Mastering Overdetection and Underdetection in Learner-Answer Processing: Simple Techniques for Analysis and Diagnosis.

Olivier Kraif, Claude Ponton, Alexia Blanchard

Mastering overdetection and underdetection in learner answers processing: simple techniques for analysis and diagnosis

Alexia Blanchard, Olivier Kraif, Claude Ponton

LIDILEM - Laboratoire de LInguistique et DIdactique des Langues Etrangères et Maternelles (http://www.u-grenoble3.fr/lidilem/labo)
Université Stendhal Grenoble 3 - BP 25, 38 040 Grenoble Cedex 9 - France

{Alexia.Blanchard, Olivier.Kraif, Claude.Ponton}@u-grenoble3.fr

Abstract

This paper presents a "didactic triangulation" strategy to cope with the problem of reliability of NLP applications for Computer Assisted Language Learning (CALL) systems. It is based on the implementation of basic but well mastered NLP techniques, and put the emphasis on an adapted gearing between computable linguistic clues and didactic features of the evaluated activities. We claim that a correct balance between *false positives* (i.e. false error detection) - and *false negatives* (i.e. undetected errors) is not only an outcome of NLP techniques, but of an appropriate didactic integration of what NLP can do well - and what it cannot do. Based on this approach, ExoGen is a prototype for generating activities such as gapfill exercises. It integrates a module for error detection and description, which checks learners' answers against expected ones. Through the analysis of graphic, orthographic and morphosyntactic differences, it is able to diagnose problems like spelling errors, lexical mix-ups, errors prone agreement, conjugation errors, etc. The first evaluation of ExoGen outputs, based on the FRIDA learner corpus, has yielded very promising results, paving the way for the development of an efficient and general model adapted to a wide variety of activities.

Keywords: CALL, language learning, error diagnosis, error feedback

1. Introduction

In the field of CALL (Computer Assisted Language Learning), and especially for so-called 'structural' systems (systems for repeating and training, as opposed to exploration and reference systems, cf. Wyatt, 1987; Meunier, 2000), error detection and analysis constitute a central point to compute the adapted diagnoses and feedback that are required for an interactive, autonomous and customized learning. Yet, most of popular activity generators as Hot Potatoes, Course Builder, NetQuiz, etc. simply implement basic string comparison between a given learner answer, and the expected correct answer(s). With such tools, a single missing (or unexpected) capital may yield a fully incorrect evaluation. Rézeau's (2001, p.375) remark is still of topical interest: "(…) it can be noted that almost all language learning softwares on the market at the late 90's only propose exercises of the first type [i.e. that display few visible and constrained choices], with indeed a 'minimal' analysis."[1] (our translation). These approaches may just yield true/false feedbacks, without giving the learners the opportunity to differentiate between what is correct and what is wrong - or incomplete - in their knowledge of the tongue, and to correct their productions by themselves.

As demonstrated in Heift and Schulze (2007), by addressing various dimensions of language (lexicon, syntax, morphology, semantics, etc.), Natural Language Processing (NLP) should bring solutions to this problem. But, even if many NLP-based CALL systems have been studied and developed so far, vanilla applications in the field are still rare. In addition to the lack of communication between researchers and practitioners of didactics, linguistics and computer science[2], this disaffection of NLP in CALL may result from various reasons linked to the difficulty of adapting complex technologies that were not initially designed for a didactic purpose:

- **Free text analysis**: complete systems as *FreeText* (Granger, Vandeventer, Hamel, 2001) or *Correcteur 101 didactique* (by Machina Sapiens Corporation) aim at detecting and analysing errors inside learners' free productions[3]. Yet, as a participant to FreeText Project acknowledges himself, the system fails because of "too high error overdetecting rate" (L'haire, 2004, p.5; our translation). For the sake of comparison, we tested *Correcteur 101 didactique* on extracts of the FRIDA Corpus[4], which was used in the FreeText Project. Correcteur 101 seems to suffer the same drawback of overdetection. Here is an example of a learner production. Spotted errors are underlined:

"Dans tout le monde, il y a plus que plusieurs langues étrangères [étranger]. Ce pour sa [ça], qui parler [parle] ou connaître plusieurs de langues est [sont] nécessaire. Mais [mai ?], je crois qui

---

1 "Or, on constate que la quasi-totalité des didacticiels de langues sur le marché à la fin des années 1990 proposent des exercices du premier type, et donc une analyse de réponse que nous qualifierons de « minimale »."

2 We noted that among our colleagues, few language teachers were really aware of what NLP technology could - or could not - offer.

3 *FreeText* is a research prototype including didactic content and exercices that implement advanced NLP tools and *Correcteur 101 didactique* is a commercial grammar checker specifically designed for learners of French. Both implement error detection using syntactic parsing.

4 FRIDA (*FRench Interlanguage Database)* is a corpus of French as a Foreign Language compiled within the framework of the EU-funded FreeText project (Granger & al., 2001). The corpus contains texts (around 500.000 words) written by learners of French as a Foreign Language.

parler trois ou quattre langues sont sufficents, parce qui n'est pas possible étudier [étudie?] tout le [toutes les] langues."[5] (Proposed corrections are written between brackets)

Even though unknown misspelled forms are well detected (e.g. *quattre*, *sufficents*…), most of the messages about homophones frequently point out nonexistent and very unlikely errors. Only short-range agreement rules are correctly analysed and parsing errors lead to misleading corrections, as in **plusieurs langues étranger* (for which the learner production was correct). That kind of misdetection and miscorrection does not allow the use of such a tool for autonomous learning.

◆ **Controlled texts analysis:** in order to avoid the limitations of free text analysis, some approaches focus on the processing of controlled production, for which various parameters are previously fixed by the didactic context. For instance, the Alexia system (Selva, Chanier, 2000) uses a corpus of texts that concern a specific domain (i.e. employment and unemployment); the Eleonore system (Rénié, Chanier, 1993) only addresses interrogative utterances. In these cases, analysis outputs and feedback are far more precise and relevant than for free texts. But critics can be made about the high cost of developments, for a rather small benefit in term of automation. Moreover, the system architecture lack of generality, and is not easily reusable in other contexts and for other kind of activities.

To be more extensive, this quick insight of NLP should not ignore the various uses of spell checkers and grammar correctors in the classroom framework. Various studies (Cordier-Gauthier, Dion, 2003; Charnet, Panckhurst, 1998; Désilets, 1998) have shown that these applications - though they are not designed for didactic purposes - may be useful in this context: despite a limited quality in term of *overdetection* and *underdetection*[6], and a certain unsuitability of feedback, the given analyses may constitute an interesting starting point for the teacher to bootstrap learners reflection and awareness about some detected errors; but it is not adapted for an autonomous learning, given that false error detection may be very confusing for a learner.

In short, automated analyses of learner productions are facing two major problems: first, NLP applications suffer a lack of reliability which becomes problematic in a didactic context; second, as pointed out by Rézeau (2001), research and development projects usually involve important investments, and projects remain often at the prototype stage, and sometimes they do not go beyond specification. To cope with both these hindrances, we propose a "didactic triangulation" strategy, based on the idea that a precise specification of didactic context allows getting round the pure NLP problems, and solving some ambiguities. This position follows the still topical recommendation of Bar Hillel (1964, p.183), to make "a judicious and modest use of mechanical aids". This "didactic triangulation" strategy is presented in the next section. Section 3 and 4 focus on the architecture and implementation of ExoGen, a system that illustrates this approach, using very basic NLP techniques that brings interesting improvements to learner answers analysis. The results are evaluated in section 5, allowing us to sketch preliminary conclusions and prospects regarding our approach in section 6.

---

[5] Literally "In all the world, there are more than several stranger languages. That's for this, that to speak or to know several languages is necessary. But, I believe that to speak three or four languages are sufficient, because it is not possible to study all the languages." Note that all the literal translation given in this paper only aim at giving access to the sense of French words, but they neither represent learner errors, nor correct English utterances.

[6] Considering the output of a spell checker, we call "overdetection" the proportion of "false positives" against the correctly detected errors; we call "underdetection" the proportion of undetected errors, against the total number of really occurring errors.

2. The "didactic triangulation" strategy

The previous section stresses on the current weaknesses and limitations of error analysis. However, two conclusions can be drawn from the mentioned experiments. First, under restricted conditions and in a controlled didactic framework, some NLP analyses may be reliable enough: spelling error detection, morphological analysis, lemmatization, short range agreements, etc. The second conclusion may be drawn from the feedback of the teachers who try to use these systems:

*"*Correcteur 101 *is very interesting because it brings the learner to self-assessment of its own utterances and errors. The software rarely gives the answer. Two tools are also proposed: a dictionary and a grammar. They may appear very useful to help the learner out. The language that is used is not conformant to the New Grammar, and it would be an advantage to link it to the grammar notions that are studied in the classroom. Sometimes, some errors are not mentioned."(our translation[7])*

Error detection without correction presents a pedagogical advantage because it should bring the learners to be aware of their own mistakes, and to build their own correction strategy according to the constructivist methodology. An adapted feedback that gives aids, complementary explanations, examples, etc., may ease this reflection and help the learners in finding the required solutions. As the quality and correctness of these feedbacks are strongly linked to error detection and analysis, these two steps have to be mastered with a maximal reliability. For these reasons, we propose an empirical approach based on the incremental use of NLP technologies, starting from the simplest ones, for which the results can be more simply controlled by the didactic context. That's what we call the didactic triangulation strategy, based on the following statements:

- There are reliable NLP techniques that may yield relevant error detection and analysis clues: tokenisation, POS-tagging, lemmatization, language identification, morphological analysis, etc.
- Context independent analysis is currently out of scope. By "context", we intend the didactic context of learner production. As the above mentioned NLP techniques are not fully self-standing, and may face unsolvable ambiguities[8], these limitations may be balanced by a correct definition of contextual parameters. For instance, in the case of the ExoGen system presented below, the knowledge of a possible expected answer allows ambiguity resolution and analysis process guiding.
- A 100% reliability is, and may stay in the future, an unattainable goal. Therefore it is more realistic to stress on "assisted" rather than "fully automated" approaches. For instance, NLP may help teacher during authoring process, to find examples (Antoniadis, Kraif, Ponton, Zampa, 2007) and build activities (Kraif et al., 2004), and assist learner in self-assessment and self-correction.
- When a linguistic ambiguity cannot be solved, all the possible results have to be taken into account, in order to take the right decision further in the process, at didactic level.

---

[7] Remark made by a French teacher who uses *Correcteur 101 didactique* in the classroom (http://c-rdi.qc.ca/produits/aff_fiche.asp?fiche=426, retrieved on 01/31th/2007). Most notes posted on this site are consistent with this citation.

[8] Even for grammatically correct sentences, the best parsers for French, in the present state of the art, scarcely reach a precision around 80% (Paroubek et al., 2007).

- The tools have to be designed in a modular and declarative way, in order to be reusable in various contexts, and above all, they have to be accessible to teachers in order to let them customize and adapt them without being computer experts. According to us, that is the main challenge of the didactic triangulation approach: even simple techniques may require complex tuning and parameter definition, a correct didactic integration should aim at hiding this degree of complexity.

The implementation of this approach (cf. figure 1) is comparable to the adaptation by Anctil (2005) of the "problem resolution strategy" developed by Andre (1986).

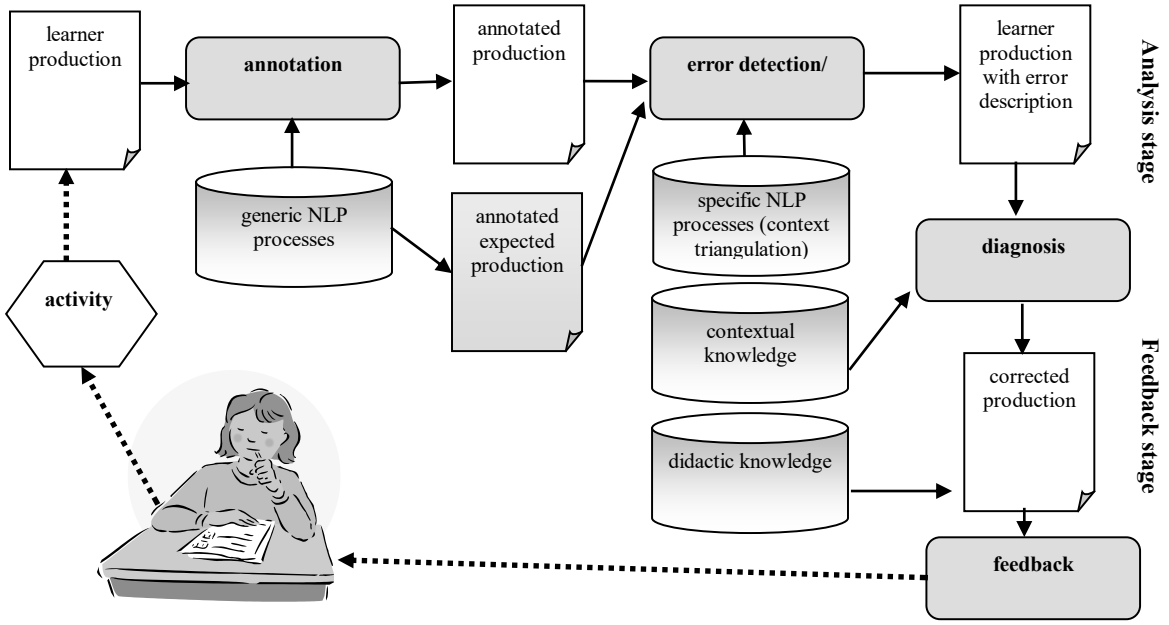


Figure 1: Didactic triangulation strategy implementation

The point consists in separating the problem analysis stage (i.e. the error analysis) from the resolution stage (i.e. the feedback production); the latter involves the design of an adapted feedback which requires further studies from a didactic point of view. The analysis stage involves three steps for detection, description and problem analysis. As Anctil, in our experiment, we have joined in the same stage detection process and error description. Based on a generic pre-processing of both learner production and expected correct production (such as tokenisation and POS-tagging), this stage consists in a fine-grained analysis of the differences between given answers (GA) and expected answers (EA). This analysis is disambiguated thanks to the triangulation of data coming from GA, EA and the context (see the description below for more details). This disambiguation involves specific NLP processes, adapted to the type of activity.

The problem analysis stage aim at identifying probable causes for identified errors, taking into account the complete knowledge about the production context: activity type and goals, typical errors, learner model, etc. This stage relies on the possibility of selecting reliable and relevant clues resulting from the previous stage. No complete diagnosis must be done if some clue is missing.

3. The ExoGen system

To validate our approach, we developed a prototype, called ExoGen, that allows generating new activities from corpora of POS-tagged and lemmatized texts (currently, we use corpora from the ConcQuest online concordancer[9]). For the moment, we propose only one type of activity that may include one or two stages: example reading, and gap-fill; but it could easily be extended to other forms of interactivity: drag and drop, classification, expression spotting, matching. The principle of generation is based on the random selection of sentences that match with specific meta-regular expressions patterns (Kraif, 2006). We call them "meta" because Perl-style regex patterns may be defined at two levels: first, we can look for sequences of characters in any feature born by a token (i.e. form, lemma, category, morpho-tags, and other additional token level tags); then, it is possible to define sequences of tokens using the same formalism[10]. For instance:

(1) <lemma=/(être|avoir)/> <cat=adv>* <tags=/ppart/>

matches with expressions containing the auxiliary verbs *être* or *avoir*, followed by any (possibly empty) sequence of adverbs, followed by a verb bearing the past participle feature. This meta-regex is designed to retrieve occurrences of *passé composé* and *plus-que-parfait* compound tenses (but also passive constructions).

(2) <lemma=/ir$/,lemma!=/(oir$|partir|venir$)/, cat=ver>

matches with verbs containing *-ir* ending (as *finir*) excluding verbs with *-oir* ending (as *pouvoir*) and irregular verbs as *venir, convenir*, and *partir*.

These patterns allow identification, in context, of compounds, idioms, phraseology and even syntactic constructions: for instance noun-verb collocations, present perfect progressive, irregular past participle forms, etc. Generated gap-fill activities are similar to those proposed by the Graz 1998 ECML workshop[11] or by Johns (1993), following the principles of Data Driven Learning, focusing on the following points: EXPLOITATION OF AUTHENTIC MATERIAL; EXPLORATORY TASK AND ACTIVITIES; LEARNER-CENTRED ACTIVITIES; EXPLOITATION OF TOOLS RATHER THAN READYMADE LEARNWARES (RÜSCHOFF, 2005, P.63). For instance, a generated activity may lists concordances where pronouns *which* and *who* appear alternatively. This part of the activity is what we call example-reading section. In the gap-fill section, the learner has to determine the correct pronoun and fill the gap, taking into account the meaning and syntactic properties of the context. The previously given examples help to solve the problem by inductive reasoning and analogy.

---

[9] The current corpora are not really designed for didactic purposes: they are multilingual parallel text collections available with the ConcQuest Concordancer (http://w3.u-grenoble3.fr/kraif/ConcQuest/concquest.php). Registered users may upload their own corpora.

[10] The meta-regular expression formalism has been implemented in PDC prolog in the ConcQuest search engine, freely downloadable from http://w3.u-grenoble3.fr/kraif/ConcQuest/concquest.php.

[11] See : http://www.ecml.at/projects/Voll/our_resources/graz_2001/data_driven_learning/bernd/index.htm

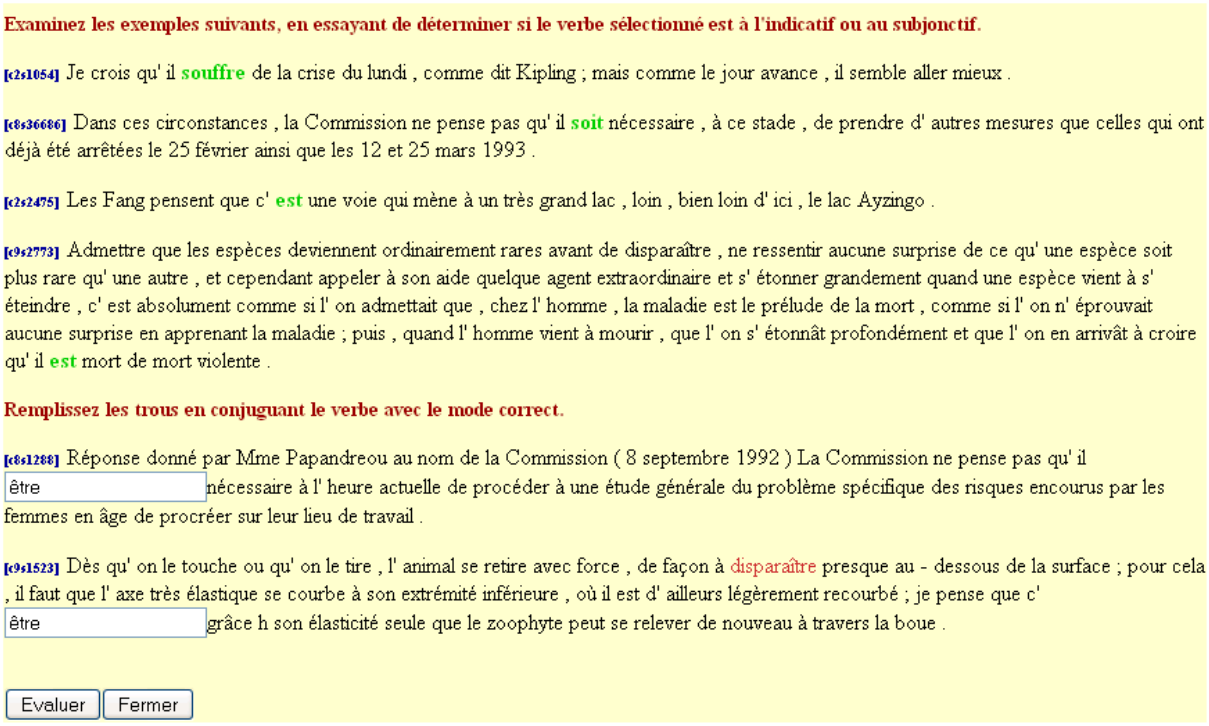


Figure 2: An activity generated by ExoGen

For the purpose of the present study, this simple model of activity is a good benchmark to evaluate the didactic triangulation strategy:

- As in Alfalex system[12], the activity are generated on-the-fly, and the complete range of correct answers cannot be defined previously in a manual way;
- Answers are short, and rather simple to process, but they may present a wide scope of ambiguities for NLP standard functions;
- Didactic context may restrict the possible interpretations of learner answers in to ways: 1/ Gaps are selected upon formal criteria. An adapted choice of these criteria, driven by didactic constraints, may reduce potential ambiguities. For instance, in French the subjunctive form is ambiguous, and may be confused with indicative for a large class of regular verbs. In an activity which focuses on the use of subjunctive, these ambiguous forms may be discarded during the gap selection (using <base!=/er!/>). This kind of "preventive" disambiguation is generally made for didactic reasons when designing an exercise, and it brings formal disambiguation during the NLP analysis. 2/ The Expected Answer EA (the expression that has been removed from the gap), as a part of the didactic context, may bring additional information to disambiguate the Given Answer (GA). Section 4 gives more details about the kind of disambiguation yielded by comparing GA and EA.

In this framework, the problem analysis is based on two different stages: 1/ difference analysis, and 2/ diagnosis.

12 http://www.kuleuven.be/alfalex/index.php?id=&ng=0

Difference analysis

The first stage aims at describing *differences* between GA and EA. These differences may affect various and possibly independent linguistic aspects: spelling, inflections and morphology, syntax, vocabulary, meaning. This stage is rather non-specific, and does not depend on the didactic context. Generic NLP tools may be implemented: POS-tagger, stemmer, wordnets, etc.

A very simple and powerful heuristic may be used to solve ambiguities during this analysis: when comparing GA and EA, one can assume that similarities are not fortuitous. We call it the *lesser difference heuristic*. Thus, if GA may be analysed in *n* different ways $\mathcal{A}_1$, $\mathcal{A}_2$,… $\mathcal{A}_n$ and EA in *p* ways $\mathcal{A}'_1$, $\mathcal{A}'_2$,… $\mathcal{A}'_p$, we can compare every possible pairs, and keep the more similar one ($\mathcal{A}_i$, $\mathcal{A}'_j$). The similarity measure can be computed from the common morphosyntactic features between EA and GA. Let's consider the following example:

il effectue/fait une opération, EA=effectue, GA=fait
(literally : he carries out/does an operation)

effectue is a verbal form bearing indicative or conjunctive feature, and *fait* may be an indicative verb or a past participle, or a noun. Feature comparison, using this heuristic, help to disambiguate both EA and GA as indicative verbs. Section 4 gives a precise evaluation of the results obtained with this heuristic.

Diagnosis

The diagnosis stage is closely related to the didactic context, because it depends on the characteristics of the activity: instructions, level, aids, linguistic context, etc. The diagnosis step aims at finding out the probable causes of learner's errors, in order to give him an appropriate feedback. It is a rather complex task that should involve, ideally, a model of the learner. In a more modest way, diagnosis may just imply the determination of typical errors, which are made by *most* learners, and for which general causes may be invoked.

To cope with NLP limitations at this stage, it is possible to rely on two simple principles:

- *Vagueness*: whenever NLP processes cannot solve ambiguities, it is safer to keep a partial analysis than to complete it at all costs. By the way, unsolved ambiguities may not be a hindrance to determine a correct feedback, which may be more or less general: it may be interesting to tell the learner precisely what error has been done, but when the system cannot draw a conclusion, a simple recall of the rules that apply in the context may fit as well.

- *Triangulation*: the analysis may take advantage of different sources of information, in order to corroborate some hypotheses. In a classic NLP context, the analyser is limited to the given answer GA, knowing its linguistic context *CO(li)*, as shown by the following diagram:

$$\text{NLP Analysis} \equiv \frac{GA}{CO(li)}$$

  For the present case, the expected answer EA and the didactic context *CO(di)* may bring additional disambiguation clues. Hypotheses resulting from different sources of information (syntactic, didactic, typical errors, etc.) may be triangulated in this way:

$$\text{Triangulated Analysis} \equiv \frac{GA}{CO(li) + CO(di) + EA}$$

Let's take the example of an activity about the past participle agreement in French, using the "passé composé" tense, with *avoir* auxiliary. To determine the correct agreement, one has to answer various questions: is there a direct object expressed before the verb, using a clitic or a relative pronoun? If so, what is the gender and number of this object? Such questions are difficult to answer using parsing technique. As shown in the following example, a parser has to determine (1) the gender/number of the relative pronoun antecedent (2) whether the verb phrase is a factitive construction (3) whether the relative pronoun has as a direct object function (4) the anaphoric antecedent of the clitic pronoun:

*... cette publication que nous avons **faite** (1) [literally: this publication that we have done]*
*... cette publication que nous avons **fait** imprimer (2) [literally: this publication that we get printed]*
*c'est pour cette publication que nous avons **fait** cela (3) [literally: it is for this publication that we did that]*
*cette fête, c'est pour cette publication que nous l'avons **faite** (4) [this party, it is for this publication that we did it]*

In the case of error diagnosis, the triangulated analysis may be far more straightforward, considering only two cases.

a) In cases 1 and 4, we do not need syntactic analysis to determine that there is an object preceding the verb, because the expected answer (in bold) has inflectional marks of feminine or plural. If a learner puts *faits* instead of *faite*, he probably makes an improper agreement between the subject and the past participle: on the FRIDA learner corpus, we have observed that most occurrences of agreement errors are due to an improper agreement with the subject. If a learner put *fait* instead of *faite*, he probably forgets to make an agreement between object and participle. Whatever error is done, a simple feedback is adapted: "As the object precedes the past participle, the participle must agree in gender and number with the object."

b) For the case 2 and 3, we cannot know without parsing if the object is before. So, if the learner answers *faits* or *faites*, another adapted feedback would recall other aspects of the rule: "the past participle must not agree with the subject when used with *avoir* auxiliary, and must not agree with the object when occurring after."

Of course, a parser would be useful to draw more precise hypotheses and discriminate between sub-cases of a) and b). The different steps of such analysis may be formalized as a decision tree. The clues which drive the analysis are hierarchically ordered, in order to process first surface indices that are always available and rather unambiguous. More complex clues are processed later when going closer to the leaves. Ambiguous cases (where it is not possible to answer *yes* or *no*) may stop the analysis at an intermediate node, or be processed in a specific branch of the tree. At each node and leaf of the tree can be attached an adapted feedback.

The example detailed on figure 3 shows how the agreement of past participle may be analysed. Note that the various leaves (round boxes) correspond to frequent error cases that have been empirically observed in the FRIDA corpus.

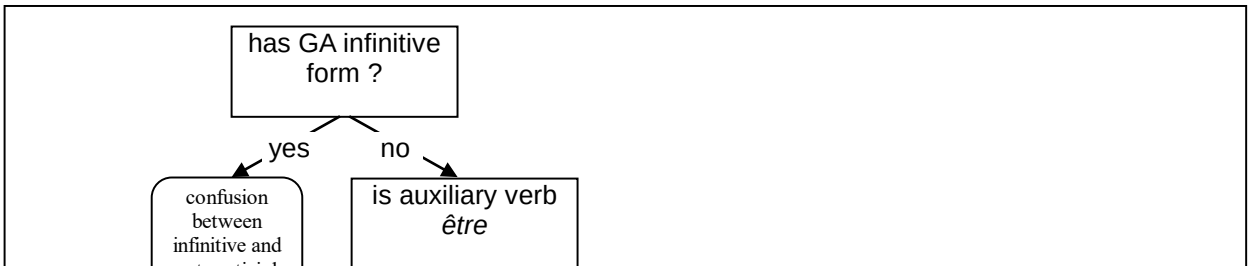


Figure 3: Decision tree for didactic analysis (Case of agreement of past participle)

4. Implementation

So far, only the first stage of difference analysis has been implemented. Anyway it is a good beginning to test whether triangulation, limited to the comparison between a given answer and an expected answer in the framework of a simple gap-filling activity, may allow the preparation of a reliable diagnosis. In order to assess the disambiguation potential of the *lesser difference* heuristic, we have tested it on non-disambiguated data. We have not used POS-tagging, lemmatizing or even parsing, because we just wanted to illustrate how the triangulated analysis could cope with linguistic ambiguity (but we do not intend to advocate a resource-poor approach of NLP). We have just used a simple tokenizer written in Perl, that we have implemented, and an inflected form dictionary available online[13]. Each entry of this dictionary consists in a simple inflected form, associated with a lemma, and possible combinations of morphosyntactic features (part-of-speech, number, gender, person, tense, mode, etc.). A sample of the dictionary records is given in figure 4.

glace	glacer	Ver:IPre+SG+P1:IPre+SG+P3:SPre+SG+P1:SPre+SG+P3:ImPre+SG+P2
glacé	glacer	Ver:PPas+Mas+SG
glacent	glacer	Ver:IPre+PL+P3:SPre+PL+P3
glacera	glacer	Ver:IFut+SG+P3
glaceraient	glacer	Ver:CPre+PL+P3

Figure 4: A sample of the inflected forms dictionary

[13] The ABU inflected form dictionary can be downloaded from : http://abu.cnam.fr/

The analysis process is driven according a hierarchical organisation of differences between EA and GA: surface similarities are processed first, because they require less computation and give safer clues. Moreover, if we take into account every linguistic aspect (morphology, syntax, meaning), these similarities involve a "smaller" subset of features, which is coherent with the lesser difference heuristic.

1. Graphic differences.

The first level concerns minor variations on surface form: spacing, case, character variants (e.g. oe *vs* œ). Usually such differences between GA and EA yield a positive feedback.

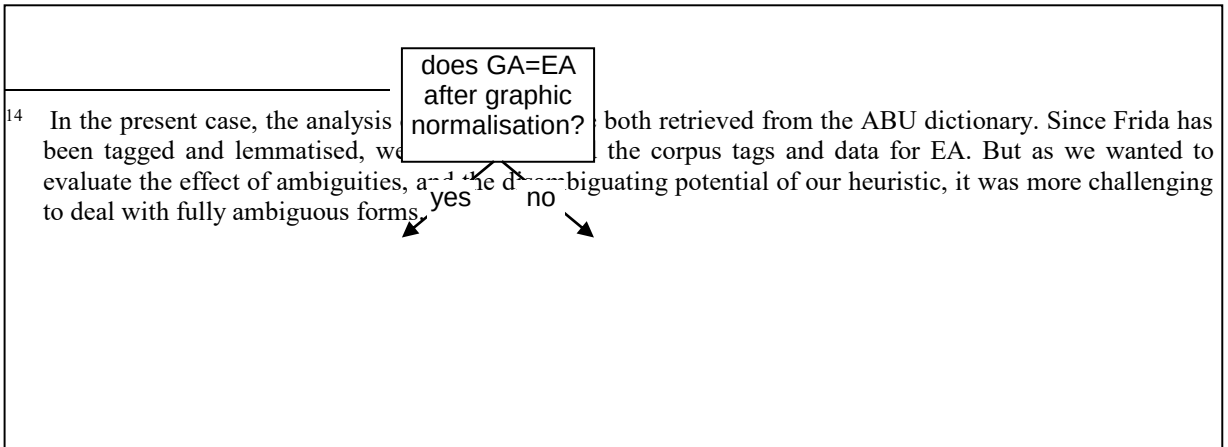
2. Spelling differences. If GA does not appear in the dictionary

a) GA and EA are similar
   i/ only diacritic signs differ (accent, cedilla, etc.)
   ii/ there is some other spelling errors
b) GA and EA are not similar
   i/ there are some neighbours (similar forms) in the dictionary
   ii/ there are no neighbour.

Surface similarity may be computed using the Levenshtein function. Here, we have used a similar method, the longest common substring computation, that we previously implemented to detect cognate word pairs successfully (Kraif, 2001). It is not difficult to see how each case of this analysis may result in a specific feedback, such as "check the accentuation", or "did you mean one of these words?", "spelling error", etc.

3. Morphosyntactic differences. When GA is found in the dictionary, it is possible to confront all its potential analyses with the analyses of EA[14]. This comparison is driven by the lesser difference heuristic, by asking, in hierarchical order, the following questions: does GA and EA correspond to a same lemma? Do they belong to a same part-of-speech? Do they share same morphosyntactic features? For known forms, all these answers are given by the dictionary.

The implemented difference analysis is illustrated by the decision tree displayed on figure 5.



[14] In the present case, the analysis … both retrieved from the ABU dictionary. Since Frida has been tagged and lemmatised, we … the corpus tags and data for EA. But as we wanted to evaluate the effect of ambiguities, and the disambiguating potential of our heuristic, it was more challenging to deal with fully ambiguous forms.

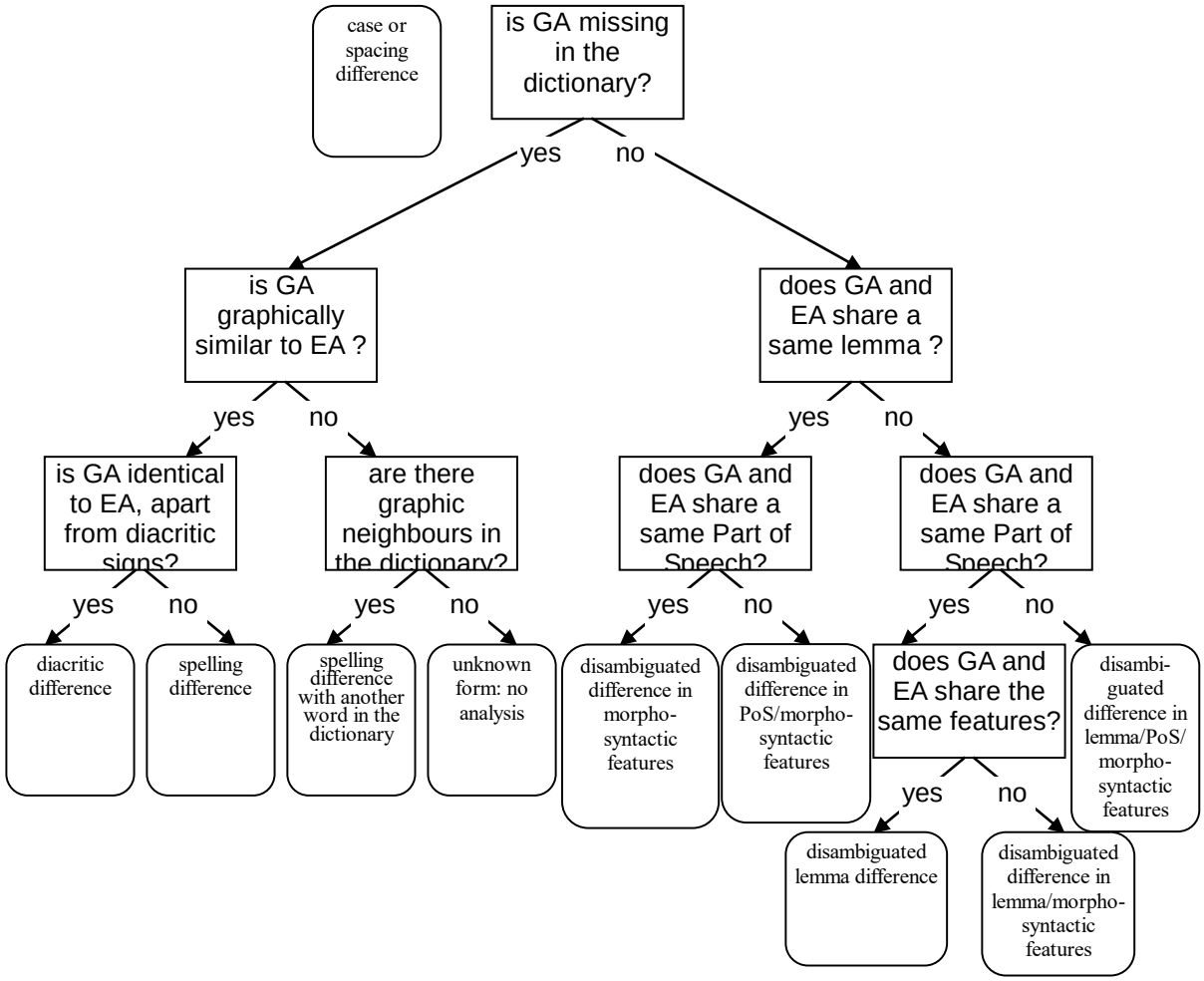


Figure 5: The first stages of difference analysis

Evaluation

The evaluation of our analysis and disambiguation method requires a learner corpus created in the framework of short-answer open response questions such as quiz or gap-fill exercises. For each given answer, we could apply error analysis based on the expected answer. Although the design of such a corpus is planned in future development of ExoGen, we do not yet have such empirical data. Accordingly, for this evaluation we have used another resource, extracted from the FRIDA corpus. It is made up of three subcorpora of similar size which contain data from English-speaking learners, Dutch-speaking learners and learners from mixed mother tongue backgrounds. All the texts had been manually error-tagged according to a three levels typology: error domain (morphology, grammar, lexicon, etc.), error category (agglutination, gender, spelling, etc.) and part-of-speech. For each identified error, annotators had indicated a correction. This corpus allows us to extract erroneous form/corrected form couples comparable to the GA/EA couples obtained in short-answer open response questions.

> *(...) une seule monnaie (l'ECU) n'**adresse** pas bien au gouvernement anglais. [literally: only one currency (the ECU) do not address well to English government]*
> Erroneous form: *adresse,* Corrected form: *convient*

This evaluation may include a bias because the relationship between GA and EA is not identical to the relationship between erroneous answer and corrected answer in our learner corpus. In an activity such as gap-filling, both expected answer and its context pre-exist to the given answer, while in Frida the corrected answer is given subsequently, on the basis of the error and its context inside the free production. However, we believe that this bias is limited from the point of view of the difference analysis, because the same kind of difference is observed, and the analyser is confronted with the same type of ambiguities (lemma, part-of-speech, features, meaning).

We used a sample of 47 productions of English-speaking learners with variable levels. We selected all errors involving two simple forms (because of limitations in our dictionary), excluding punctuation, for a total of 318 cases of errors. For each error we applied difference analysis and we obtained descriptions corresponding to 16 possible cases with precisions about lemma, part-of-speech and features.

Examples of output:

Example of error	Description (automatically processed)
(…) avant de **retourner** *(arriver)* en Angleterre. [literally: before coming back (arriving) in England]	Forme grammaticalement correcte (verbe infinitif), mais on attendait une autre forme. [Grammatically correct form (infinitive verb), but another form was expected]
et beaucoup d'**échafaide** *(échafaudages)*. [literally: and lot of scaffoldings]	Orthographe erronée ou mot inconnu du dictionnaire. [Wrong spelling or unknown word]
Je dois me **dépécher** *(dépêcher)*. [literally: I have to hurry up]	Orthographe erronée : problème d'accent. [Wrong spelling: problem with accent]
(…) sommes bien amusées et c'est **vrai** *(juste)* de dire que nous avons dansé assez bien [literally: we had fun an it is true (right) to say that we dance rather well]	Forme grammaticalement correcte (adjectif ou adverbe ou nom masculin singulier), mais on attendait une autre forme [Grammatically correct form (adjective or adverb or masculine singular), but another form was expected]
C'était **désespéré** *(désespérant)* mais c'était la seule chance (…) [literally: It was despaired (despairing) but it was the only chance]	S'il s'agit du verbe *désespérer* : cas 1 [masculin singulier] : On attend un participe présent et non un participe passé. [If it is the verb *despair* (masculine singular) : present participle is expected instead of past participle]
Pour moi **l'** *(cette)* image crée une ambiance délassante [literally: For me the (this) image creates a relaxing atmosphere]	Forme grammaticalement correcte sur le plan de la catégorie (déterminant), mais on attendait une autre forme avec d'autres traits. [Grammatically correct part-of-speech (determiner), but another form was expected with other features.]
le Premier ministre reste toujours un **britannique** *(Britannique)* [literally: the Prime minister remain anyway a british man]	Exact, mais il faut une majuscule à l'initiale. [Correct, but the initial letter should be a capital]

Table 1: Examples of errors (corrections between parentheses) and corresponding descriptions

One notes that in some cases disambiguation is partial, however a relevant description can be given. For a quantitative assessment of results, we manually evaluated correcting statements related to various analyses. In addition, we observed for all cases where forms (erroneous and corrected) encompassed ambiguities (multiple analyses), if disambiguation is full, partial or null (see table 2).

Case	**Every cases**	non ambiguous	fully disambiguated	partially disambiguated.	not disambiguated.
Wrong/Correct	**6 / 312**	1 /187	5 /104	0/14	0/7
Precision	**0,981**	0,995	0,954	1	1

Table 2: Evaluation of the correction of error descriptions
Precision = Correct / (Correct+Wrong)

One notes that precision, which expresses the proportion of correct analyses, is very satisfactory. The disambiguation heuristic, effective on 1/3 of cases, very often leads to a full

disambiguation with less than 5% wrong. In many cases, heuristic yields a spectacular reduction of ambiguities:

> *une seul monnaie (l'ECU) n' **adresse** (convient) pas bien au gouvernement anglais.*

In this example, *adresse* may correspond to two different lemmas (*adresse* and *adresser*), to two different parts of speech (noun and verb) and to several features (*Nom:Fem+SG, Ver:IPre+SG+P1, IPre+SG+P3, SPre+SG+P1, SPre+SG+P3, ImPre+SG+P2*). The comparison with *convient* permits to keep the only common representation: verb, present indicative tense, third person singular (*Ver:IPre+SG+P3*).

Concerning the erroneous analyses, they are due to two phenomena:

- Dictionary lack (2 cases): in the following example, *futur* is not recorded as a potential noun but only as an adjective.

> *le **futur** (avenir) de l'Angleterre [literally: the future of England] -> "On attendait une autre forme, d'une autre catégorie grammaticale (Nom # Adjectif)." [Another form was expected, with another part-of-speech]*

- Wrong disambiguation (4 cases): in the following example, the corrected form is interpreted as the determinant *tous* and not as a pronoun:

> *l'heure de se joindre et de parler **tout** (tous) d'une voix [literally: time to join each other and to speak every (all) together] -> "S'il s'agit du déterminant **tout** on a : cas 1 [Masculin] : On attend un pluriel et non un singulier." [If it is the determiner* every *(masculine) : plural is expected instead of singular]*

Note that even if disambiguation is wrong, the feedback given to the learner can present an analysis as hypothetical, in order to avoid a state-against truth. In addition, some ambiguities can be reduced by selecting the information sent to the user. Let's consider the following example:

> *Soudain, nous avons **entendus** (entendu) un bruit [literally: Suddenly, we heard a noise] -> "S'il s'agit du verbe **entendre** [participe passé masculin], on attend un singulier et non un pluriel ; S'il s'agit de l'adjectif **entendu** [masculin] on attend un singulier et non un pluriel ; s'il s'agit du nom **entendu**[15] [masculin] on attend un singulier et non un pluriel." [If it is the verb 'to hear' [past participle masculine], singular is expected instead of plural; if it is the adjective 'heard' [masculine], singular is expected instead of plural; if it is the noun 'innuendo' [masculine], singular is expected instead of plural]*

The result is ambiguous (verb, adjective or noun) but the analysis of the features is always the same, and the following feedback may be produced, as implemented in the actual version of ExoGen: "singular is expected and not plural." It is possible to satisfy oneself with this information, incomplete but reliable, focusing on the error committed by the learner.

[15] Note that our fullform dictionary records erroneously *entendu* as a noun, because it is a part of the noun *sous-entendu.)*

5. Conclusion and prospects

We have presented a general framework for learner answer analysis, based on the comparison between the given answer and the expected answer in a particular didactic context. We propose to cope with the lack of reliability of NLP by a correct didactic integration of generic techniques such as tokenisation, POS tagging and lemmatizing, morphological analysis, etc. A specific analysis stage, specially designed for a given activity, can take advantage of these low-cost generic processes, and disambiguate their results by taking into account contextual information such as activity instructions, expected answer(s) and activity type. We call this approach: the "didactic triangulation strategy". Spelling errors, lexical confusions, agreement problems and improper conjugations easily fall in the scope of this strategy.

To illustrate and partially validate this approach, we have implemented a very simple method of answer analysis in the context of a gap-fill exercise. The triangulation strategy has allowed the development of a disambiguation heuristic based on the confrontation of given answer against expected answer. The results are encouraging, with a precision in the error description higher than 98%. Such analysis could go further, doing comparison on semantic features: when GA and EA correspond to distinct lemmas that belong to the same part-of-speech, it may be interesting to look at the semantic similarities between them: they may share some senses, and be linked by synonymic, hyponymic or other semantic relationship. Register may be another interesting dimension for this comparison: when comparing GA=*job* and EA=*work*, the system should identify that familiar register was not expected in the correct answer... We plan to use a French wordnet to complete our analyser on these aspects.

The following stage will consist in developing rules for diagnosis, in order to determine probable causes of errors (for instance, in French, an error-prone agreement of past participle with the subject in the context of *avoir* auxiliary). To make a generic system, adaptable to a wide range of activities, it is important to define a simple and declarative language to express these rules. According to us, it is a real challenge, essential to allow teachers to define themselves the content and the goals of diagnosis, in order to prepare adapted automatic retroactions.

For this purpose, we plan to develop first finer disambiguation techniques, based on Given Answer / Expected Answer / linguistic context triangulation. This NLP module, going further than standard techniques such as POS-tagging, should be relatively generic and autonomous, in order to be applied to various activities of CALL.

References

Anctil D. (2005). *Maîtrise du lexique chez les étudiants universitaires : typologie des problèmes lexicaux et analyse des stratégies de résolution de problèmes lexicaux*. Mémoire de M.A. Faculté des Sciences de l'Education. University of Montréal (Québec).

Andre T. (1986). *Problem solving and education*. San Diego, CA: Academic Press.

Antoniadis G., Kraif O., Ponton C., Zampa V. (2007) (in press). Un outil exploratoire de corpus d'apprenants. *Proceedings of UNTELE'07*. University of Compiègne (France). 29-31 mars 2007.

Bar-Hillel, Y. (1964) The future of Machine Translation. *Language and Information : Selected Essays on their Theory and Application*. London, Addison-Wesley. 180-184.

Charnet C., Panckhurst R. (1998). Le correcteur grammatical : un auxiliaire efficace pour l'enseignant ? Quelques éléments de réflexion. *ALSIC,* 1 (2). 103-114.

Cordier-Gauthier C., Dion C. (2003). Correction et révision de l'écrit en français langue seconde : médiation humaine, médiation informatique, *ALSIC,* 6 (1). 29-43.

Désilets M. (1998). Que penser de l'utilisation des logiciels correcteurs à l'école? *Vie pédagogique*, 107. 9-11.

Granger, S., Vandeventer, A., Hamel, M.-J. (2001). Analyse des corpus d'apprenants pour l'ELAO basé sur le TAL. *TAL*, 42 (2). 609-621.

Heift T., Schulze M. (2007). *Errors and Intelligence in Computer-Assisted Language Learning: Parsers and Pedagogues*. Routledge.

Johns T. (1993). Data-driven learning: an Update, *TELL&CALL* 1993/2. 4-10.

Kraif O. (2006). Extraction automatique de lexique bilingue : application pour la recherche d'exemples en lexicographie. *Journées du CRTT, Université Lyon 2, Lyon (France).*

Kraif O., Antoniadis G., Echinard S., Loiseau M., Lebarbé T., Ponton C. (2004). NLP Tools for CALL: the Simpler, the Better. *Proceedings of InSTIL/ICALL 2004 Symposium, nlp and speech technologies in advanced language learning systems*. 37-40

Kraif O. (2001). Exploitation des cognats dans les systèmes d'alignement bi-textuel : Architecture et évaluation. *TAL*, 42 (3). 833-867.

L'Haire S. (2004). Vers un feed-back plus intelligent, les enseignements du projet Freetext. *Proceedings of TALAL*. 1-12. [Towards a more intelligent feedback: what have been learned from the Freetext Project]. Retrieved May, 28, 2008, from http://w3.u-grenoble3.fr/lidilem/talal/actes/JourneeTALAL-041022-lhaire.pdf

Meunier L. E. (2000). La typologie des intelligences humaine et artificielle : complexité pédagogique de l'enseignement des langues étrangères dans un environnement multimédia. In L. Duquette et M. Laurier (Eds*). Apprendre une langue dans un environnement multimédia*. Les éditions Logiques: Outremont (Québec). 211-253.

Paroubek P., Vilnat A., Robba I., Ayache C. (2007). Les résultats de la campagne, EASY d'évaluation des analyseurs syntaxiques du français. *Actes des Ateliers de la 14e, Conférence annuelle sur le Traitement Automatique des Langues Naturelles (TALN),* France, vol. 2. 242-252.

Rénié D., Chanier T. (1993). La modélisation de l'acquisition, une étape dans la construction de systèmes d'EIAO des langues: le cas des interrogatives en français langue seconde. In M. Baron, R. Gras, J.F. Nicaud (Eds). *Environnements Interactifs d'Apprentissage avec Ordinateur*. Tome 1. Eyrolles: Paris, 123-134.

Rézeau J. (2001). *Médiatisation et médiation pédagogique dans un environnement multimédia. Le cas de l'apprentissage de l'anglais en Histoire de l'art à l'université*. PhD Thesis. University of Bordeaux II (France).

Rüschoff B. (2005) DATA-DRIVEN LEARNING (DDL): THE IDEA. IN FITZPATRICK T., LUND A., MORO B., RÜSCHOFF B. (EDS). *INFORMATION AND COMMUNICATION TECHNOLOGIES IN VOCATIONALLY ORIENTED LANGUAGE LEARNING*, Council of Europe (retrieved on 30th July

2008: http://www.ecml.at/documents/pub131aE2003_Fitzpatrick_withoutBookmarksAndCover.pdf)

Selva, T., Chanier, T. (2000). Génération automatique d'activités Lexicales dans le système ALEXIA. *Sciences et Techniques Educatives*, *(STE)*, 7 (2). 385-412.

Wyatt D. H. (1987). Applying pedagogical principles to CALL courseware development. *Modern Media in Foreign Language Education*. Wm. Flint Smith (dir.). Lincolnwood, IL : National Textbook. 85-9