REUSABILITY OF NLP TOOLS FOR DETECTING RULES AND CONTEXTS WHEN MODELLING LANGUAGE LEARNERS' KNOWLEDGE

A. Díaz de Ilarraza, M. Maritxalar, and M. Oronoz.
Dept. of Computer Languages and Systems.
University of the Basque Country. 649 Postakutxa. 20080 Donostia. Spain.
{jipdisaa,jipmaanm,jiboranm}@si.ehu.es

Abstract

This work shows how the reusability of computational tools previously developed for language studies is possible in the research of second language acquisition. Reusability is possible if the tools are independent from the language. We have adapted some tools (lemmatiser, spelling checker-corrector, morphological disambiguator ...) in order to model the grammatical competence (Interlanguage) of Spanish learners of Basque, based on corpus analysis. The methodology for modelling includes the use of subsets of corpora for defining the models and the evaluation of these models by means of new subsets of corpora and expert teachers' opinion. Experiments and interviews with experts lead us to see the need of identifying the contexts where the grammatical rules are applied. We have made changes in this direction.

1 Introduction

The main aim of this paper is to show in which way we have adapted Natural Language Processing tools previously developed by our group for the treatment of Basque. We use the adapted tools in a system created for the study of second language acquisition. We will focus on the reusability of the morphological analyser, the Lexical Database and the morphological and morphosyntactic disambiguator based on the Constraint Grammar. We have adapted these tools in order to model the grammatical competence (Interlanguage) (Selinker 92) of Spanish learners of Basque, based on their written texts. We have selected the study of the morphology in the learning process because of its difficulty due to the fact that Basque is an agglutinative language with rich morphosyntactic information inside the word. One of the main changes in the morphological analyser (Agirre et al. 92) is the modification of the automata used in the implementation of the Two-Level Morphology (Koskenniemi 83). The adapted automata permit us to identify rules for standard phenomena for each language level, as well as rules for deviant phenomena, and the contexts (e.g. part of speech) where the rules are activated.

2 Corpus Studies for Modelling Interlanguage Structures

For modelling language structures at each interlanguage level we use a top-down methodology, that is, we start from the modelling of high levels and continue to lower ones. 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.

We have collected corpora and in order to codify the texts we have established a 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, **s**ummary). At the same time we have developed a database for gathering the relevant information about the students' learning process .

When modelling language structures we differentiate between general knowledge and contextual knowledge. Besides, we identify two types of knowledge based on the acquisition rate: *fixed knowledge*, composed of the structures which become stable from a crucial moment during the learning process and *variable knowledge* whose structures change through such process.

The methodology used for modelling is structured in three steps:

1. Different subsets of the corpus corresponding to different language levels are used for establishing the linguistic models at each language level concerning the deviant and correct knowledge.

2. New experiments using new corpora are performed in order to validate the models previously established.

3. Experts in teaching language are asked about the correctness of the results obtained in step 2.

Our studies lead us to specify the concept of deviation as an alternative to the idea of error. We will say that a language structure is a deviation

(Maritxalar et al. 93) when one of the following three conditions is fulfilled: A) the structure is incorrect – we usually call it 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?'). In our proposal deviations are part of the interlanguage and they are represented in the same way as standard structures.

We consider that a linguistic phenomenon has different perspectives. Before showing the tools used in order to detect language structures and their contexts, let us give a brief idea about these different perspectives: linguistic and superficial perspectives. For explaining them, we will use the examples below:

a) **oihanan* *(=> oihan**e**an)*
(= in the forest)

b) **dakilako* *(=> daki**e**lako)*
(= because he/she knows it)

In the case of **oihanan* as well as in **dakilako* from a purely **superficial** perspective we could codify the deviant structure as *LEDIE (Delete the E Letter Inside the word). However, from a **linguistic** perspective each of one has a different interpretation: while in the (a) case the student has deleted the epenthetic *-e* used for linking morphemes, the (b) case shows that the student has only replaced the causal morpheme *-lako* instead of *-elako.* This shows us that different layers must be considered when modelling knowledge of language learners. Our system detects both perspectives in order to infer the appropriate one.

3 Adapting the Existing Tools

We have two main reasons for adapting the tools we have, instead of using the original ones :

1. Some of the deviant language structures used by second language learners are different to those native Basque speakers use. The contexts of application are not the same in both cases. That is why in some cases we have modified the automata we had for representing typical errors of native Basque speakers. Moreover, we needed to add some new rules for identifying each new deviant phenomenon detected in language learners, e.g. those derived from the loan of words from Spanish.

2. In the original tools the context of application of the rules remained ambiguous. In our case, 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 .

Let us explain in this section the tools we have adapted. In the Figure 3. we can see the implemented programs: an adapted morphological analyser based on the Two-Level Morphology (divided in two modules Learner_Analyser + Postprocess), an adapted lemmatiser for Basque (learner_lemmatiser), a disambiguation tool for discarding unlikely interpretations of the language knowledge level (Language_Level_Based _Disambiguator), and the morphological and morphosyntactic disambiguator for each level based on the Constraint Grammar (Context_Based Disambiguator) .

New attributes such as stabilisation value; linguistic, textual and thematic contexts and the list of language levels have been added in the design of the adapted Lexical Database - we have based our design on the lexical database for Basque (EDBL) previously developed by our group and composed of 65,000 entries. Besides, it is possible to define more than one entry for the same morpheme. For example, the -a morpheme, which had one entry for the ergative case in EDBL, will have a second entry for the absolutive one to make possible the detection of deviations. It will be necessary to mark, for each language level, the entries which represent typical deviations.

Statistics and Interlanguage_Models complete the set of programs developed. Statistics is an interactive program that selects subsets of the corpora and offers us some data such as: 1) the number of interpretations where a concrete rule has been applied, 2) the place of the word (lemma/morpheme) where rules appear, and so on. This program helps us to define disambiguation rules for the Context_Based_Disambiguator. Interlanguage_Models creates a previous version of the interlanguage structures belonging to the interlanguage level of the corpora that we study. In the near future, we will use this information in IDAZKIDE, a learning environment for students of Basque, in order to help on the diagnosis of deviant structures detected during the interaction between the system and the learners.

3. 1 Modifying the Automata

As we said above, the original morphological analyser was based on the Two-Level Morphology. The implementation of the morphophonological rules was made by automata. In the original analyser we had 30 automata: 11 of them were used for analysing standard rules but their activation was never shown; the other 19 which represented deviations 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 Appendix). 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 in others that detect in a more specific way some morphological phenomena very interesting for the study of language acquisition of a second language. 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 but 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.]

3. 2 Rules in their Context

Experiments and interviews with experts show us the need of refining the models of the students' interlanguage by means of identifying the language level of the students and the contexts where the morphological rules (for standard or deviant phenomena) are applied. We identify these contexts adding to the adapted morphological analyser (Learner_Analyser) some characteristics as the place (lemma/morpheme) where the rules are applied, the length of the word and the type of the last letter (vowel/consonant) of the root (Postprocess).

Next we will see an example of the output of the tools before showing how we disambiguate interpretations:

In the analysis of the word *arriskurik* (any danger) we find two interpretations: *arriskurik* 'any danger' and *harrizkorik* 'any (thing *->elliptical*) made of stone' which doesn't make sense as a deviation.

```
/<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)
...
```

Figure 1: Morphophonologically analysed text: set of possible lemmas at the language level

We take into account two different ways to discard unlikely interpretations of a word:

1. The context of application of the rules (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 possibility is not possible for high language level students, so we discard this interpretation 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 in a word that begins with the d letter and the part of speech is a verb the number of interpretations in the analysis of the word is reduced to a half.

2. The number of applied rules (standard/deviant) and the suitable rules at each language level (Language_Level_Based_Disambiguator).

Here we will see two ways of discarding not suitable interpretations taking as basis the language level:

a) In the first phase of this module a filter of interpretations has been implemented. For some language levels there are rules for deviant phenomena which have sense only when all the interpretations of the word are deviant. When the word has deviant and standard interpretations the standard are the suitable ones

in the cases the applied deviant rules don't correspond to the language level, however, in the cases the deviant rules are possible at such level the standard and deviant interpretations must be taken into account. The parameters of the filter are: the language level and the set of deviant rules which are possible at such level.

b) A word changes into another one quite different as a result of the application of an excessive number of deviant rules. From the study of the corpora we conclude that for example at high levels we can discard interpretations where three or more deviant rules are activated.

For example in the analysis of the word *arriskurik* 'any danger' (see Figure 1.) we discard the interpretation *harrizkorik* which doesn't make sense (3 deviant rules applied).

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

Figure 2: Disambiguated text using specific features for each learning level.

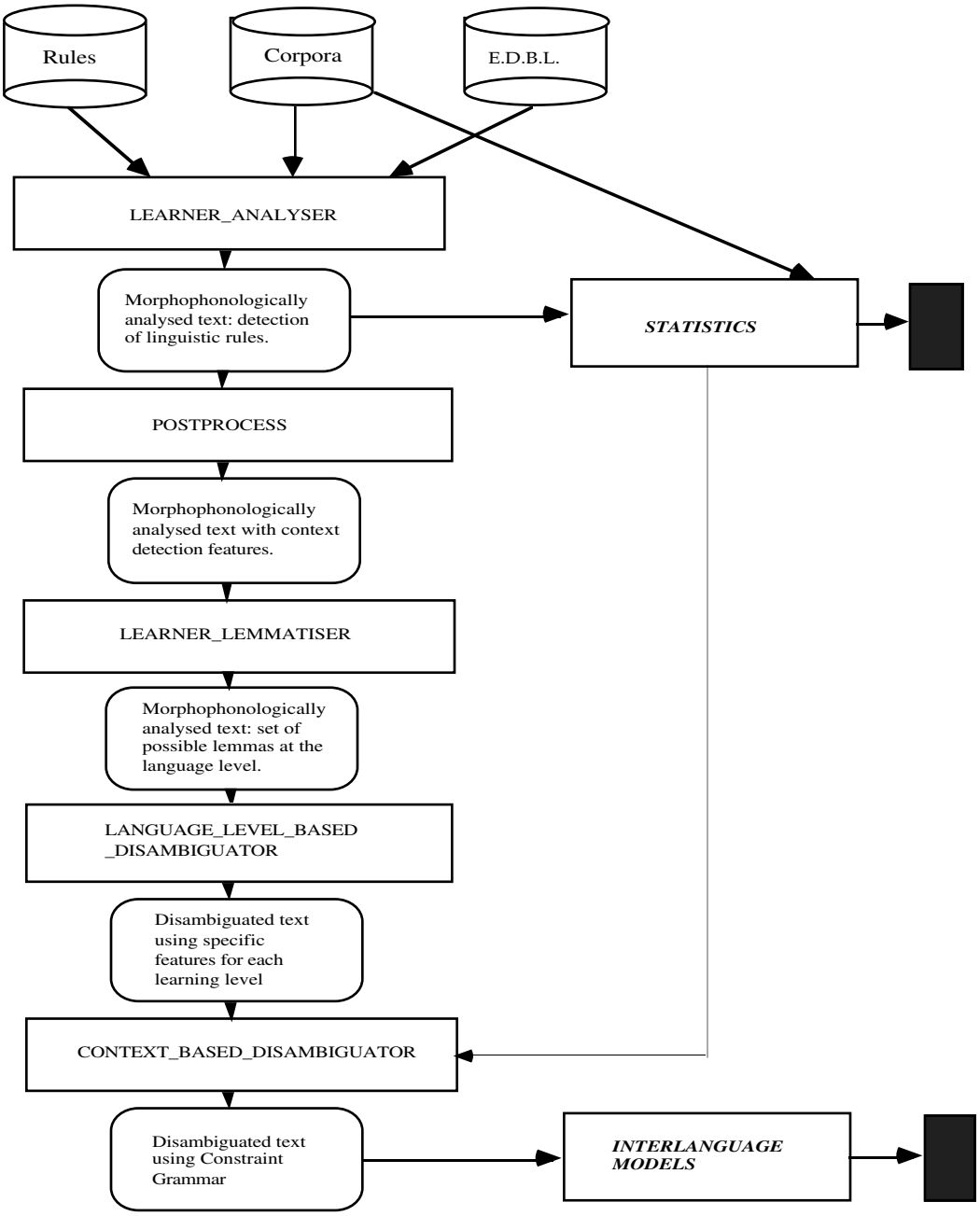

Figure 3: Process for the definition of Interlanguage Models

3. 3 Morphosyntactic Rules

By means of the submodule of the Constraint Grammar (see Figure 3.) for studying the morphosyntax (Alegria et al. 92; Karlsson et al. 95) we disambiguate morphosyntactic interpretations that are not possible at word level. We will see two examples of agreement between subject and verb. In the first case we define a rule for solving the ambiguity at any language level:

Example 1:
As said above, the morpheme -a has two different entries in the lexical database of language learners, the first one represents the ergative case and has associated from the lexicon the SUBJ syntactic function (@SUBJ), the second one represents the absolutive case and, from the syntactic point of view, has associated two different syntactic functions: subject and direct object (@SUBJ @OBJ):

("-a"... ERG ... @SUBJ)
("-a"... ABS ... @SUBJ @OBJ)

In the sentence *haurra etorri da* 'the child has come', we discard the first interpretation of *haurra* by means of the following rule of the Context_Based_Disambiguator:

(@w =0 (ERG) (0 ABS) (*1C CLB *LC)
(*LC IZAN))

[We discard (=0) the first interpretation (ERG) because on the right of *haurra* there is the intransitive verb *da* (IZAN) inside the same clause (CLB). The agreement ergative-intransitive verb is not possible].

Example 2:
When a learner of Basque says: *haurra ekarri du* we can understand two different meanings: a) **the child has carried it* or b) *He has carried the child*. Both interpretations are possible in second language learners.

At high language levels, after applying the morphosyntactic disambiguation only one interpretation will remain:

((lemma "haur") (morph "-a")... ABS... @OBJ)

The ambiguity is solved at high language levels but remain unsolved at intermediate levels. The reason is that the first interpretation of the -a morpheme defines a deviant phenomenon that can not occur in short sentences at high levels; in the case of intermediate levels, however, we would need pragmatic information for solving the ambiguity. Therefore, at intermediate language levels we will still have the following two interpretations :

((lemma "haur") (morph "-a") ... ERG ... @SUBJ)
((lemma "haur") (morph "-a") ... ABS ... @SUBJ @OBJ)

4 Conclusions and Future Work

The presented work shows how the reusability of computational tools previously developed for language studies is possible in the research of second language acquisition. Reusability is possible if the tools are independent from the language. In our case, we have changed the morphological and morphosyntactic rules of the analyser, by means of creating new automata and new morphosyntactic rules. Besides, some changes have been made in the module of the analyser which recognised the automata and, the rest of modules were reused without changing their implementation. As a result of the experiments and the study of the corpus, we identified different language perspectives of the same phenomenon. The idea behind our study is to recognise semi-automatically the grammatical competence of Spanish learners of Basque when writing texts.

The results obtained using the developed tools provide us statistical information about intuited phenomena previously detected by the teachers and the psycholinguists. Quite a lot of phenomena have been demonstrated whereas some can't be corroborated using the results. For example, in studies of high language level 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 fact has not been detected in the results of the corpus).

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. 96), 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 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 will be improved in order to identify the learners' models taking into account characteristics about the learning process of the particular learners.

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

(Agirre et al. 92) E. Agirre, I. Alegria, X. Arregi, X. Artola, A. Díaz de Ilarraza, M. Maritxalar, K. Sarasola and M. Urkia, *XUXEN: A spelling checker/corrector for Basque based on Two-Level Morphology*. Proceedings of the Third Conference ANLP (ACL) (pp. 119-125). Trento, Italy, 1992.

(Alegria et al. 96) I. Alegria, J.M. Arriola, X. Artola, A. Díaz de Ilarraza, K. Gojenola, M. Maritxalar and I. Aduriz, *A Corpus-Based Morphological Disambiguation Tool for Basque,* Actas del XII. Congreso de la SEPLN Sevilla, 1996.

(Andueza et al. 96) A. Andueza, A. Díaz de Ilarraza, M. Maritxalar, J. Martiarena and I. Pikabea, *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, 1996.

(Karlsson et al. 95) F. Karlsson, A. Voutilainen, J. Heikkilä and A. Anttila, *Constraint Grammar: a language-independent system for parsing unrestricted text.* Mouton de Gruyter, 1995.

(Koskenniemi, 83) K. Koskenniemi, *Two-level Morphology: A general Computational Model for Word-Form Recognition and Production*. University of Helsinki, Department of General Linguistics. Publications no. 11, 1983.

(Maritxalar et al. 93) M. Maritxalar and A. Díaz de Ilarraza, *Integration of Natural Language Techniques in the ICALL Systems Field: The treatment of incorrect knowledge*. (Internal report. UPV/EHU/LSI/TR 9-93). Informatika Fakultatea (UPV/EHU), 1993.

K. Oflazer, *Two-level description of Turkish morphology*. Literary and Linguistic Computing, vol. 9, No. 2, 137-148, 1994.

(Selinker 92). K. Selinker, *Rediscovering interlanguage*. London: Longman, 1992.

Appendix

@OBJ: Object.
@SUBJ: Subject.
ABS: Absolutive.
CLB: Clause Boundary.
COMM: Common. i.e. Common Noun.
DEC: Declension morpheme.
DETER: Determination.
ERG: Ergative.
INDEF: Indefinite.
INS: Instrumental.
LC: Left Context.
LGE: Locative Genitive.
morph: Morpheme.
PAR: Partitive.

• 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:
A: Anywhere in the word.
B: At the Beginning of the root.
E: At the End of the root.
I: Inside the word
Examples:
LEAES: Add the S LEtter at the End of the root.
LERAZS: Replace Z by S Anywhere in the word.