From Psycholinguistic Modelling of Interlanguage in Second Language Acquisition to a Computational Model

Montse Maritxalar
DeptofComputer
LanguagesandSystems
UnivoftheBasqueCountry
649postakutxa20080
DonostiaSpain
jipmaanm@si.ehu.es

Arantza Díaz de Ilarraza
DeptofComputer
LanguagesandSystems
UnivoftheBasqueCountry
649postakutxa20080
DonostiaSpain
jipdisaa@si.ehu.es

Maite Oronoz
DeptofComputer
LanguagesandSystems
UnivoftheBasqueCountry
649postakutxa20080
DonostiaSpain
jiboranm@si.ehu.es

Abstract

The present article demostrates the implementation of a psycholinguistic model of second language learners' interlanguage in an Intelligent Computer Assisted Language Learning (ICALL) system for studying second language acquisition. We have focused our work on the common interlanguage structures of students at the same language level. The Interlanguage Level Model (ILM) is made up of these structures.

In this paper we explain the conceptual model of ILMs, we present the experimental method followed for collecting written material, we describe the output of the modelling tool, and finally some conclusions and future work are outlined.

1 Introduction

The main goal of this article is to show the implementation of a psycholinguistic model of second language learners' interlanguage in an Intelligent Computer Assisted Language Learning (ICALL) system for studying second language acquisition. The ICALL system, and the computational model for representing interlanguage, have been designed after previous work on psycholinguistics, artificial intelligence, and computational linguistics. This article will be focused on the computational model of interlanguage. Description of the ICALL system can be found in (Maritxalar et al., 1994).

The concepts transitional dialects (Corder, 1971) and approximate systems (Nemser, 1971) are precursors to interlanguage (Selinker, 1972) (Selinker, 1992). Their aim was to define communicative and grammatical competence in second languages. All of them have these common features: 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. In our case, the above mentioned characteristics are the basis for modelling the interlanguage. For example, as 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. Therefore, in our ICALL system we find a module where the different Interlanguage Level Models (ILMs) of each language level are represented. Representation of the ILMs will be presented in this article.

Our work is based on corpus analysis. At this moment we have focused on the implementation of 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. We are aware of the limits of the modelling only taking into account written material. Spoken material could also be collected for a better modelling. However, as the computational tools we have for studying Basque are only for writing studies, we have discarded the treatment of the spoken output of the learners.

In the next section we will explain the conceptual model of ILMs by means of the KADS Domain Description Language (DDL) proposed in Schreiber (Schreiber et al., 1993). Next, we will give an idea of the experimental method used in order to collect written material and the top-down methodology used in order to model interlanguage in support of the selected corpora. After that, we will describe some of the implemented tools for modelling ILMs.

Then, we will describe the output of the modelling tool, in order to compare the information given by the computational tool and the conceptual model of ILMs proposed before. Finally, some conclusions and future work are presented.

2 A Conceptual Model for Interlanguage Level Models

In this section we show the psycholinguistic model for Interlanguage Level Models (ILMs). The ILMs characterise different grammars of the interlanguage the ideal learner can have for the second language at each language level. It must be noted that, although some linguistic structures are particular to each student, others are common to all students at the same level. These common structures are those which will be represented in the ILMs.

We represent ILMs by means of two different types of sequences of models: *consecutive* sequences and *embodied* sequences. For each level, two different kinds of knowledge are represented: *the variable knowledge*, the knowledge the student is supposed to be learning; and, *the fixed knowledge*, the knowledge the learner has already acquired. The knowledge the language learners have acquired at a concrete level includes the knowledge acquired at previous levels. However, the knowledge they are learning at each level follows a different structure: each model from each level is independent from other levels, although intersections between knowledge at different levels can occur at times. The set of different variable language structures at each language level consists of consecutive sequences of models, while the set of fixed ones consists of embodied sequences (see Fig. 1). When we say fixed knowledge or variable knowledge we refer to the language structures the learners have in their interlanguage, without making any distinction between the representation of correct and "incorrect" structures. When we specify interlanguage all the language structures, right and "deviant", are represented in the same way. It must be noted that while a language structure used by high level learners can be considered deviant, the same structure in the case of beginners could be seen as correct at their level. For example a deviation at 10th level like avoiding the word *ote* 'could' (i.e. *nor da?* 'who is?' instead of *nor ote da?* 'who could be?') is not considered a deviation in lower levels of the language.

Language structures of the interlanguage are context-dependent (Selinker, 1992). Language learners create discourse domains as contexts for interlanguage development.

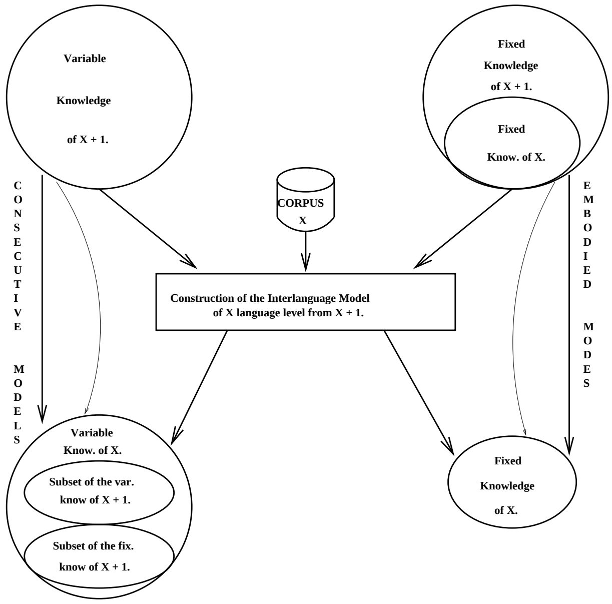

Fig 1. Construction of the Interlanguage Model of X from X+ 1.

Such domains are constructed in connection with life experiences that have importance for the learner, containing prototypical interlanguage forms associated with the content area by the learner. Interlanguage may be developing in one domain while at the same time it may be stabilised, or possibly fossilised, in another (Selinker, 1992). So, language structures of the interlanguage include information about context where they appear. Some structures appear in the interlanguage in specific contexts, however, others appear in any context. That is why we can see, in the representation below that structures can be context-dependent or context-independent. A first approximation of the representation (using DDL) for the ILMs is as follows:

structure Interlanguage_Level_Models;
parts:
models:
set(instance(Interlanguage_Level_Model));

structure *Interlanguage_Level_Model;*
parts:
fixed_knowledge:
set(instance(interlanguage_model));
cardinality:
min 0 max *HIGHEST_LEVEL;*
variable_knowledge:
set(instance(interlanguage_model));
cardinality:
min 0 max *HIGHEST_LEVEL;*

structure *interlanguage_model;*
parts:
structure_context_indep:
set(instance
(interlanguage_structure_context_indep));
structure_context_dep:
set(instance
(interlanguage_structure_context_dep));
properties:
language_level:
integer-range(1, *HIGHEST_LEVEL);*

structure
interlanguage_structure_context_indep;
subtype-of: interlanguage_structure,

structure *interlanguage_structure_context_dep;*
subtype-of: interlanguage_structure,

structure *interlanguage_structure;*
subtypes:
interlanguage_structure_context_indep,
interlanguage_structure_context_dep;
parts:
phenomena:
set(instance(linguistic_phenomenon));
conditions:
set(instance(interlanguage_condition));
properties:
description: **string;**
contextualization: **boolean;**
deviation: **boolean;**
stabilization:
function(stabilization of the rules
associated to each phenomenon);
axioms:
If contextualization=False
then conditions=<>.
The same interlanguage structure can be
deviation=True at the X language level
and deviation=False at the Y language level.
The value of *stabilization* is
{rarely, sometimes, usually, always}

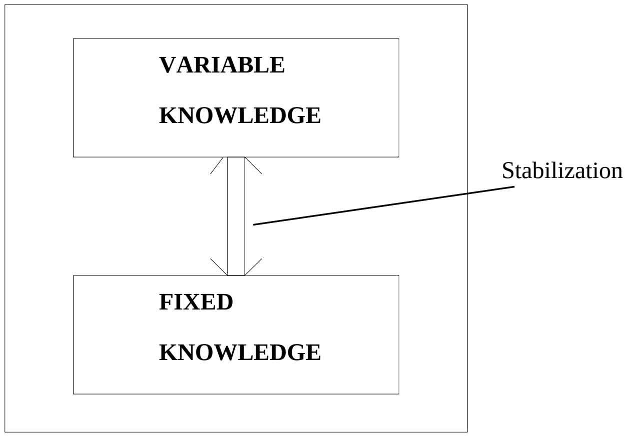

Fig 2. Interlanguage.

The interlanguage structures we propose are composed of *linguistic phenomena* which occur under some *conditions*. The properties of the interlanguage structures are: *description* of the structure; *contextualization*, that is, context dependent or context independent; *deviation*, which marks if the structure must be considered deviant at the related language level and, finally, *stabilization*.

The stabilization property is a qualitative value which represents the acquisition level of the structure for a particular language level. It is by means of this property that a language structure inside the student's interlanguage is considered fixed knowledge or variable knowledge. When the value of the stabilization property is *always* it means that the language structure has been assimilated by the learner and in the future it would be quite difficult making some changes. However, when the value is *rarely*, there is a high variability of the language structure, and, therefore, a good teacher (could be an ICALL system) should be able to help the learner to take away those structures in case they would not direct the student towards a target language.

In the first approximation of the representation for the ILMs we have represented interlanguage structures as a set of linguistic phenomena depending on interlanguage conditions. Now, we will explain those conditions, and, we will also describe the definition of linguistic rules and replacements which define the linguistic phenomena.

structure *linguistic_phenomenon;*
parts:
rules:
set(instance(linguistic_rule))
| **set(tuple**(replacement_rule);
properties:
type:
function(type of the rules or replacements
associated to each phenomenon)
description: **string;**
global: **boolean;**
lexical_entry: **set(instance**(morpheme));
axioms:
If global=False then lexical_entry = < >

```
structure interlanguage_condition;
  parts:
   linguistic_context:
     set(instance(linguistic_condition));
   non_linguistic_context:
     set(instance(non_linguistic_condition));
```

Which kind of conditions must be considered in the interlanguage? We said above that, inside the interlanguage, language structures are context-dependent or context-independent (Selinker, 1992). Referring to context-dependent structures it is very important to differentiate between two types of contexts: non-linguistic and linguistic.

Non-linguistic conditions are related to a discourse domain. Interlanguage varies depending on the domain. The linguistic structures that we activate when writing a story are different to those we activate when we write scientific-technical texts (thematic_conditions). Activity types, such as fielding questions, writing a letter, translating a sentence, conversing in a group etc. are also non-linguistic conditions.

Interlanguage forms can vary from one activity type to another, even though the discourse domain is the same. The activity types can be related to the structure of the whole text (text_conditions); in addition there are also some non-linguistic conditions (e.g. length of the sentence) related to a particular language structure in the text (structure_conditions). For example, sometimes students mark agreement between verb and complements in small simple sentences, and, at the same time they forget it in long sentences. In our case, structure conditions are studied by means of the corpus; text conditions, however, are detected by means of interviews with the teachers and learners.

```
concept condition;
  subtypes:
   linguistic_condition, non_linguistic_condition;
  properties:
   description: string;

concept linguistic_condition;
 subtype-of: condition;
  properties:
   word_level:
     list(part_of_speech, declension_case ...)
   sentence_level: list(use_of_subordinates ...)
```

```
concept non_linguistic_condition;
  subtype-of: condition;
  subtypes:
     textual_condition, thematic_condition;

concept textual_condition
  subtype-of: non_linguistic_condition;
  properties:
   text_conditions:
     list(summary, formal_letter,
        translation ...);
   structure_conditions:
     list(long_sentence_based, long_word_based,
        apl_place_lem, apl_place_mor ...);

concept thematic_condition
  subtype-of: non_linguistic_condition;
  properties:
   type: {general, technical};
```

Linguistic conditions are studied by means of a corpus. In the corpus composed of texts written by high level and upper intermediate language students (350 texts collected between 1991 and 1995) we found linguistic influence in some language structures of the interlanguage. For instance, the presence of plurality of some components in the sentence can cause verbs to agree with such components, whether or not these must be in agreement with the verb (the phenomenon of plurality has also been observed in second language learners of French (Lessard et al., 1994)). Finally, we would like to claim that this way of putting linguistic structures in context could be applied similarly when modelling first languages.

After defining which kind of conditions must be taken into account when linguistic phenomena are identified, we will see which type of linguistic rules can be found inside the phenomena.

It is usual that language learners know only some linguistic rules corresponding to a particular linguistic phenomenon, and not all of them. In the case of language natives, however, they know, in most cases, all linguistic rules. That is the reason for representing explicitly in the student's interlanguage the set of linguistic rules related to the linguistic phenomena of the language structure. In other cases, as in natives, it should not be necessary to make their linguistic rules explicit as the linguistic phenomena define, implicitly, the set of linguistic rules.

In the same way, we could say that linguistic rules identify the corresponding linguistic phenomenon, however, it is necessary to make it explicit, as a linguistic phenomenon could be detected in a student's interlanguage, but the linguistic rules would not be identified until after some interactions with the student. The student model is dynamic, so, the language structures are also dynamic. Therefore, in the first interactions with the student a linguistic phenomenon can be detected inside the student's interlanguage before eliciting the corresponding linguistic rules. Next we shall define the structure of the linguistic rules.

structure *linguistic_rule;*
subtypes: morphological_rule, syntactic_rule;
parts:
implemented_by: **set** (rule_identifier);
cardinality: min 1;
conditions: **set(instance** (interlanguage_condition));
properties:
type: {morph, syn};
description: **string;**
example: **string;**
stabilization:
{rarely, sometimes, usually, always};

In the experiments the teachers have identified three types of language phenomena in the students' interlanguage: first, simple interlanguage structures composed of a set of linguistic phenomena and linguistic rules; second, avoiding interlanguage structures defined as linguistic rules that the student usually avoids; and, third, replacement language structures where the learner uses a structure too often (or rarely) instead of using other structures (e.g. when a person uses the conjunction *and* all time and rarely uses structures such as *however, nevertheless, thus* ...). Consequently, we can say that there are relationships between linguistic rules of the interlanguage. That is why structures which represent linguistic phenomena also have a set of replacement tuples, which represent the replacement relations between linguistic rules.

3 Using Corpus Linguistics in order to Model Interlanguage

In this section we will explain first the experimental method used in order to collect written material, and second the top-down methodology used in order to model interlanguage in support of the selected corpora. We model Interlanguage Level Models by means of automatic tools which use the collected material as input. It must be noted that some information of the interlanguage models, for example text conditions (see section 2), are detected semiautomatically with the help of the teachers and learners.

Before explaining modelling based on corpus analysis, 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 such 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, the learner's code, and the type of exercise proposed (e.g. s, summary). The last feature is what we have called text condition in section 2. At the same time, a database for gathering the relevant information about the students' learning process was developed. We retrieved such information from interviews with the students and the teachers (Andueza et al., 1996). The corpus collected from 1990 to 1995 is made up of 350 texts. This corpus has been divided in subsets depending on the language level. At the moment we have defined three language levels of study that we call low, intermediate, and high levels.

Before designing and implementing the automatic tools, three different 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) were carried out. 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 on studying the corpora during the 94/95 academic year. The language learners had five hours of language classes per week, and they wrote one composition every week or every fortnight.

For modelling interlanguage at different language levels we use a top-down methodology, that is, we start from the modelling of high levels and continue to lower ones (see Fig 1). The reasons for

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

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 is more written material at high levels than at low ones.

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 deeply and automatically studied.

The steps we followed using the tools we have adapted in order to build the interlanguage model for each N language level were:

1. Design of the lexical database for the Nth language level.
2. Selection of the corpus (CORPUS-N) and subsets of CORPUS-N to be used in the next steps. This selection is based on the criteria, explained before, for collecting material.
3. Definition of the morphology and morphosyntax based on a subset of CORPUS-N.
4. Identification of the fixed knowledge and the variable knowledge, considering the contexts defined in section 2.
    (a) Evaluation of the reliability of the model using other subsets of CORPUS-N.
    (b) Evaluation of the results by a language teacher of N level.

For example, in studies of high language modelling, a teacher evaluated the results at word level, that is, the type of rules detected and the contexts where they were applied. The evaluation was 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. This phenomenon has not been detected in the results of the corpus).

4 Implementation

We have adapted tools previously developed in our computational linguistic group during the last ten years. These tools are: the Lexical Database for Basque (EDBL) (Agirre et al., 1994), the morphological analyser based on the Two-Level Morphology (Agirre et al., 1992), the lemmatiser (Aldezabal et al., 1994) and some parts of the Constraint Grammar for Basque (Alegria et al., 1996) (Karlsson et al., 1995).

We have two main reasons for adapting these tools:

1. Some of the deviant linguistic structures used by second language learners are different to those native Basque speakers use. The context of application, i.e. structure conditions at word level of some rules, are not the same in both cases. Moreover, we need to add some new rules to have one rule for each linguistic structure, and we also need some of these rules in order to detect deviant linguistic phenomena, e.g. loan words from Spanish (see section 4.1).
2. In the original tools, the context of application of the rules remained ambiguous. As we explained in section 2, the context of application is important to us for modelling the grammatical competence of the students, so we disambiguate such contexts by means of our adapted tools (see section 4.2). In the figure below we can see a scheme of the way in which we have used these adapted tools (Diaz et al., 1997):

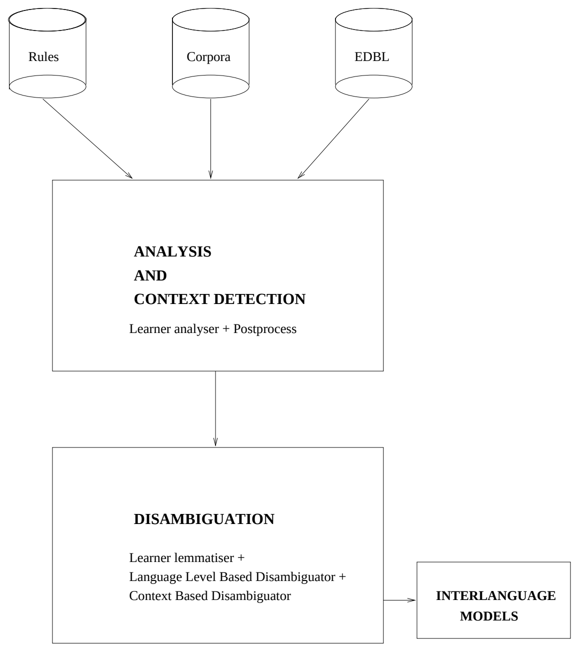

Fig 3. Modelling Process.

LEARNER ANALYSER = The adapted morphological analyser (detection of structural contexts).

POSTPROCESS = Context detection tool (detection of structural contexts and linguistic contexts).

LEARNER LEMMATISER = The adapted lemmatiser.

LANGUAGE LEVEL BASED DISAMBIGUATOR = Disambiguator for each language level (Based on the number of rules in each interpretation).

CONTEXT BASED DISAMBIGUATOR = Disambiguator for each language level.

(Based on subsets of the Constraint Grammar for Basque + disambiguation rules based on the context of application).

INTERLANGUAGE MODELS = Interlanguage Level Models (ILMs)

4.1 Redesigning the automata

As we said above, the original morphological analyser was based on the Two-Level Morphology. The implementation of the morphophonological rules were made by the automata. In the original analyser we had 30 automata: 11 of them were used for analysing standard words, but their activation was never detected; the other 19 which represented a deviation were identified but the type of deviation remained unknown. Moreover, the context of application remained ambiguous.

In the adapted analyser we find 59 automata, which represent 59 different types of phenomena (these are codified as we can see in the table below). We have modified the automata of the morphological analyser in order to detect which rules have been applied and the contexts (structure conditions) 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 in others that detect, in a more specific way, some morphological phenomena which are very interesting for the study of second language acquisition. An example will illustrate this fact:

Rules in the original analyser

h:0 => R:= (+:=)+ _ (Rule for standard phenomenon)
h:0 => [V:V — (0:*) :] _ V:V (Rule for deviant phenomenon)

[The application of the rule for standard phenomenon is not detected, however, the competence error represented in the rule for deviant phenomenon is identified even though the context of application remains ambiguous]

Rules in the adapted analyser

h:0 => R:= (+:=)+ _ (LEDEH: Delete the H LEtter at the End of the root)
h:0 => V:V _ V:V (LEDAH: Delete the H LEtter Anywhere in the word)
h:0 => (0:*) : _ V:V (LEDBH: Delete the H LEtter at the Beginning of the root)

[All three rules of the interlanguage and the context of application are detected when they have been activated. We repeat the same automata three times and mark as negative in each automaton the states which correspond to the activation of the rule.]

4.2 Rules in their Linguistic Context

Experiments and interviews with experts lead us to see the need of identifying the linguistic context where a morphological rule (for standard or deviant phenomenon) is applied. We identify this context by adding to the adapted morphological analyser *(Learner_Analyser)* some characteristics such as the place (lemma/morpheme) where the rule is applied, the length of the word and the type of the last letter (vowel/consonant) of the root *(Postprocess)*.

We have two main aims in mind:

1. Disambiguate unlikely interpretations of a word (Context_Based_Disambiguator).

   There are two ways to do this:

   * Discarding interpretations in which a morphological rule has been applied in a part of the word (lemma/morpheme) where it never appear in real life examples. For example, the deviant word **analisis* has two interpretations *(analisi/analisiz)*: the rule to add an s at the end of the lemma / the replacement of z by s in the morpheme. The second interpretation is not possible for high language level students, so we discard it at such level.

   * Discarding interpretations in which a morphological rule is applied within an unusual part of speech. The rule that detects the replacement of t by d is a good example of this. The rule is never used in verbs starting with d. After discarding all interpretations of the words, where the replacement rule has been applied and the part of speech is a verb, the number of interpretations in the analysis of the word is reduced to a half.

2. Refine the model of the student's Interlanguage (Language_Level_Based_Disambiguator).

   A word changes into another one quite different as a result of the application of an excessive number of rules who represent deviant phenomena. From the study of the corpora, we can determine the exact number of possible deviation rules for an interpretation that makes sense at each language level. At the moment, we have determined it for some levels (i.e. the highest) and we are working on the others.

5 The Output of the Modelling Tool

In this section we will show an example of an interlanguage structure in order to see the relationship between the output of the modelling tool and the description of interlanguage structures explained in the conceptual model.

In this example we will see a detected linguistic phenomenon in the corpus of high level learners. The description of the phenomenon is: "when learners want to construct relative clauses and the last letter of the verb is t, for example *dut* (auxiliary verb for composed verbs), when adding the suffix -n for constructing relative clauses, the t is replaced by d and the a letter is added. That is, *dut + -n = dudan*".

e.g. *Ikusi dudan haurra atsegina da.* 'The boy who I have seen is nice.'

Ikusi dudan	haurra	atsegina	da.
I have seen who	boy the	nice	is.

This example shows that Basque syntactic information is found inside the word. That is why in the modelling of the LC_1 linguistic condition (see the example) the REL feature (relative clause) is at word level, and not at sentence level.

An example of the output of the modelling tool:

```
(interlanguage_structure IE_1
  (phenomena (LF_1))
  (conditions (LC_1 TC_3))
  (description ?)
  (contextualization True)
  (deviation False)
  (stabilization usually))

(linguistic_phenomenon LF_1
  (rules (LEAIA_1 LERATD1_1))
  (type (morph))
  (description ?)
  (global False)
  (lexical_entry <>))

(linguistic_rule LEAIA_1
  (implemented_by (9))
  (conditions (LC_1 TC_1 TC_3))
  (type morph)
  (description
     "Add the A LEtter Inside the word")
  (example "dudan")
  (stabilization usually))

(linguistic_rule LERATD1_1
  (implemented_by (8))
  (conditions ( LC_1 TC_2 TC_3))
  (type morph)
  (description
     "Replace T by D Anywhere in the word")
  (example "dudan")
  (stabilization usually))

(textual_condition TC_1
  (text_conditions <>)
  (structure_conditions (apl_place_mor))
  (description
     "rule applied in the morpheme"))

(textual_condition TC_3
  (text_conditions <>)
  (structure_conditions (end_word_t))
  (description "the last letter of the word is t"))

(linguistic_condition LC_1
  (word_level (V REL))
  (sentence_level <>)
  (description "verb for relative clause"))

(textual_condition TC_2
  (text_conditions <>)
  (structure_conditions (apl_place_lem))
  (description "rule applied in the lemma"))
```

In the process of modelling, first, we identify the linguistic rules, second, we detect groups of linguistic rules which occur in the same context and define the linguistic phenomenon, and last, the interlanguage structure is identified.

If we compare this output with the conceptual model given in section 2, we can see how the information needed is reached automatically, except the description of the linguistic phenomenon and the interlanguage structure (see question marks in the example). Such information will be completed by the psycholinguist who will use the ICALL system. The information given by the psycholinguist will be reused in future modellings.

6 Conclusions and Future Work

The presented work proves that the implementation of psycholinguistic models of interlanguage is viable from the adaptation of the linguistic-computational tools we have for the automatic study of Basque. Modelling at word level is important in our case because Basque is an agglutinative language with rich morphosyntactic information within words.

The results obtained using the developed tools for language learning studies provide us statistical information about phenomena that teachers and psycholinguistics knew by intuition. Therefore, we can say that corpus analysis is a good technique for modelling ILMs.

In the near future we will develop new tools in order to model each student's knowledge. The detection of contexts will be improved in order to identify contexts related to the characteristics of the particular learners. Moreover we will work on the detection of interlanguage structures at sentence level . Field studies have been carried out in this sense (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 (Andueza et al., 1996), where in a final test the hypothesis obtained in the study carried out in the 94/95 academic year were contrasted with the students. In this work, we also 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.

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

7 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.

References

Eneko Agirre, Iñaki Alegria, Xabier Arregi, Xabier Artola, Arantza Díaz de Ilarraza, Montse Maritxalar, Kepa Sarasola and Miriam Urkia. 1992. XUXEN: A spelling checker/corrector for Basque based on Two-Level Morphology. In *Proceedings of the Third Conference ANLP(ACL)*, pages 119–125, Trento, Italy.

Eneko Agirre, Xabier Arregi, Jose Mari Arriola, Xabier Artola and Jon Mikel Insausti. 1994. Euskararen Datu-Base Lexikala (EDBL) *Internal Report. UPV/EHU/LSI/TR 8-94*, Computer Science Faculty: University of the Basque Country.

Izaskun Aldezabal, Iñaki Alegria, Xabier Artola, Arantza Díaz de Ilarraza, Nerea Ezeiza, Koldo Gojenola, Itziar Aduriz and Miriam Urkia. 1994. EUSLEM: Un lematizador/etiquetador de textos en euskara. In *Proceedings of the X. Conference SEPLN*, Córdoba, Spain.

Iñaki Alegria, Jose Mari Arriola, Xabier Artola, Arantza Díaz de Ilarraza, Koldo Gojenola, Montse Maritxalar and Itziar Aduriz. 1996. A Corpus-Based Morphological Disambiguation Tool for Basque. In *Proceedings of the XII. Conference SEPLN*, Sevilla, Spain.

A. Andueza, Arantza Díaz de Ilarraza, Montse Maritxalar, Josune Martiarena and Iñaki Pikabea. 1996. Hizkuntza baten Ikaskuntza Prozesuari buruzko landa lana. Sistema Informatiko adimendun baten oinarria. In *Internal Report. UPV/EHU/LSI/TR 8-96*, Computer Science Faculty: University of the Basque Country.

S. Bull. 1994. Student Modelling for Second Language Acquisition. In *Computers & Education 23*, pages 13–20.

S. Corder. 1971. Idiosyncratic dialects and error analysis. In *International Review of Applied Linguistics 9*, pages 147–159.

Arantza Díaz de Ilarraza, Montse Maritxalar and Maite Oronoz. 1997. Reusability of Language Technology in support of Corpus Studies in an CALL Environment. In *Language Teaching and Language Technology Conference,* Groningen, The Netherlands.

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

G. Lessard, D. Maher and I. Tomek. 1994. Modelling Second Language Learner Creativity. In *Journal of Artificial Intelligence in Education 5(4).*, pages 455–480.

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

Montse Maritxalar and Arantza Díaz de Ilarraza. 1994. An ICALL System for Studying the Learning Process. In *Computers in Applied Linguistics Conference.* Iowa State University.

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

Montse Maritxalar, Arantza Díaz de Ilarraza, Iñaki Alegria and Nerea Ezeiza. 1996b. Modelización de la Competencia Gramatical en la Interlengua basada en el Análisis de Corpus. In *Proceedings of the XII. Conference SEPLN*, Sevilla, Spain.

W. Nemser. 1971. Approximate Systems of Foreign Language Learners. In *International Review of Applied Linguistics 9.*, pages 115–123.

G. Schreiber, B. Wielinga and J. Breuker. 1993. KADS: A Principled Approach to Knowledge-Based System Development. Academic Press.

L. Selinker. 1972. Interlanguage In *International Review of Applied Linguistics 10.*, pages 209–231.

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