we consider the types of texts written by the learner, tools used during the session, the different requests made, deviations detected, and so on. Referring to the general features of the student, and taking as a basis our conclusions about the different types of students, extracted from empirical work with language teachers, we differentiate between static and dynamic characteristics (fixed and changeable). Among the static characteristics we include the mother tongue, reasons for learning the language, knowledge about other languages. The interlanguage and the learning strategies used (the student usually pays attention to the grammar or not; the student pays attention to the meaning or not, and so on) are part of the dynamic knowledge.

The study of the interlanguage is very important in our work, because we will give help to the students depending on such knowledge. As we said in the previous section, we will use the standard interlanguage model of the language level in order to represent, as a whole, the set of correct and deviant linguistic structures generally held collectively by the students, and the contexts in which they appear. We want to point out the fact that we speak about deviant structures instead of about errors. We think that the term ‘error’ may not be an error at all the language levels, because the student could have not learnt the concept. We will only consider deviant structures depending on the knowledge level of the student.

Once a new student begins working with our system, the standard model is updated and customised according to the user. Texts and queries the users present are analysed from the perspective of their model.

3.2. The scheduler and the advisor

The system can provide three types of learning: guided learning, free exploration and combined free-guided learning, allowing the students to navigate through the system as they wish.

The scheduler decides which linguistic tool must run when using guided learning or combined free-guided learning. In the guiding process, the scheduler takes into account the students' exploration, i.e. the tools they are using, and the different characteristics gathered in the student model. When using free exploration the student decides which linguistic tool the scheduler will call.

The advisor adapts the system's answer to a student’s enquiry using the information from the student model. However, the scheduler activates the advisor.

The architecture of the system is shown in the following figure 1:

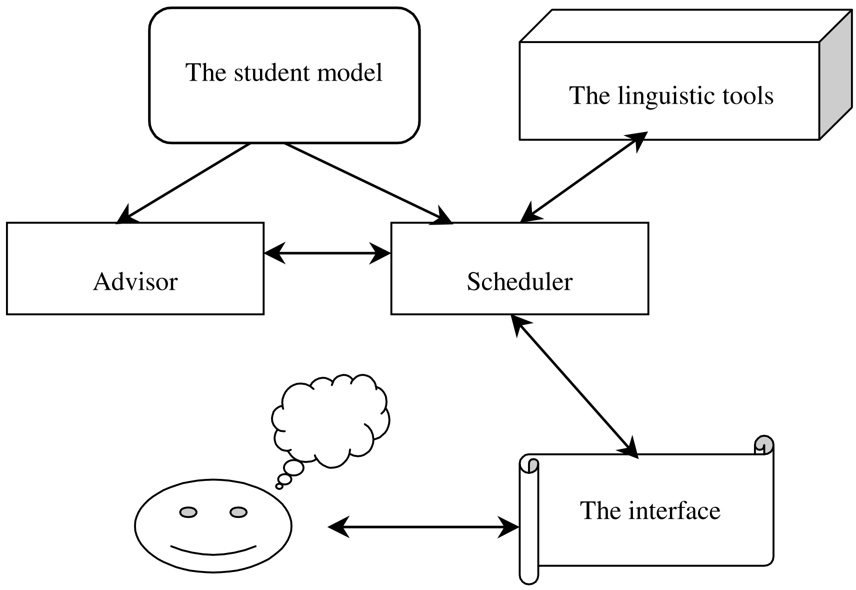

Fig. 1 Architecture of IDAZKIDE

4. *Integration of linguistic tools: NLP in IDAZKIDE*

The work we present in this paper has been developed within a general project to develop NLP tools for Basque. The linguistic tools used in IDAZKIDE were designed and implemented using the tools previously developed in our research group. Thus, the tools do not use a list of words as a basis for analysing, correcting or generating words, but we only store a lexical database of about 70,000 lemmas and morphemes, and from that, the system generates or analyses the words. Reusability in NLP is a very important issue and, in our research group it is the basis for developing different types of applications: a lemmatiser (Aduriz et al., 96:17-26), Psycholinguistic Modelling of Interlanguage (Maritxalar et al. 97), and of course, IDAZKIDE. For us it is essential to reuse and, in some cases, adapt the original tools. The reasons for adapting some of the general tools are twofold: firstly, deviant language structures are different in second language learners from those of native Basque speakers (the context of application of the rules changes and we need new rules for new phenomena), and secondly, the context of application of the rules in the case of the morphological analyser, for example, remained ambiguous and we needed this information for a correct diagnosis of the deviation. The Natural Language Processing tools that the student has access to are:

- XUXEN: a spelling-checker corrector (Aduriz et al., 97:282-288; Agirre et al., 92: 119-125) which detects deviant words in a text or, given a word-form, shows whether it is correct or not. The adapted corrector will correct students' texts depending on their knowledge level of Basque and their characteristics. For example, it will not mark those words that are out of the language level corresponding to the student. This feature will not be available until the communication with HITES is implemented.

- EDBL: a lexical database for Basque (Agirre et al. 95:299-313.; Aduriz et al. 98:821-826) with 70,000 entries that offers the possibility of consulting the category and subcategory of words together with examples of use. It gives lexical information about the word even if the word is inflected. Before processing the word as consulting the dictionary the system lemmatises the word in order to find the entry in the lexical database. We must consider that we need a lexical database for each language level; as a consequence, we propose to include new attributes in the database in order to store this kind of information. Moreover, information about the context of application of words should be included.

- A dictionary where the student can look up the definition of the chosen word. In this case, as in EDBL, the tool lemmatises the word given by the student before looking it up.

- A correct word-form proposer that, given a deviant word-form, assists the student by offering similar correct proposals, taking into account particular characteristics of the student, language level, objectives in the session, etc.

- The morphological analyser (Alegria et al., 96) which could be useful for students who want to know the grammatical rules used in Basque morphology in a straightforward manner. The morphological analyser is based on the Two Level Morphology (Agirre et al. 92). It gives the morphological analysis of the given word even if the word is unknown to the system. In the analyser the control and the knowledge are independent, and thus it is possible to adapt it for new goals, which in our case is a learning system. Principally we have changed the knowledge, and we have also made some changes in the control. For example, the initial analyser did not detect the applied rules in the analysis. The adapted analyser, however, gives for each interpretation which rules have been applied with a comprehensible code for the student. In that way the student should be able to realise which deviation they have used and correct it by themselves. If the student does not know why the word is a deviation, they can ask for proposals. In this way, the student gets feedback on their work.

- A correct word-form generator with which the students can get all possible declension-forms associated with a given lemma or, given a lemma, the desired case for its declension and the number. The system will propose the correct word-form even if the lemma is new for the system. The morphological analyser we have chosen works in both senses, that is, analysis and generation, so the generator we have developed is based on that analyser. This is one of the most interesting features of the Two-Level Morphology (Koskenniemi, 83) that has been used as the theoretical basis of the analyser. The generator is explained in detail in the next section by means of an example.

Next, we will explain what kinds of changes we have made to the original morphological analyser in order to adapt it for a language learning system. As with the original analyser, in the adapted one the implementation of the morphophonological rules has been done by using an automata. In the original analyser we had 30 automaton: 11 of them were used for analysing standard rules and their activation was never shown; the other 19 which represented deviant structures were identified, but the type of deviation remained unknown. In the adapted analyser we have 59 automata, which represent 59 different types of phenomena. We have modified the automata of the morphological analyser in order to detect which rules have been applied, and the contexts where they have been activated. We have also made some changes in the module of the analyser which recognised the automata. The number of automata has increased due to the addition of new rules for detecting new deviations of language learners and to the division of some original automata that detect in a more specific way some morphological phenomena which are very interesting for the study of language acquisition.

5. The interface in a session with IDAZKIDE

In this section we will show some of the most interesting options a student can access in a session using IDAZKIDE. Before that, we will present some ideas about the interface.

IDAZKIDE has a Motif-style graphical user interface. Since this is a standard for many X Window System applications, the time needed to learn how to interact with IDAZKIDE is minimised.

Two different platforms are offered in the interface to work with the tools: one in which the students can pose questions about a specific word and check its morphology, and another one for doing the same with the entire text.

In the student model, we strive to model the Basque language knowledge of the students without forgetting the modelling of the preferences and the objectives of the students, and their behaviour in different work sessions. What we understand as behaviour is referred to as the type of deviations the students make and the tools they use. Taking this into account, the system will adapt the advice for the students depending on their profile, hence making the interface adaptative. Thus, the students can use the system with a double purpose: to check their doubts or to correct their texts.

Next we present a typical session of a student using IDAZKIDE. When the students start the work session, they find the screen shown in figure 2.

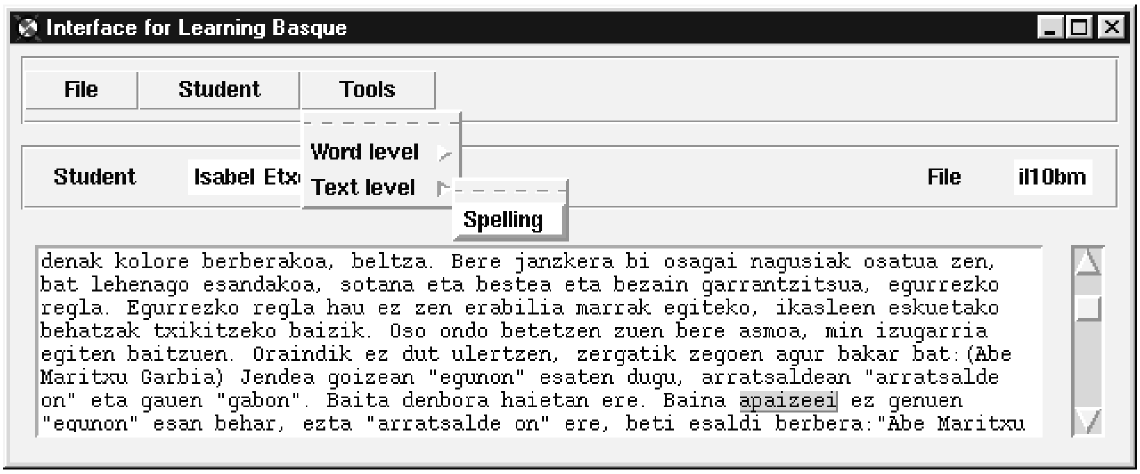

Fig. 2: The main screen of IDAZKIDE.

In this screen the students have the option of working with a previously written text or with a text that is currently being edited. They can use the options of a typical text editor such as file management (open, close, save, save as...) and editing (cut, paste...). With the chosen text, the system could give help in writing the text using the NLP tools mentioned above. The students can spell-check the whole text or select a word to work with the tools mentioned in the third section: corrector, generator, dictionary, proposer, analyser and lexical database.

The generator and the proposer are of special interest due to the features explained in subsection 1.1. In figure 3, we can see the generator.

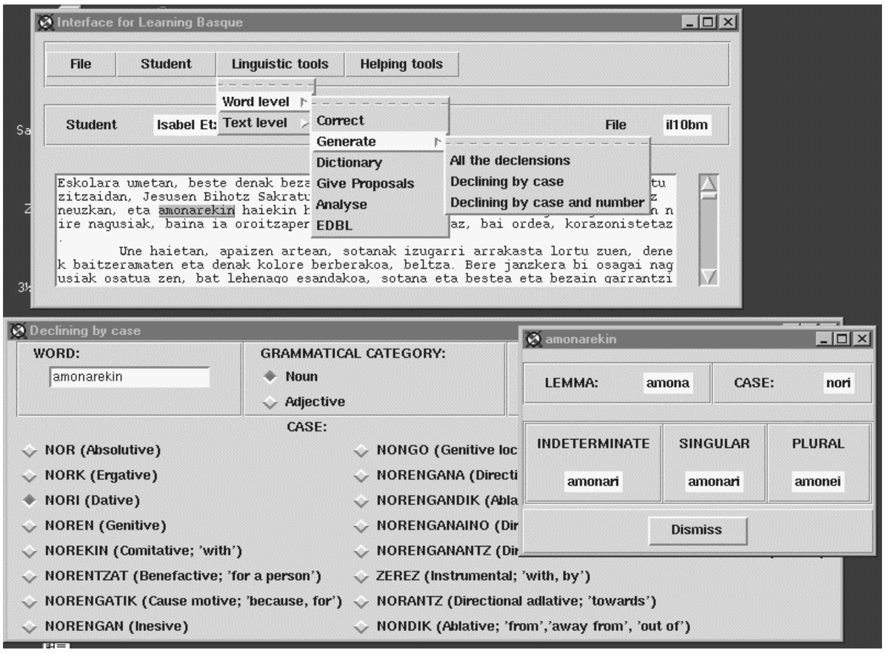

Fig. 3: Declining the word amona.

The students can choose different options when they select the generator to decline a word form. In this example, the student wants to decline the word amonarekin 'with the grandmother' in a specific case 'dative' (Declining by case in figure 3) but s/he doesn't know which is the lemma of the word. The system is able to find the lemma of the word -amona 'grandmother'- using the lemmatiser and then decline it for the requested case. The result of the action is shown in the little window in figure 3 - amonari 'to any grandmother', amonari 'to the grandmother' and amonei 'to the grandmothers'. Due to the agglutinative characteristic of Basque, we found a problem when choosing the lemma for amonarekin. In the example below, we see that there is more than one possibility for that word:

amonA + arekin ‘with the grandmother’

ama + on + arekin => *amA + on ‘of us, the mothers’*
amA+ on + arekin ‘with the one of us, the mothers’

In the first option there is only one case suffix and in the second one there are two. After some experiments with students we concluded that, with little knowledge of the language, the use of the simplest option is more satisfactory. That is, it is easier and more logical to select the longest lemma because it will have fewer suffixes. However, this heuristic must be better studied and may be revised.

There are another two options available in the generator. First, as well as in the option explained before, the students can select the declension case and they can also choose the number (indeterminate, singular and plural)(Declining by case and number). Second, given a word form, declined or not, the students get all the possible declensions for the first level of declension (see subsection 1.1)(All the declensions).

We have said that the generator is useful for those students who do not know how to decline a word. But it is possible that the students do not know the lemma, so they can use the proposer, as one of the functions is to find the correct lemma starting from a deviant one (declined or not). Let us see the proposer:

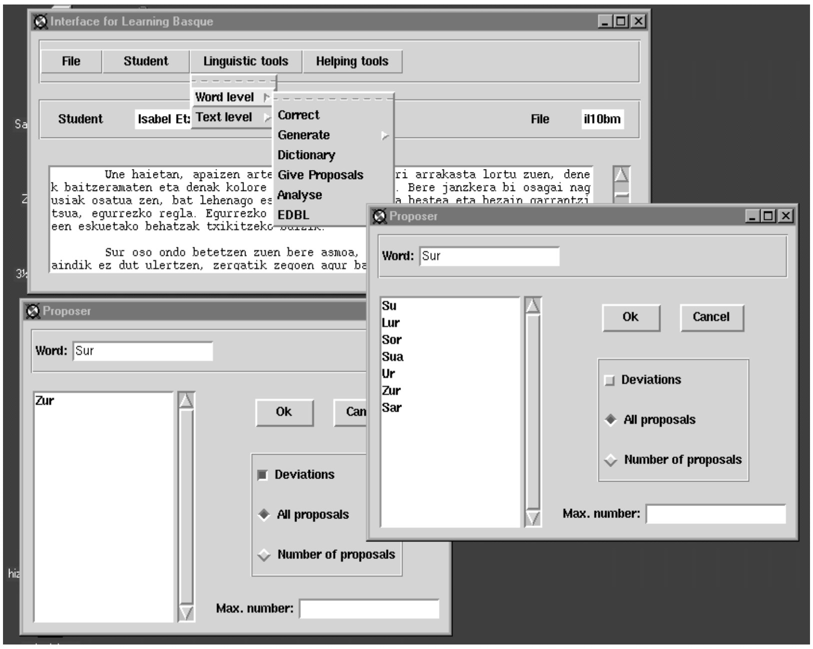

Fig. 4: Proposals for the word *sur
(the correct word is zur 'wood').

In another example, the student does not know with certainty how to write ‘wood’ (zur). They know that it is written similarly to *sur, so they can call the proposer with that word. They will get all the possible proposals for that word or they can limit the number of answers. For limiting the answers they can specify a maximum number (with this option the system will propose the more similar words) or select the option for typical errors. In the example above, when asking for all the proposals, the student has received seven words, but when asking for typical errors, they have got only one word. That occurs because the only word among the seven with a typical deviation is zur (substitution of ‘s’ for ‘z’, a very typical alternation error by in Basque learners).

6. Conclusions and future work

In the article we have explained IDAZKIDE, a prototype of an intelligent language learning environment for learners of Basque. It runs on UNIX and the interfaces have been implemented using TCL/TK while the tools are coded in C++. In the system some linguistic tools are accessible to the students in order to help them write their texts. These are adapted tools developed in the context of a wide framework of natural language processing research into Basque. The tools are independent from each other, and they have been integrated into the learning system. We have seen in our experiments that systems for learning languages must have a strong basis in the computational linguistics field. In any other case, the learning system will not be able to be easily adapted for future versions.

IDAZKIDE gathers information on all sessions where the students ask questions about the language or correct their texts. At this moment we do not use such information, but in the future that information will be analysed in order to be used by language teachers. A database for the language teacher has been developed too, but at the moment we only include in it information about characteristics of the student and language structures used in the texts. However, the information gathered about the student sessions has not yet been included in that database. In addition, students are asked about their age, mother tongue, language level, motivation and goals for learning Basque, preferences when learning a language (speaking, writing and so on), personal characteristics (meticulous, likes to probe structures learned before, and so on). These characteristics can be filled in by the students themselves or by their teachers. This information will be used in the future in the diagnosis process.

Acknowledgements

This work is being developed in the framework of a research project sponsored by the University of the Basque Country and the Government of the Province of Gipuzkoa. The authors would like to thank both institutions.

We would like to thank Jon Patrick, Professor of the University of Sydney, for his suggestions regarding this paper.

References

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.

Aduriz I., Aldezabal I., Ansa O., Artola X., Díaz de Ilarraza A. & Insausti J.M. (1998) "EDBL: a Multi-Purposed Lexical Support for Treatment of Basque". First International Conference on Language Resources and Evaluation, Granada, Spain,Vol II, 821-826.

Aduriz I., Aldezabal I., Alegria I., Artola X., Ezeiza N. & Urizar R. (1996) "EUSLEM: A lemmatiser/Tagger for Basque". EURALEX'96, Part 1, 17-26. Göteborg (Sweden).

Aduriz I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K. & Maritxalar M. (1997) "Morphosyntactic disambiguation for Basque based on the Constraint Grammar Formalism". Recent Advances in NLP (RANLP97), Tzigov Chark, Bulgaria, 282-288.

Agirre E., Arregi X., Arriola J.M., Artola X., D¡az De Ilarraza A., Insausti J.M. & Sarasola K. (1995) "Different issues in the design of a general-purpose Lexical Database for Basque". First Workshop on Applications of Natural Language to Databases, Versailles, 299-313.

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

Alegria I., Artola X., Sarasola K. & Urkia M. (1996) "Automatic morphological analysis of Basque". Literary & Linguistic Computing 11 (4), Oxford: Oxford University Press.

Díaz de Ilarraza A., Maritxalar M., Oronoz M. (1998) "An Implemented Interlanguage Model for Learners of Basque". In Jager S, Nerbonne J & van Essen A (eds), Language Teaching and Language Technology, Lisse: Swets and Zeitlinger, 149-166.

Andueza, A., Díaz de Ilarraza, A., Maritxalar, M., Martiarena, J. & Pikabea, I. (1996) Hizkuntza baten Ikaskuntza Prozesuari buruzko landa lana. Sistema Informatiko adimendun baten oinarria. Internal Report. UPV/EHU/LSI/TR 8-96. Computer Science Faculty: University of the Basque Country.

Downton A. (1991) Engineering the human-computer interface. The Essex Series in Telecommunication and Information Systems, London: McGraw-Hill.

Gojenola K. & Sarasola K. (1994) "Aplicación de la Relajación Gradual de Restricciones para la Detección y Corrección de Errores Sintácticos". Boletín de la SEPLN, grupo: 4 Análisis sintáctico y semántico, Córdoba.

Karlsson F., Voutilainen A., Heikkil J. & Anttila A. (1995) Constraint Grammar: a language-independent system for parsing unrestricted texts, Mouton de Gruyter.

Koskenniemi K. (1983) “Two level morphology” a General Computational Model for Word-Form Recognition and Production. University of Helsinki. Department of General Linguistics. Publications no. 11.

Maritxalar M., Díaz de Ilarraza A. & Oronoz M. (1997) “From Psycholinguistic Modelling of Interlanguage in Second Language Acquisition to a Computational Model”. Computational Natural Language Learning (CoNLL97). In conjunction with ACL'97/EACL'97, Madrid, Spain, 50-59.

Maritxalar M. & Díaz de Ilarraza A. (1993) Integration of Natural Language Techniques in the ICALL System Field: The treatment of incorrect knowledge. Internal Report. UPV/EHU/LSI/TR 9-93. Computer Science Faculty: University of the Basque Country.

1