An Implemented Interlanguage Model for Learners of Basque

A. Díaz de Ilarraza, M. Maritxalar & M. Oronoz

University of the Basque Country

1 Introduction

Intelligent Computer Assisted Language Learning (ICALL) systems should be used, in our opinion, with a double purpose: helping the user in his/her learning process and, helping in the research of Second Language Acquisition theories. These systems can be used in the study of special phenomena of Language Acquisition, such as fossilisation and learning strategies. Some work in this direction has already been developed in Bull et al. (1994) and Lessard et al. (1994). Our research takes into account their perspective.

An ICALL system must follow an interdisciplinary approach which includes experts in the fields of Psycholinguistics, Computational Linguistics, and Artificial Intelligence. The techniques used in the Computational areas can be useful in these types of systems.

In this paper we mainly show the way in which we have adapted the Natural Language Processing tools for Basque previously developed by our group for the detection of both, deviant and correct linguistic structures at word level; this is the basis for the internal representation of the student's language knowledge, that is the INTERLANGUAGE. We have refined these tools in order to study the learning process of Basque as a second language.

The basic tools developed by our computational linguistic group which we have based our research on are: a Spelling/Checker for Basque (XUXEN), in which a morphological analyser is included, a syntactic parser for Basque, and a Lexical

Database (EDBL) with 65,000 entries, in which general information about the morphology and syntax of Basque words is included .

The ICALL system we are working on consists of two Subsystems: the Teachers's Assistant and the Student's Assistant. The function of the Teacher's Assistant is to gather information about the users' learning process from written texts, in order to show their evolution and to help the language teacher make hypotheses about, among others, the reasons for using deviant language structures. This information will constitute part of the user model.

The Student's Assistant helps the users in their learning process giving hints according to their level and based on the recorded information we have about the students' language knowledge and learning features as native language, language study evolution, learning experience ...

The work is based on corpus analysis, and we will focus it on the morphological and morphosyntactic competence at word level. Work at word level is important in our case because Basque is an agglutinative language with rich morphosyntactic information within words. We have studied texts written by Spanish students of Basque.

In this paper we will explain the construction of the grammatical competence in the interlanguage, a module of the Student's Assistant.

2 The Grammatical Competence in the Interlanguage

The concepts transitional dialects (Corder 1971) and approximate systems (Nemser 1971) are precursors to interlanguage (Selinker 1972, 1992). Their aim was to define communicative and grammatical competence in second languages. All of them have these common characteristics: a) a student's discourse is **independent from the native language (L1) and the target language (L2)** and it is the product of a **structured linguistic system**; b) the linguistic system is **variable** during the learning process and it is **very similar in students of the same language level**, with the exception of some differences, results of a person's learning experience. The above mentioned characteristics will be the basis for modelling the interlanguage. For example, as we have said that the linguistic system is very similar in students of the same level, we can infer that we will have an interlanguage model for each level.

Before seeing the different perspectives to explain a linguistic phenomenon, we consider it important to specify the concept of DEVIATION as an alternative to the idea of error. We will say that a linguistic structure is a deviation (Maritxalar et al. 1996) when one of the following three conditions is fulfilled: A) the linguistic structure is incorrect – we usually call that error –; B) the structure is overused, replacing more suitable structures the student wants to avoid – e.g., to use the conjunction *eta* 'and' instead of *ordea* 'however'–; C) a specific structure is avoided (e.g., *nor da?* 'who is?' instead of *nor ote da?* 'who could be?'). Deviations are part of the interlanguage and they are represented in the same way as standard structures. For example a deviation at 10th level like *nor da?* (C type) is not considered deviation in lower levels of the language.

2.1 Linguistic and superficial perspectives for the same phenomenon

After determining the linguistic phenomena in the interlanguage, we will explain now different perspectives from which to describe the same linguistic phenomenon: the superficial perspective and the linguistic one. We can take as example these A type deviations:

a) **oihanan (=> oihanean)* b) **dakilako (=> dakielako)*
(= in the forest) *(=because he/she knows it)*

In the case of **oihanan* as well as in **dakilako* from a purely SUPERFICIAL perspective we could codify the structure as *LEDIE (Delete the E Letter Inside the word). However, from a LINGUISTIC perspective we would say that in the (a) case the student has deleted the epenthetic *-e* used for linking morphemes, whereas in (b) the causal morpheme *-lako* has been used instead of *-elako*.

Our tools detect both perspectives by means of different interpretations produced by the adapted morphological analyser (Learner_analyser) and the lexical database for language learning. At this time we have no explicit information in order to distinguish the perspectives, but future work is being developed in that direction. At the moment, the adapted analyser includes 59 morphophonological general rules (18 for standard and 41 for deviant phenomena). It must be noted that we don't include among the 59 general rules for particular phenomena like: replacing the causal morpheme -lako instead of -elako. The high number of general rules is due to the fact that Basque is an agglutinative language.

As we have said previously, in the future both perspectives, superficial and linguistic, will be represented for all linguistic phenomena whenever they can be identified. We must take into account that for some phenomena only one perspective will exist and for others we will find more than one possible interpretation in the same perspective. We can see an (A) type deviation in figure 1.

Example: **institutoako (=> institutoko)* 'of the school' (locative)

Two superficial interpretations:	Two linguistic interpretations:
1) *LEAEA	1) Lemma *institutoa* +
(Add the A Letter at the End of the lemma of the word)	morpheme *-ko*
2) *MOA-A	2) Lemma *instituto* +
(Add the A MOrpheme)	morpheme *-a* +
	morpheme *-ko*

Fig.1.

Considering the variability of the linguistic system in the interlanguage, experienced language teachers have diagnosed that students could have in their knowledge base both lemmas *institutoa* and *instituto* at the same time.

2.2 Variable Knowledge versus Fixed Knowledge

As we said before, the grammatical competence in students of L2 is variable. Based on the study performed we detected that there are some structures that become stable during the learning process. Such structures make up the FIXED KNOWLEDGE whereas changing structures make up the VARIABLE KNOWLEDGE. Variable knowledge can either develop into new variable structures or become part of the fixed knowledge. When we analysed corpora, we found four types of structures depending on the use made of them: attempts, systematic structures, changeable structures, and slips (Maritxalar et al. 1996). Systematic structures are part of the fixed knowledge and changeable ones belong to variable knowledge. However, attempts, incomprehensible outputs of the student, and slips, momentary deviations caused by lack of concentration, are not considered part of the grammatical competence of the student. We propose a STABILISATION value in interlanguage structures in order to determine when a structure belongs to fixed knowledge and when it belongs to variable knowledge.

On the other hand, the variability of interlanguage changes, depending on the context in which the linguistic structure is activated (Selinker et al. 1992). So, the specification of a linguistic structure is characterised by: linguistic features, stabilisation value, and context. We distinguish three types of contexts: linguistic (e.g. subordinates in a sentence), textual (e.g. sentence or length of the word; type of text: composition, letter, and so on), and thematic (e.g. subject of the text, exercise: story or scientific-technical creation, summary …). The linguistic structures we activate when writing a story are different to those we activate when we write scientific-technical texts. This way of putting linguistic structures in context could be applied similarly when modelling first languages.

3 Using Corpus Linguistics in order to Model Interlanguage

In this section we will explain the application of a top-down methodology to model interlanguage at different language levels of mastery. In our case, interlanguage modelling is based on the study of the results of corpus analysis carried out in Maritxalar et al. (1993), Andueza et al. (1996), and Maritxalar et al. (1996). The applied methodology is based on the following criteria: a) the linguistic system for language learners at the same level is similar for all of them (Corder 1971, Nemser 1971, Selinker 1972,1992); b) the fixed knowledge in the interlanguage of one level includes the whole fixed knowledge at lower levels (Maritxalar et al. 1994). The applied methodology starts at high language levels, then we study the interlanguage of lower levels based on the results for higher levels. In this way, we examine the evolution of the interlanguage through out different levels. The reasons for a top-down methodology are that most computational tools for Basque we have (lemmatiser, spelling checker-corrector, morphological disambiguator …) can be easily adapted for high language levels; besides, usually computational tools for analysing written texts of high language levels are more robust than those of low levels and, finally, there usually exists more written material at high levels than at low ones.

Before explaining modelling based on corpus analysis and showing some examples, we would like to make some comments about the criteria for defining the corpus: we collected written material from different language schools (IRALE[1], ILAZKI) and grouped this material depending on some features of the texts as, 1) the kind of exercise proposed by the teacher (e.g. abstract, article about a subject, letter ...) and 2) the student who wrote the text. Those are students with a regular attendance in classes and with different characteristics and motivations for learning Basque (e.g. different learning rates, different knowledge about other languages, mother tongue ...).

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. 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). At the same time, a database for gathering the relevant information about the learning process of the students was developed. We got such information from interviews with the students and with teachers (Andueza 1996). The corpus is made up of 350 texts written by students of IRALE and ILAZKI from 1990 to 1995. This corpus has been divided in subsets depending on the study level. At the moment we have defined three language levels of study that we call low, intermediate, and high levels. We have automatically analysed subsets of corpora in intermediate and high levels. Choosing a text as a unit of study, groups of sixteen texts have been studied deeply and automatically, in the way we explain below. A language teacher has evaluated the results: the type of rules detected and the contexts where they are applied. The evaluation has been successful, even though in some cases the perception of the teacher was not the same as the results inferred from the automatic study of the corpora (e.g., in the opinion of the teacher the students are used to deleting the h letter more usually than adding it). It must be taken into account that the design and implementation of the automatic tools were carried out based on three different previous studies of corpora during 90/91 (i.e. 50 texts semiautomatically analysed), 93/94 (i.e. 20 texts semiautomatically analysed), and 94/95 (i.e. 100 texts semiautomatically analysed). These studies were done, in the first case, by teachers who didn't know the students, and in the other two cases by teachers who knew the students. In the first two cases the work lasted two months. In the third case, however, texts were collected every week from September until June, and two teachers worked five hours per week in studying the corpora during the 94/95 academic course. The language learners had five hours of language classes per week, and they wrote one composition every week or every fortnight.

3.1 Modelling interlanguage based on corpus analysis

The tools we are adapting have been previously developed in our computational linguistic group during the last ten years. These tools are: the Lexical Database for Basque (Agirre et al. 1994), the morphological analyser (Agirre et al. 1992), the lemmatiser (Aldezabal et al. 1994) and some parts of the Constraint Grammar system (Alegria et al. 1996, Karlsson et al. 1995).

---

1. IRALE and ILAZKI: schools specialised in the teaching of Basque

The steps we followed using these adapted tools in order to build the interlanguage model for the highest language level (H level) were:

1) Design and implementation of the lexical database for the H language level.

2) Selection of the corpus (CORPUS-H) and subsets of CORPUS-H to be used in the next steps. This selection will be based on the criteria explained earlier.

3) Definition of the morphology and morphosyntax based on a subset of CORPUS-H.

4) Identification of the fixed knowledge and the variable knowledge, considering the contexts defined in section 2.

5) a) Evaluation of the reliability of the model using other subsets of CORPUS-H.

b) Evaluation of the results by a language teacher of H level.

The grammatical competence of the interlanguage consists of general and specific knowledge dependent on the context. However, we have not explicitly represented such differences in figure 2. In this figure we have symbolised the development of the fixed and variable knowledge through consecutive levels. We can obtain the model of X from the model of X+1 and the study of the CORPUS at level X (CORPUS-X).

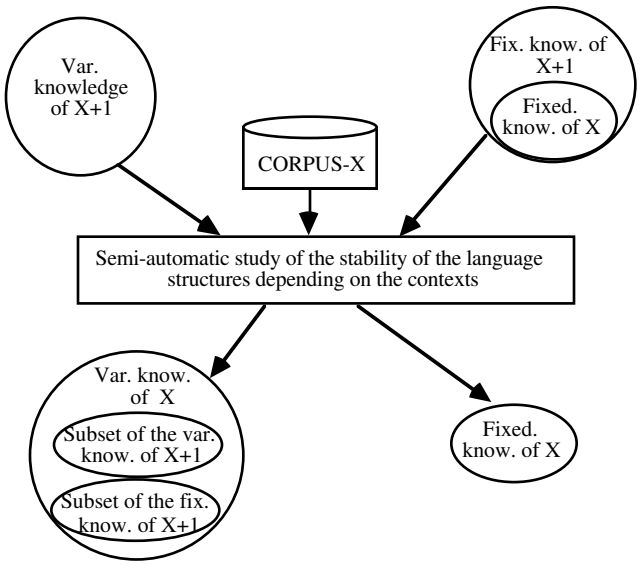

Fig2. Construction of the interlanguage model of X from X+1

This process will be repeated to represent the interlanguage models of lower language levels.

In the next three subsections we will explain, in more detail, the different steps for the construction of the interlanguage model and the tools we have developed in order to study the collected corpora.

3.2 Design of the lexical database for each language level

In the design of the lexical database for the study of interlanguage, we propose to add new fields associated to the entries of EDBL (Lexical Database for Basque): stabilisation value, linguistic context, textual context, thematic context, and a list of levels. For building the database, we intend to use the lemmatiser (Aldezabal et al. 1994) along with a context detection environment. In the database for L2 we found morphemes with more than one entry; for example, the morpheme -a which only exists as absolutive in EDBL will have a second entry for the ergative in our case because it is a typical deviation in students of Basque.

At this time we are using EDBL with 65,000 entries that our group implemented for studying language phenomena in an environment for Basque as first language.

3.3 Definition of the morphology and morphosyntax

In this section we will explain some of the tools we have developed in order to create interlanguage models by means of corpus analysis. We used C language and Perl for implementing the tools. At this time we concentrate our work on the definition of the morphology and morphosyntax (word level) because, as we said above, in Basque a big part of the syntax is included within the word (see the example in section 2.1. (*dakielako* 'because he/she knows'). In future work we will study the modelling of some aspects of the syntax at sentence level, as for example concordance. Next, we will describe the tools that appear in figure 3.

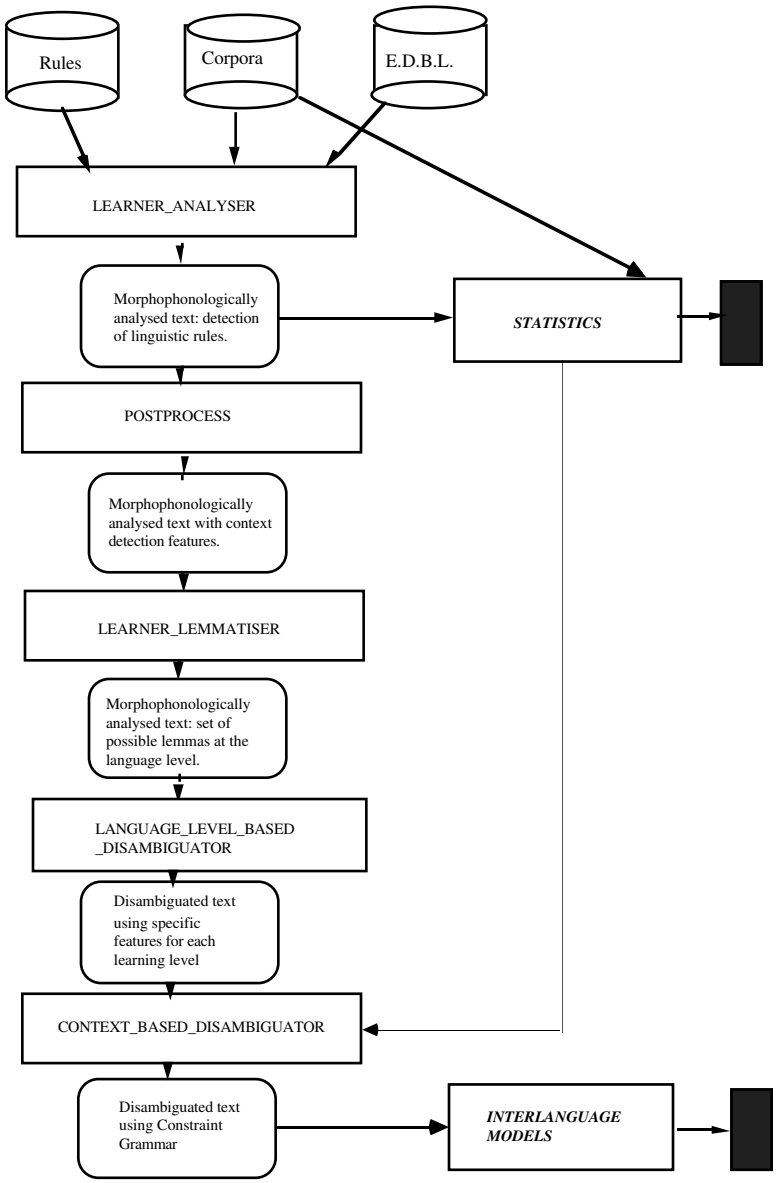

Fig. 3. Process for the definition of Interlanguage Models

Let us see a description and an example of each tool:

LEARNER_ANALYSER: the adapted morphological analyser detects the morphophonological rules (for standard and deviant phenomena) applied in each interpretation of the word and the total number of rules applied in the interpretation. The analyser uses two-level morphology (Agirre et al. 1992) and gives us morphological and morphosyntactic information.

When we analyse the word *arriskurik* 'any danger' we find 5 interpretations of the word. As an example we will only represent the most interesting interpretations for our explanation.

```
/<arriskurik>//
((form "arriskurik")
%T:rule2:1
 ((anal $ 1)
      ((lemma "arrisku")((ENTRY arrisku)(P-O-S NOUN)(S-P-O-S COMM)))
      ((morph  $ "Rik")((ENTRY ik)(P-O-S DEC)(CASE PAR)
                (DETER INDEF)(SF1 @OBJ)(SF2 @SUBJ))(2)))
%T:rule22,:rule29,:rule36,:rule2:4
 ((anal $ 2)
      ((lemma $ "harri")((ENTRY harri)(P-O-S NOUN)(S-P-O-S COMM))(22))
      ((morph $ "Ez")((ENTRY z)(P-O-S DEC)(CASE INS)(DETER INDEF)
                (SF1 @ADV_COMPL))(29))
      ((morph $ "ko")((ENTRY ko)(P-O-S DEC)(CASE LGE)
                (SF1 @ADJ_COMPL>)(SF2 @<ADJ_COMPL)
                (SF3 @ADV_COMPL))(36))
      ((morph $ "Rik")((ENTRY ik)(P-O-S DEC)(CASE PAR)
                (DETER INDEF)(SF1 @OBJ)(SF2 @SUBJ))(2))) ... )
```

Fig.4. Morphophonologically analysed text: detection of linguistic rules (see fig. 3). See linguistic codes in Appendix 1.

POSTPROCESS: this postprocess identifies the rules and shows the place (lemma/morpheme) in which each rule is applied in the morphological interpretation. This tool also identifies other features such as word length and type of last letter (vowel/consonant) of the root. Our experience during the interviews performed with teachers show us the importance of these aspects.

```
/<arriskurik>//
((form "arriskurik")
%T:rule2:1
 ((anal $ 1)
      ((lemma "arrisku")((ENTRY arrisku)(P-O-S NOUN)(S-P-O-S COMM)
            (LENGTH LENGTH10)(END-ROOT VOWEL)))
      ((morph $ "Rik")((ENTRY ik)(P-O-S DEC)(CASE PAR)
            (DETER INDEF)(SF1 @OBJ)(SF2 @SUBJ))(2)(PLACE MOR_LEAAR1)
%T:rule22,:rule29,:rule36,:rule2:4
 ((anal $ 2)
      ((lemma $ "harri")((ENTRY harri)(P-O-S NOUN)(S-P-O-S COMM))
            (22)(PLACE LEM_LEDBH)(LENGTH LENGTH10)
            (END-ROOT VOWEL)
      ((morph $ "Ez")((ENTRY z)(P-O-S DEC)(CASE INS)(DETER INDEF)
            (SF1 @ADV_COMPL))(29)(PLACE MOR_LERAZS)
      ((morph $ "ko")((ENTRY ko)(P-O-S DEC)(CASE LGE)
            (SF1 @ADJ_COMPL>)(SF2 @<ADJ_COMPL)(SF3 @ADV_COMPL))
           (36)(PLACE MOR_LERAOU)
     ((morph $ "Rik")((ENTRY ik)(P-O-S DEC)(CASE PAR)
            (DETER INDEF)(SF1 @OBJ)(SF2 @SUBJ))(2)(PLACE MOR_LEAAR1)
... )
```

Fig.5. Morphophonologically analysed text with context detection features.

LEARNER_LEMMATISER: this is an adaptation of the lemmatiser for Basque (Aldezabal et al. 1994). In our case we have suppressed some filters of the original lemmatiser in order to show the set of possible lemmas at the language level. The reason for suppressing the filters is that in the case of corpora written by language learners, the number of deviant structures is obviously higher than in those written by Basque native speakers. It must be taken into account that the goal of the original lemmatiser was to find the lemma of words, that is why it discarded deviant interpretations when it exists at least one standard interpretation of the word.

```
/<arriskurik>/
      ("arrisku" /arrisku/ NOUN COMM LENGTH10 VOWEL +
                DEC PAR INDEF MOR_LEAAR1)
      ("harri" /harri/ NOUN COMM LEM_LEDBH LENGTH10 VOWEL +
                DEC INS INDEF MOR_LERAZS + DEC LGE MOR_LERAOU +
                DEC PAR INDEF MOR_LEAAR1)
                                        ...
```

Fig.6. Morphophonologically analysed text: set of possible lemmas at the language level.

LANGUAGE_LEVEL_BASED_DISAMBIGUATOR: the goal of this tool is to discard interpretations that are not possible in the language level which we analyse. Thus, the rules for disambiguating the analysis of the word will depend on the language level. For example, in high levels the likely number of deviation rules that can be applied in a morphological interpretation, that makes sense, is not higher than two. However in low levels we have found possible interpretations that make sense where the number of deviation rules is higher than two. We have determined the exact number of possible deviation rules for an interpretation that makes sense for some language levels (the highest) and we are working on the others.

For example, in the analysis of the word *arriskurik* (any danger) we discard the interpretation *harrizkorik* 'any (thing ->*elliptical*) made of stone' which doesn't make sense (4 rules applied). Four interpretations from the original five are discarded using this type of disambiguation rules.

```
/<arriskurik>/
 ("arrisku" /arrisku/ NOUN COMM LENGTH10 VOWEL +
                DEC PAR INDEF MOR_LEAAR1)
                                        ...
```

Fig.7. Disambiguated text using specific features for each learning level.

CONTEXT_BASED_DISAMBIGUATOR: this tool will use some parts of the Constraint Grammar (CG) module for Basque (Alegria et al. 1996, Karlsson et al. 1995). It will also disambiguate unlikely interpretations in which some morphological rules have been applied in parts of the word (lemma/morpheme) where they never appear in real life examples. Some interpretations will also be discarded depending on the part of speech of the word where a rule has been detected. Others will be deleted because of the high rate which results from comparing the

number of rules and the length of the word. There will be a submodule of rules that will depend on the language level. A lot of the rules of the Context_Based_Disambiguator are decided based on the results of the tool Statistics. The Context_Based_Disambiguator module has been designed but not completely implemented yet. At the moment, in some cases, we disambiguate unlikely interpretations by hand.

STATISTICS: this is an interactive program that selects subsets of the corpora following different options: language level, type of exercise, student, specified text or specified list of texts; then, it offers you some data such as: 1) the number of interpretations where a concrete rule has been applied, 2) the rules that appear only in morphemes, 3) the number of applications of a rule depending on the average of the length of the words in the corpora, and so on. This program works with the results of learner_analyser (the adapted morphological analyser) and helps to define disambiguation rules of the Context_Based_Disambiguator. For example, if we determine using Statistics that a specific rule always appears in lemmas in interpretations that make sense, we can discard an interpretation where the rule is applied in a morpheme. This occurs with the LERAZS rule (Replace Z by S Anywhere in the word). That rule to make sense must appear in lemmas, but as we can see below it has been applied in a morpheme when we have analysed the word *analisis* 'analysis'. That interpretation is not probable in the mind of a student, so we discard it by means of rules in the Context_Based_Disambiguator.

```
/<analisis>/
    ("analisi"  /analisi/  NOUN COMM  LENGTH8  VOWEL +
                DEC INS INDEF MOR_LERAZS)     0  DISCARDED
    ("analisi"  /analisi/  NOUN COMM LEM_LEAES  LENGTH8  VOWEL +
                DEC ABS INDEF)
    ("analisi"  /analisi/  NOUN COMM LEM_LEAES  LENGTH8  VOWEL)
```

Fig.8. Disambiguated text using Constraint Grammar.

INTERLANGUAGE_MODELS: the aim of this tool is to create a previous version of the interlanguage structures belonging to the interlanguage level of the corpora we are studying. By means of this tool, after disambiguating unlikely interpretations, we identify for each applied rule the context in which it is applied (place of the word, length of the word, last letter of the root (vowel/consonant), part of speech, type of exercise ...). The program is implemented even though the results are not definitive because we are still obtaining some small results from the module Context_Based_Disambiguator.

The output of Interlanguage_Models is as follows:

```
(Ident "LEAAR1")(Descrip " Adding the Epenthetic r")
(Example (arriskurik arrisku))
(Left_context(  ([V:=(=:0)|V+-:==-](+:=)+)  ()  ()  (/:0(+:=)+)))
(Right_context(  ()  ([+:=|$:$])  (%:0[+:=|$:$])  ()))
(Operator((<=>)(<=>)(<=>)(=>)))
(Linguistic_context(( NOUN DEC MORPH*)))
(Textual_context ( VOWEL  LENGTH>=  S))
```

Fig.9. An interlanguage structure for a standard phenomenon.

```
(Ident "LEAES")(Descrip " Adding the s letter at the End of the root ")
(Example (analisis analisi))
(Left_context( (i:i)))
(Right_context( ([+:=|$:$])))
(Operator((=>)))
(Linguistic_context((NOUN LEMMA)))
(Textual_context ( VOWEL LENGTH>= S))
```

Fig.10. An interlanguage structure for a deviation phenomenon.

Left context, right context and operator are two-level information. LENGTH>= means words whose length is higher or equal to the average length of words in the corpora. S is the type of exercise of the text where the rule is applied.

3.4. Identification of the fixed and variable knowledge and Evaluation of the results taken from the analysis of the corpora.

When we detect interlanguage structures from the results of the corpora, how do we decide when a structure belongs to the fixed knowledge or to the variable knowledge? We will comment on some criteria we have applied. These criteria are the outcome of some experiments we have done up to now.

Firstly, we have detected that for all rules, for standard and deviation phenomena, the frequency of use is not the same, even in native speakers. Therefore, the probability of a rule been fixed depends on the specific rule. Secondly, we have also identified some rules that disappear at high levels. The consequence is: if a rule which exists in low levels disappears at a given level X and at all the higher levels, it means that this rule will always be in the variable knowledge for any level lower than X. In some cases when a variable rule disappears at a given level it may mean that another rule has been fixed at the same time (see the examples of the next section). Third, the number of different roots of words of the analysed corpora has influence in deciding fixed and variable knowledge; for example, if in a corpus where few different roots of words appear we don't find a rule, we can not say that in that level that rule doesn't exit. Maybe we haven't found it because of the low variety of roots. And, the last criteria we have tested is that we must take into account the type of exercise when deciding fixed and variable knowledge.

The evaluation of the results obtained from the analysis of the corpora is made by means of other subsets of the corpus of the same language level and the opinion given by a language teacher. We have compared the results of a subset of the corpus (16 texts with 5318 words in total) and the evaluation of the teacher at high language levels. In the comparison the results are quite successful. Nevertheless, in some specific cases the results from the corpora and the opinion of the teacher are not the same: for example, when analysing the use of the h letter in the lemma of the word, the teacher said that in deviation structures it is more likely to delete the h letter when it should appear than to add it when it should not be added; however, in the results of the corpora, even though the hypothesis of the teacher is fulfilled the difference in the application rate between the two rules is not so clear.

4 Some results and examples in the study of stability

In this section we will comment on some phenomena detected when comparing the results between two corpora of upper intermediate and high language levels. The aim of presenting these results is to show some cases that show the development of the fixed and variable knowledge. We will see some examples of rules which:

a) Belonged to variable knowledge at upper intermediate, but have disappeared at high level.

b) Belong to variable knowledge of both levels.

c) Appear in the fixed knowledge in the high level, but they didn't exist at upper intermediate.

d) Belonged to variable knowledge at upper intermediate, but have disappeared at high level similar to (a) case but are also related to other rules that at the same time, have changed from variable to fixed knowledge. This is due to the interference between rules in the process of knowledge compilation.

In the (a) case we have detected the use of the c and v letters in loanwords from Spanish. The c and v letters don't exist in most Basque words (except for proper nouns …). Both are deviation rules that don't appear at high level.

When borrowing words from Spanish the g and j letters must sometimes be maintained and some others must change from g to j and vice versa. This rule belongs to (b) case.

An example of rules of the c type is the linking of verbs in the first and second singular persons that ends in t and n respectively plus the suffix -n in order to make relative clauses. (e.g. *dut + n = dudan; dun + n = dunan*).

Finally we will comment on one example of d type where we can see interferences between rules in the process of knowledge compilation:

Let us explain the next three morphological rules (the first two rules are standard, while the third can be a deviation depending on the language level):

1) LEAEE (Add the E epenthetic LEtter at the End of the root of the word when the last letter of the root is a consonant and a declension morpheme starting with a will be added on the right) - e.g. *oihan + an -> oihanean* 'in the forest'.

2) e+e = ee (when the last letter of the root is an e and the first letter of the next morpheme is also an e, deletion doesn't happen) - e.g. *etxe + ekin -> etxeekin*) 'with the houses'.

3) *LEAIE (Add the E epenthetic LEtter Inside the word when the last letter of the root is a consonant and the declension morpheme to be linked on the right starts with e) - e.g. *euskaldun + ekin -> *euskalduneekin* 'with Basque speakers'.

In the development of the knowledge from upper intermediate to high level we have detected in the corpus analysis that the students of upper intermediate use both *oihanean* and **oihanan*. They also use *etxeekin* and **etxekin, *euskalduneekin* and *euskaldunekin*. At high level, however, they only use *oihanean, etxeekin*, and *euskaldunekin*. Words as **euskalduneekin* don't appear at this level.

upper intermediate level	LEAEE (oihanean/*oihanan) e+e=ee (etxeekin/*etxekin) *LEAIE (*euskalduneekin/euskaldunekin)	
high level		LEAEE (oihanean) e+e=ee (etxeekin)

Fig. 11.

When students write **euskalduneekin* they are unconsciously applying the LEAEE rule without taking into account the conditions for application. They also apply the rule in the standard version (e+e=ee). As a consequence of this interference the deviant *LEAIE rule has appeared at upper intermediate level and disappears when the first two rules are correctly learned (fixed) at high level.

5 Conclusions and future work

The presented work proves that the implementation of a semiautomatic system for the study of interlanguage is viable from the adaptation of the linguistic-computational tools we have for the automatic study of Basque.

The results obtained using the developed tools for language learning studies provide us statistical information about phenomena teachers and psycholinguistics had a sense of. Quite a lot of phenomena have been demonstrated whereas others can't be corroborated using the results.

Field studies have been carried out (Maritxalar et al. 1993; Andueza et al. 1996). At the moment we are studying some aspects of the morphosyntax and syntax for L2 taking as a basis the results obtained in Andueza et al. (1996), where in a final test with the students the hypothesis obtained in the study carried out in the 94/95 academic year were contrasted. In that work, we analysed the reasons for using some language structures in addition to the detection of their context. We plan in the future to add to the system such knowledge about the diagnosis of structures.

In the near future we will develop new tools in order to model the student's knowledge about a second language. The detection of contexts by means of Interlanguage_Models (see section 3.3) will be improved in order to identify contexts related to the characteristics of the particular learners. Some examples explained in the section 4 showed that the phenomena of borrowing words from the mother tongue is important in the study of the fixed and variable knowledge. Therefore, in the future, we will also think about including knowledge of the mother tongue. The computational tools which we have are independent from the language, that is why Spanish morphology could be included without changing the implementation of the tools.

Finally, we would like to remark that together with the experiments explained in the article two environments are being prepared: the Teacher's Assistant and the Student's Assistant. The Teacher's Assistant helps language teachers to make hypotheses about, among others, the reasons students have for using deviant language structures. The Student's Assistant guides users in their learning process giving hints according to their language level.

Appendix 1

@ADJ_COMPL: Adjectival Complement.
@ADV_COMPL: Adverbial Complement.
@OBJ: Object.
@SUBJ: Subject.
ABS: Absolutive.
anal: Analysis.
COMM: Common. i.e. Common Noun.
DEC: Declension morpheme.
DETER: Determination.
INDEF: Indefinite.
INS: Instrumental.
LGE: Locative Genitive.
morph: Morpheme.
P-O-S: Part of speech.
PAR: Partitive.
S-O-P-S: Sub-Part of speech.
SFn: Syntactic Function.

- The words with @ are syntactic codes.

- Codification of the linguistic rules:

LE: LEtter.
LEIX: The X LEtter has been Invented.
LECX: The position of the X LEtter has been Changed.
LEDPX: Delete the X LEtter in the P position.
LEAPX: Add the X LEtter in the P position.
LERPXY: Replace X by Y in the P position.
P position could be:

B: At the Beginning of the root.
E: At the End of the root.
I: Inside the word
A: Anywhere in the word.

Examples:

LEAES: Add the S LEtter at the End of the root.
LERAZS: Replace Z by S Anywhere in the word.

References

Agirre, E., Alegria, I., Arregi, X., Artola, X., Díaz de Ilarraza, A., Maritxalar, M., Sarasola, K., Urkia, M. 1992. XUXEN: A spelling checker/corrector for Basque based on Two-Level Morphology. Proceedings of the Third Conference ANLP (ACL). 119-125. Trento, Italy: ACL.

Agirre, E., Agirre, X., Arriola, J.M., Artola X., Insausti, J.M. 1994. Euskararen Datu-Base Lexikala (EDBL), Internal Report. UPV/EHU/LSI/TR 8-94. Computer Science Faculty : University of the Basque Country.

Aldezabal I., Alegria I., Artola X., Díaz de Ilarraza A., Ezeiza N., Gojenola K. Aduriz I., Urkia M. 1994. EUSLEM: Un lematizador/etiquetador de textos en euskara. Proceedings of the X. Conference SEPLN. Córdoba : SEPLN

Alegria I., Arriola J.M., Artola X., Díaz de Ilarraza A., Gojenola K., Maritxalar M., Aduriz I. 1996. A Corpus-Based Morphological Disambiguation Tool for Basque. Proceedings of the XII. Conference SEPLN. Sevilla : SEPLN

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.

Bull, S. 1994. Student Modelling for Second Language Acquisition. Computers & Education 23.13-20.

Corder, S. 1971. Idiosyncratic dialects and error analysis. International Review of Applied Linguistics 9. 147-59.

Karlsson, F., Voutilainen, A., Heikkilä, J., Anttila, A. 1995. Constraint Grammar: a language-independent system for parsing unrestricted text. Mouton de Gruyter.

Lessard, G., Maher, D. & Tomek, I. 1994. Modelling Second Language Learner Creativity in Journal of Artificial Intelligence in Education 5(4). 455-480.

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

Maritxalar, M., Díaz de Ilarraza, A. 1994. An ICALL System for Studying the Learning Process. Computers in Applied Linguistics Conference. Iowa State University.

Maritxalar, M., Díaz de Ilarraza, A. 1996a. Hizkuntza baten Ikaskuntza-Prozesuan zeharreko Tartehizkuntz Osaketa: Sistema Informatiko baten Diseinurako Azterketa Psikolinguistikoa. Internal Report. UPV/EHU/LSI/TR 7-96. Computer Science Faculty : University of the Basque Country.

Maritxalar, M., Díaz de Ilarraza, A., Alegria, I., Ezeiza, N. 1996b. Modelización de la Competencia Gramatical en la Interlengua basada en el Análisis de Corpus. Proceedings of the XII. Conference SEPLN. Sevilla : SEPLN

Nemser, W. 1971. Approximate Systems of Foreign Language Learners. International Review of Applied Linguistics 9. 115-23.

Selinker, L. 1972. Interlanguage. International Review of Applied Linguistics 10. 209-31.

Selinker, L. 1992. Rediscovering interlanguage. London: Longman.