COMBINED EVALUATION OF A VIRTUAL LEARNING ENVIRONMENT: USE OF QUALITATIVE METHODS AND LOG INTERPRETATION TO EVALUATE A COMPUTER MEDIATED LANGUAGE COURSE

Mariona Estrada, Raquel Navarro-Prieto, Martí Quixal
Barcelona Media
Barcelona, Spain
mariona.estrada@barcelonamedia.org, raquel.navarro@barcelonamedia.org, marti.quixal@barcelonamedia.org.

Abstract

In this paper we are going to present the process of evaluation of AutoLearn a blended language learning platform. The data collected through this evaluation in real instruction settings allow us to discuss the usefulness and suitability of a combined methodology to evaluate language learning platforms. The evaluation is mainly focused in the usability of the platform but taking into account also some didactic and linguistic aspects concerning the language learning process. With the aim of gathering all this data we are carrying out an evaluation combining exploratory methods, such as questionnaires and observations, and the analysis of the user navigation through the platform (logs). In these sense, the objective of our work is test the usefulness of this combined evaluation framework in AutoLearn, and, subsequently, develop a methodology for the analysis and evaluation of computer mediated language learning environments in general.

Keywords - VLE evaluation, blended language learning, quantitative-qualitative methods, log analysis.

1 INTRODUCTION

In this paper we present the process of evaluation of AutoLearn, a language learning platform suited both for blended learning and distance learning (available at http://parles.upf.edu/autolearn). The data collected through a series of testing actions in real teaching/learning environments allow for a discussion of the usefulness and suitability of a methodology combining qualitative and quantitative techniques to evaluate language learning platforms. The evaluation is mainly focused in the usability of the platform but taking into account also some didactic and linguistic aspects concerning the language learning process.

AutoLearn, funded by the EU under the Lifelong Learning Program as the project “AUtomatic tTuTor for lifelong language LEARNing” (AUTOLEARN), is a project conceived to define, carry out and evaluate an ICT-based paradigm for Foreign Language Teaching/Learning targeting at different educational levels. The main goal of the project is to integrate in a Virtual Learning Environment a set of didactic materials designed for training learners of English and German in skills and topics relevant for the Business and Finance domain. The learning approach, which is task-oriented, specifically addresses the use of Natural Language Processing (NLP) techniques to automatically generated feedback. Therefore, learning units are conceived from the very beginning to facilitate the use of computational techniques to develop a virtual tutor.

AutoLearn’s consortium consists of a set of partners with complementary expertise: researchers in Human-Computer Interaction (HCI), Foreign Language Teaching and Second Language Acquisition, and Natural Language Processing (NLP). The interaction between these three disciplines allows for an evaluation methodology that takes into account the perspective of the user, the learner, the teacher and also the overall performance of the NLP-based automatic correction module.

AutoLearn builds upon previous work developed in several national and international projects (MeLLANGE (F/04/B/P/LA-151212, http://mellange.ilia.jussieu.fr/), Uni-Deutsch project, ProGram-iK, ARQUÍTEXT (HUM2004- 05321-C02-02). The most relevant one is ALLES (IST-2001-34246, [1], [2]), a project whose aim was to show the feasibility to create more intelligent and individualised automatic correction modules on the basis of state-of-the-art linguistic technology. From a pedagogical point of
view, ALLES learning materials aimed at fostering the development of language in use rather than language as a formal system of signs and rules.

Interestingly, the communicative setting made up in a learning environment provides a naturally restricted domain for those NLP tasks to be performed beyond the morphology and syntax levels. AutoLearn uses part of the techniques previously developed and tested and extends their application to a wider range of target users, since ALLES was limited to university level learners of a foreign language.

In this paper we are going to present a summary of the evaluation of AutoLearn in one of the learning institutions where it was tested. Section 2 explains the methodological approach to the evaluation; Section 3 describes the evaluation setting(s) and its participants; Section 4 presents the results and the analysis obtained; finally, Section 5, presents some preliminary conclusions of work and some discussion and future work.

1.1 A brief description of AutoLearn’s didactic contents

AutoLearn currently consists of two English courses and two German courses. For each language one of the courses is targeted at C1 learners [3] and the other is targeted at B2 learners. Each of these courses consists of a range of units (between 5 and 6), all of which are conceived as enabling tasks leading to a final communicative task. The courses offered are:

- German, intermediate (B2): Weiterbildung [Education and training]
- German, advanced (C1): Geschäftsberichte [Business reports]
- English, intermediate (B2): Customer service and international communication
- English, advanced (C1): Business reports

Learning units include skill-oriented activities comprising all communicative skills except for speaking. As for the kind of automatic feedback, AutoLearn includes two different types of exercises: exercises with simple correction strategies (character and string matching) developed using the Hot Potatoes software, and exercises corrected with Natural Language Processing tools (fully automatic linguistic analysis at the level of word, morphology, syntax and content). The former are referred to as Hot Potatoes exercises and the latter are referred to as AutoTutor exercises.

2 METHODOLOGICAL APPROACH

According to Ardito et al. [4] the techniques most commonly used to evaluate usability derive from heuristic evaluation; other techniques such as field studies, observation, questionnaires, interviews or system logs have also been used, although less frequently. When evaluating ICT-enhanced learning platforms some authors [5] proven for the need to include end-user experience, not only the experience of the usability expert. From this point of view, Hartson et al. [(6)] state that in educational environments “because usability is ultimately determined by the end user, not an expert evaluator, realness of problems needs to be established by the user” (2003:159). In accord with these authors, we decide to focus our evaluation on real users in real environments not only to evaluate the usability aspects but the other aspect related to the learning context.

In the selection of the data collection methodologies to be used during the evaluation we also took into account the specific goals of AutoLearn, namely carry out and evaluate an ICT-based paradigm for foreign language learning. Following the work of Enric Mor [7], we decided to collect and add to the evaluation the navigation through the platform (logs) with the aim of observing the learner’s behaviour when they were doing their exercises. In spite of the proven usefulness of data mining in many areas of HCI [8], this method has been rarely applied to e-learning. However, we think that combined with qualitative techniques, such as questionnaires and interviews with users, can help us to get very useful information to improve and analyse the gain and limitations of using a language learning platform such as AutoLearn. Recently [7] has validated that need for this combined approach and conducted a study to show its usefulness in the area of virtual learning platforms.

In our view, the need to investigate using this methodological approach arises from the fact that we must be able to look beyond the final grades/marks obtained by learners in a course. In doing so we want to be able to investigate aspects related with the ease of use and the usefulness of learning platforms in real learning settings, as well as on the consequences that this might have on certain pedagogical issues.
2.1 Goals of the evaluation and data collection methods

Our evaluation procedure consists of three phases: definition of the evaluation criteria [9], selection and description of the testing actions (place, course, learning approach, etc.), and execution of the testing actions in real learning environments.

In the first step we decided to focus on three dimensions of the evaluation procedure: HCI and usability aspects, user satisfaction, and pedagogical and linguistic aspects. Thus, the main objectives of the evaluation have been:

- To evaluate the usability of the platform. Following the work [10], our usability evaluation will focus on the presentation, the application activity and the different multimedia issues to evaluate the HCI aspects. With this goal we wanted to assess the efficiency and effectiveness of several of the platform resources (course guides, exercises, and activities with automatic feedback) as well as the usefulness and supportiveness for learning and teaching.

- To evaluate the user satisfaction with the platform and the language courses. With this goal we wanted to assess the end user's perceptions and valuations as well as their needs and suggestions to improve AutoLearn. The user's satisfaction in connexion with all the aspects of the platform allows us to know the real needs and problems of end users.

- To evaluate the pedagogical and linguistic aspects of the courses. Through this dimension we wanted to evaluate if the user's learning gain was adequate all through their learning experience with the platform, and if the language and the metalanguage used were suitable for the learning/training purposes.

As we have said before, the data collected in the evaluation come from real users in a real learning context. To collect the exploratory data we have used initial questionnaires to know the previous experiences and expectations of the users; final questionnaires to know the ease of use and the usefulness in learning English of the platform perceived by the users; and observations of the learning scenario. In addition, quantitative metrics we have been collecting, mainly through logs, the navigation in the main pages, the time spent in each course activity or the frequency of use (number of visits) of a given exercise.

All these techniques have allowed us to collect information and see different aspects of the use of AutoLearn use:

- Initial questionnaire for learners: to get information of the learner’s profile (language level profile, ICT user profile) before the use of AutoLearn and their expectations about computer assisted language courses.

- Activity logs: to get information of the user’s activity when they are using the platform and analyse the supportiveness for the user needs. The log files provide information about what the user is doing and their navigation through the platform but it not provides information about their intentions, objectives or needs.

- Informal interviews with the teachers: to get information about their needs when using AutoLearn in their classes.

- Final questionnaire for learners: to collect user perception and valuations of 1) the effectiveness, usefulness and general satisfaction with the platform; 2) the ease of use of all the items of the platform; 3) the pedagogic and linguistic aspects.

3 DESCRIPTION OF THE EVALUATION SETTING AND PARTICIPANTS

As stated before, the evaluation started with the preparation and execution of a set of learning scenarios to be evaluated in terms of usability and adequacy of the VLE (moodle, http://www.moodle.org), the training materials and the feedback provided by the NLP-based virtual tutor. Three major tasks were carried out: i) recruitment of institutions to participate in the testing actions executed in real instruction settings; ii) preparation of the technical resources for the execution of the evaluation experiments; and iii) execution of the testing actions. Each testing action included a first session to present briefly the goals and basic characteristics of the project and the platform, a series of (voluntary) sessions working on the learning platform, and a final session to answer a user satisfaction questionnaire. During testing actions teachers and learners were assisted in terms of
technical problems with the platform, and a set of observations were carried out during some the face-to-face sessions.

To be able to define and characterise the learning scenarios evaluated we required testing institutions and subjects (teachers and learners) to answer respectively a preliminary questionnaire. From these questionnaires we obtained information related to knowledge field of the English class (general or specific), educational level, learning approach (distance or blended), teacher background, degree of integration of the AutoLearn activities in the course programme, and language learner profile. Before the actual testing we met with the teachers responsible of each class and settled together the actual contents (units) to be tested. The choice was made in accordance with the language level of the learners and the appropriateness of the contents in relation the course topic (since there was no placement test, both decisions were left to the teacher). The evaluation procedure of the platform and the testing actions (criteria and techniques for data collection) were defined internally by project members.

All in all, five educational centres participated in the testing actions: two in Barcelona (the Programa d'Esenyament d'Idiomes of the University Pompeu Fabra, and the Fundació Llor – a primary and secondary school –, two in Istanbul (University of Bogaziçi and University of Istanbul), one in Edinburgh (Heriot-Watt University) and one in Saarbrücken (the Sprachenzentrum of the University of the Saarland). All in all, it amounted to 289 testing subjects (13 of which were teachers) divided in 14 course groups.

### 3.1 Preparation and description of testing action and subjects in PEI

In this paper we only present a preliminary analysis of the testing action run in the Programa d'Esenyament d'Idiomes (PEI) of the University Pompeu Fabra, which is conceived as a centre to promote linguistic competence in Foreign Languages among the academic and research community. In addition, PEI’s approach to language training is to develop the command of oral skills and to foster learner autonomy. The courses that participated in the testing actions were a group of learners in PEI Nivell 3 (B1.2 CEF level), three groups in PEI Nivell 5 (B2.2 CEF level) and one group in PEI Nivell 7 (C1.2 CEF level).

The testing actions in PEI involved a total of 69 learners and five teachers. All learners answered an initial questionnaire to find out about their profile as language learners, computer users and about their expectations on computer-mediated courses (the learner profile basically adapted from Freed et. al [11]; two out of the five teachers answered the corresponding questionnaire for teachers, and they were (collectively) interviewed in three occasions before the actual testing started.

**A. Teacher profile/definition**

When organising the testing action, we met teachers in three preparation meetings, where they were told about the objectives of the testing actions, the contents of the learning units and some basic details on the user interface of the application. Along the preparation period teachers had a look at the contents of the courses offered. The courses offered where only the English ones:

- English, intermediate (B2): Customer service and international communication
- English, advanced (C1): Business reports

From the materials in these courses, teachers were required to select those that they considered better suited for the purposes of their course programmes, including at least two units (each of which required between 30 min or 45 min of learner work).

According to the interviews and the questionnaires answered teachers had little experience using Virtual Learning Environments, and their expectations with regard to the benefits of using a VLE for their language classes were between moderate and low.

**B. Learner profile**

Out of the 69 testing subjects (learners) 27 (39%) were male and 42 (61%) female. 50 (72%) learners stated to speak Catalan as a native language, 18 (27%) stated to speak Spanish and one (1%) German. Similar percentages were obtained for the language they usually spoke at home (48, 70%, Catalan and 20, 29%, Spanish). All of them except one person learned English at school (93% of at least for three years). 36 of them (52%) learned English, in addition, in a foreign language school, and 11 (16%) of them had a working experience and en English-speaking environment.
As for their profile as computer users, 65 (94%) subjects declared to use computers on a daily basis, so in principle they are frequent users of Information and Computer Technologies (ICT) and 41% them said to regularly use the Internet for the purpose of their studies, which might be relevant in the sense that they are regular users (consumers) of web-based applications.

86% of the testing subjects never went through the experience of using computer-assisted language learning (CALL) systems. However, 41% of them had the impression that CALL systems could be a good way to learn foreign languages, 58% stated they were unsure whether it would be useful or not, and only 1% thought that this would not be a good way to learn a foreign language.

Through open questions in the questionnaire we found out that the most common preconceptions among the testing subjects were that CALL systems (most frequent preconceptions appear first):

- Represent a flexible way to learn, since they can be used anytime and anywhere. Also because users can freely decide on the time devoted to the learning activity, as well as on the kinds of activities.
- Result in a lack of communication, since using computers represents to spend less time talking to their instructor and classmates, less time in face-to-face learning.
- Require more discipline on the learner side, since the learner is more often left on their own.
- Can nicely be used as a complement to face-to-face instruction but not to replace it.
- Is more interactive and fun, since it allows teachers/learners to use movies, interactive exercises, as learning materials.

Generally speaking, computer-assisted courses are perceived as an individual activity (as opposed to collective) way to learn languages. This feature can be perceived as a positive one: users can decide/adapt the regularity of their learning activity (a flexible way to learn) and even choose the kinds of exercises (as a first step into user adaptive learning). But it can also be perceived as negative feature: it is seen as a menace to communication with the teacher and classmates. This perception might reinforce the appropriateness of using learning platforms such as AutoLearn in combination with face-to-face instruction, that is, to go for the blended learning approach.

4 RESULTS OF THE EVALUATION

As previously explained, the aim of this evaluation has been to assess the usability of the platform, to evaluate the pedagogical and linguistic appropriateness of the courses, and to assess user satisfaction with the platform and the language courses. With this aim we collected both the answers to a final questionnaire on the experience of using the platform and the navigation through the platform (activity logs). The combination of these two techniques has brought important information to know how the courses have been used in real learning environments.

4.1 Learner’s behaviour using the platform (logs)

During the testing actions in the language institution, we have been collecting the logs of the students using the information stored in Moodle’s database. Moodle captures all platform activities performed by learners and teachers through a relational database system. For the purpose of the AutoLearn’s evaluation a set of tables were added to the platform’s database in order to be able to have user activity organised in terms of sessions (a session was defined as the action of logging in and out of the platform), which is not implemented in standard Moodle installations. We also developed a series of tables and algorithms to capture the reaction of learners in front of exercises with automatic correction, since the virtual tutoring module is one of AutoLearn’s contributions to the platform.

In order to visualize the logs we have used a tool of information visualization [12]. With this tool we are able to see the structure of the courses and exercises in each course and get an image of the user’s behaviour when navigating through the courses. In Figure1, the centre of the circle corresponds to the initial page of the institution (PEI). In the first circle (near to the centre) there are the two levels (advanced and intermediate). In the following circle there are the five courses and, in the most external circle, there is the representation of the different exercises and instructions.
Fig. 1. Visualization of courses structure

This tool allows us to visualize the number of visits in every page of the platform, classify them taking into account the different CEF levels, or even the kind of exercise, and see the time spent in each page. For this paper we have chosen to analyse two basic metrics related with user learning activity: frequency of use and time spent.

With the frequency of use we wanted to evaluate some didactic and linguistic aspects of the exercises: which kinds of exercises are more attractive to learners? Do they do all the exercises of the same unit or they jump from one unit to the other? In Fig. 2 we see the frequency of use of each exercise, represented by a bar (the larger the more number of visits it had). The figure also distinguishes the sort of automatic feedback used: simple feedback (Hot Potates) in dark grey, and more complex NLP-based feedback (AutoTutor) in light grey.
Through this graphic (and other more detailed graphics we did not include for space reasons) we observed that Hot Potatoes (using simple correction strategies) were the exercises with a higher number of visits in all five courses. At this stage, we were yet unable to tell whether this was because of their content or because of their appearance and layout. As we will see, questionnaires will give us some further evidence to interpret this fact.

We also observed that usually students did all the exercises in one unit (each course having five or six units) and, that some units were not visited at all. This fact is very possibly a consequence of the testing instructions. In the testing actions run in the PEI, students were required to go through at least two units to complete the testing, and the option to go through the other three or four units was optional. We interpret this as a mechanism of economy: learners do not do any exercise beyond those they are required to, and as soon as they comply with the testing requirements their motivation decreases and the platform itself is not an incentive in its own.

Another interesting aspect is the decreasing frequency of visits of exercises within the same pedagogical unit: in most units the first exercise was the most visited and then there was a reduction of the number of visits. The interpretation of this is not easy, and cannot be further contrasted with other data collected: it could be because as they progress, the difficulty of the task decreases, but it could, among others, again a lack of motivation.

In Fig. 3 the average of the time spent in each exercise is represented with a circle. This metric allowed us to evaluate some relevant aspects concerning the didactic and linguistic dimension of the exercises and usability: is there any kind of exercise that requires more time that others? Do users spend a lot of time in one exercise because of its difficulty or because the lack of usability? Is the language level of the exercises suited to the learner’s level? Or are they too difficult? In the visualization we could see clearly which exercises required more time in contrast with others.
It is important to say that we did not have a preconception as to the time needed to do each given exercise, so we are not able to say if the time spent is too much or too little. But with the visualization of this metric we can compare between different kinds of exercises, between similar exercises and, besides, between the time spent in one given exercise in each of the courses (testing groups).

In this sense, the visualization tool helped us to realize that in all courses users spent more time in Hot Potatoes exercises (in dark grey in the diagram) than in AutoTutor exercises (in light grey). Moreover, among all the existing instances of a Hot Potatoes exercise, participants spent a lot of time in one particular exercise concerned with the vocabulary related to the Internet.

It is also interesting to remark that there was one course in which all the activities were done in less time (in the left-hand side of the picture, the course main node marked with a black star) in contrast with the time spent by learners in other courses. In this case, this can be attributed to the fact that the learner’s level of member of the group was higher than the one for which the course was designed: they were C1 CEF level learners going through an exercise corresponding to a B2 CEF level.

### 4.2 User acceptance towards AutoLearn

At the end of the testing action, learners were requested to answer a final questionnaire about the usability of the platform, some pedagogical and linguistic aspects of the courses, and their general satisfaction with AutoLearn. All questions were in form of an assertion that learners had to grade using a Likert scale between 1 and 7: 1 meaning “I don’t agree” and 7 meaning “I fully agree”.

To describe the results, we have grouped the answers related to three criteria: usability, feedback and didactic and linguistic aspects.
A. Usability

The average of the punctuation of the questions related to usability was 5.3. Among the questions related to usability, we have also made three different groups: questions on the appearance of the course, questions on the ease of use of the feedback buttons (check button, hint button, etc.), and on the ease of use of the navigation facilities. With this classification, we could see that the ease of use of the feedback and the navigation facilities through the platform received a good judgment (5.4 and 5.5), while the judgment of the appearance of the courses was lower (4.7).

Complementing these results with the learner’s profile, we can say that the navigation and the applications usability were not a problem for the participants due their acquaintance with ICT and the ease of use of the platform. However, the appearance of the pages used to access certain important locations (main platform page, main course page, platform structure pages, etc.) and the activities did not satisfy users. This can be added as a negative effect on the general motivation of users towards the courses offered in AutoLearn.

B. Feedback

The average of the punctuation related to the usefulness of the automatic feedback was 5.3. Dividing the results related to Hot potatoes exercises and AutoTutor exercises, we see that the former were better perceived more positively (5.9) than the latter (5). Among the different kinds of feedback functionalities available in the activities, users are pleased with the “check” button, one of the most useful ones. This button provides an instant feedback, confirming if the answer is right or wrong. Other types of feedback functionalities (the Hint button) provided a hint for the correct answer. In this sense, we noted that the users put more value on the possibility to check their own answers than on the possibility to receive help for the correct answer.

C. Didactic and linguistic aspects

From the final questionnaires we also get users’ impressions on the didactic and linguistic benefits of AutoLearn. The level of difficulty of the activities (suitability to their language level) was rated with an average score of 5.1 and the score to the performance of the automatic correction tools was 5.6. Regarding these punctuations we can say that users considered that activities were appropriate to their level. However, they found problems in doing some activities because the average score for the statement “In the activities, what to do and how to do was always very clear.” was quite low (4.7) in contrast with the suitability or the performance of the automatic correction tools (almost one point lower).

Besides, we asked users about how AutoLearn helped them in improving their linguistic competence. The average of the punctuation related to improving vocabulary was higher (5.1) compared to the average impression learners had with respect to how AutoLearn helped them improving their grammar (4.1), their listening skills (4.5) or their reading skills (4.2).

5 CONCLUSIONS AND DISCUSSION

The combination of the information collected from user navigation and the information obtained through the final questionnaires gave us very relevant information about end user preferences, behaviour and needs.

Sometimes this combination of information types provided us with a double source of evidence for certain facts. For instance, in the final questionnaires learners stated that they preferred Hot Potatoes activities rather than AutoTutor activities. In addition, in the logs visualization we also saw that Hot Potatoes activities were more frequently used. We also know that these two pieces of evidence and are not related with usability or appearance. We have to further investigate to find out whether this could be related to the proportion of Hot Potatoes and AutoTutor exercises (usually in a 5:1 ratio in each unit) or this could be related to aspects such as the quality of the feedback provided, or more general aspects of human behaviour as suggested by the findings of Heift [13].

Another interesting aspect is that the information obtained through the different collection methods could be used to further refine interpretations or discard certain hypotheses. For instance, we saw that users spent a lot of time in vocabulary activities, particularly in those activities were the vocabulary was related to the Internet. Our first impression was that this kind of activity was very difficult for the learners; but it could also have been that the topic was very interesting to them. In the final questionnaires, when users evaluated the topics of learning activities, they gave them an average
score of 5.1 (out of 7). Moreover, they thought that courses were useful to improve their vocabulary, again 5.1 (out of 7). Given all these pieces of information, we tend to think that they found it interesting, and that it probably was adequate to their learner level.

These preliminary results show the importance of get different type of information for language learning platforms. At present we are carrying out a more detailed analysis of the data specially to examine the usefulness of using NLP techniques to automatically generated feedback.

6 ACKNOWLEDGEMENTS

We are grateful to Víctor Pascual and Sergi Mínguez for the collaboration and help with the use and the setting up of the environment and the tool for data visualization.

7 REFERENCES


