A Learning Environment for English for Academic Purposes based on Adaptive Tests and Task-based Systems

Jean Piton Gonçalves[1], Sandra Maria Aluisio[1], Leandro H. M. de Oliveira[1], Osvaldo N. Oliveira Jr. [1,2]

[1]Núcleo Interinstitucional de Lingüística Computacional (NILC),

ICMC-University of São Paulo, CP 668, 13560-970 São Carlos, SP, Brazil,

Phone number: +55 16 273 9663

[2]Instituto de Física de São Carlos, USP, CP 369, 13560-970 São Carlos, SP, Brazil,

Phone number: +55 16 273 9825

jpiton@yahoo.com, sandra@icmc.usp.br, landroh@nilc.icmc.usp.br, chu@if.sc.usp.br

Resumo

A necessidade de jovens pesquisadores lerem correta e rapidamente uma grande quantidade de textos escritos em inglês, que é a *língua franca* da ciência, representa uma barreira considerável para eles. Dada essa necessidade, em 2001, o programa de mestrado em Ciências da Computação e Matemática Computacional do ICMC-USP passou a avaliar a proficiência em inglês (PI) dos alunos quanto à habilidade de reconhecerem o gênero de textos científicos em inglês, com a estrutura e as convenções que lhe são características. O site do exame de PI disponibiliza Exames Modelos com correção automatizada para os alunos saberem antecipadamente como será o exame formal, também informatizado. Porém, a prática com o Exame Modelo fornece apenas um relatório com o escore do aluno, não oferecendo um ambiente propício para ele relembrar seus conhecimentos, receber instrução ou ainda rever conceitos errôneos. Para avaliar a hipótese que a integração de um teste adaptativo informatizado (TAI) e um ambiente computacional de tasks (ACT) geraria um ambiente baseado na abordagem aprender-fazendo que fosse útil para ensinar a estrutura e as convenções do gênero de textos científicos a jovens pesquisadores iniciando a carreria acadêmica, foi projetado, implementado e parcialmente avaliado um Ambiente Computacional de Aprendizagem (ACA) para o inglês instrumental. Embora o ACA tenha sido instanciado com o conhecimento do inglês instrumental dos EPI do ICMC-USP ele é genérico o suficiente para servir de modelo para outros domínios e/ou para ser usado em outros programas de mestrado que avaliam o inglês instrumental e por jovens pesquisadores que desejam conhecer a estrutura e as convenções do gênero de textos científicos em inglês.

Abstract

This papers introduces the environment CALEAP-Web that integrates an adaptive testing system, ADEPT, with a task-based environment, CATESE, in the domain of English for academic purposes. It is aimed at assisting graduate students at ICMC-USP for the proficiency English test, which requires them to be knowledgeable of the conventions of scientific texts. Both testing (ADEPT) and learning (CATESE) system comprises four modules dealing with different aspects of Instrumental English, which include a bank of items for evaluation with 140 questions. These modules were based on writing tools for scientific writing, which employ case-based reasoning. In CALEAP-Web, the students are

assessed on an individual basis, taking full advantage of the adaptive nature of ADEPT. They are then guided through appropriate learning tasks of CATESE to minimize their deficiencies, in an iterative process until the students perform satisfactorily in the tasks. The robustness of CALEAP-Web was demonstrated with simulations, working as specified in all stages. An analysis was made of the item exposure in the adaptive testing, which is crucial to ensure high-quality assessment. Though conceived for a particular domain, the rationale and the tools developed for CALEAP-Web may be extended to other domains.

1. Introduction

There is a growing need for students from non-English speaking countries to learn and employ English in their research and even in school tasks. Only then can these students take full advantage of the enormous amount of teaching material and scientific information in the WWW, which is mostly in English. For graduate students, in particular, a minimum level of instrumental English is required, and indeed universities tend to require the students to undertake proficiency exams. There are various paradigms for both the teaching and the exams which may be adopted. In the Institute for Mathematics and Computer Science (ICMC) of University of São Paulo, USP, we have decided to emphasize the mastering of English for academic purposes. Building upon previous experience in developing writing tools for academic works (Aluisio & Oliveira, 1995, Aluisio & Gantenbein, 1997, Aluisio et al, 2001), we conceived a test that checks whether the students are prepared to understand and make use of the most important conventions of scientific texts in English (Aluisio et al, 2003). This fully-automated test, called CAPTEAP[1], consists of objective questions in which the user is asked to choose or provide a response to a question whose correct answer is predetermined (Mckenna & Bull, 1999). CAPTEAP comprises four modules (Aluisio et al, 2003), explained in Section 2. In order to get ready for the test – which is considered as an official proficiency test required for the MSc. at ICMC, students may undertake training tests that are offered in the CAPTEAP system. However, until recently there was no module that assisted students in the learning process or that could assess their performance in their early stage of learning. This paper describes the *Computer-Aided Learning of English for Academic Purposes* (CALEAP-Web) system that fills in this gap, by providing students with adaptive tests integrated into a computational environment with a variety of learning tasks.

CALEAP-Web employs a computer-based adaptive test (CAT), *Adaptive English Proficiency Test for Web* (ADEPT), with questions selected on the basis of the estimated knowledge of a given student, being therefore a fully customized system. This is integrated into the *Computer-Aided Task Environment for Scientific English* (CATESE) (Gonçalves, 2004) to train the students about conventions of the scientific texts, in the approach known as *learning by doing* (Schank, 2002).

2. Computer-based Adaptive Tests

The main idea behind adaptive tests is to select the items of a test according to the ability of the examinee. That is to say, the questions proposed should be appropriate for each individual, being therefore different from one examinee to the other. An examinee is given a test that adjusts to the responses given previously. If the examinee provides the correct answer for a given item, then the

[1] http://www.nilc.icmc.usp.br/capteap/

next one is harder. If the examinee does not answer correctly, the next question can be easier. This allows a more precise assessment of the competences of the examinees than traditional multiple-choice tests because it reduces fatigue, a factor that can significantly affect an examinee's test results (Olea & Prieto, 1999). Other advantages are an immediate feedback, the challenge posed as the examinees are not discouraged or annoyed by items that are far above or below their ability level, and reduction in the time required to take the tests.

2.1 Basic components of a CAT

According to Conejo et al. (2001), Adaptive Testing based on Item Response Theory (IRT) comprises the following basic components: a) an IRT model describing how the examinee answers a given question, according to his/her level of knowledge. When the level of knowledge is assessed, one expects that the result should not be affected by the instrument used to assess, i.e. computer or pen and paper; b) a bank of items containing questions that may cover part or the whole knowledge of the domain. c) level of initial knowledge of the examinee, which should be chosen appropriately in order to reduce the time of testing. d) method to select the items, which is based on the estimated knowledge of the examinee, depending obviously on the performance in previous questions. e) stopping criteria that are adopted to discontinue the test once the pre-determined level of capability is achieved or when the maximum number of items have been applied, or if the maximum time for the test is exceeded, among others.

2.2 ADEPT

ADEPT provides a customized test capable of assessing the students with only a few questions. It differs from the traditional tests that employ a fixed number of questions for all examiees and do not take into account the previous knowledge of each examinee.

2.2.1 Item Response Theory

This theory assumes some relationship between the level of the examinee and his/her ability to get the answers right for the questions, based on statistical models. ADEPT employs the 3-parameter logistic model (Lord, 1980) given by the expression:

$$P(\theta) = c + (1-c)\frac{1}{1+e^{-1.7a(\theta-b)}}$$

where a (discrimination) denotes how well one item is able to discriminate between examinees of slightly different ability, b (difficulty) is the level of difficulty of one item and c (guessing) is the probability that an examinee will get the answer right simply by guessing

2.2.2 Item calibration

To calibrate bank items means to specify the parameters (a,b,c) for them. The bank of items employed by ADEPT contains questions used in the proficiency tests of the ICMC in the years 2001 through 2003, for Computer Science, Applied Mathematics and Statistics. There are 30 tests, with about 20 questions each. The insertion in the bank and checking of the questions were carried

out by the first author of this paper. Without considering reuse of an item, there are 140 questions with no repetition of texts in the bank.

The proficiency test contains four modules:

1. Module 1 - conventions of the English language in scientific writing, It deals with knowledge about morphology, vocabulay, syntax, the verb tenses and discourse markers employed in scientific writing. Today, this module covers two components of Introductions[2], namely *Gap* and *Purpose*;
2. Module 2 - structures of scientific texts. It deals with the function of each section of a paper, covering particularly the *Introduction* and *Abstract*;
3. Module 3 - text comprehension, aimed to check whether the student recognizes the relationships between the ideas conveyed in a given section of the paper.
4. Module 4 - strategies of scientific writing. It checks whether the student can distinguish between rhetorical strategies such as definitions, descriptions, classifications and argumentations, covering two parts of the Introduction, namely *Setting* and *Review of the Literature*.

The questions for Modules 1 and 4 are simple, independent from each other. However, the questions for Modules 2 and 3 are interdependent, associated with a text in English (i.e. they are testlets) (Oliveira, 2002).

Calibration of the items is carried out with the algorithm of Huang (1996), viz. the Content Balanced Adaptive Testing (CBAT-2), a self-adaptive testing which calibrates the parameters of the items during the test, according to the performance of the students. In the ADEPT, there are three options for the answers (choices a, b, or c). Depending on the answer (correct or incorrect), the parameter b is calibrated and the updating of the parameters R (number of times that the question was answered correctly in the past), W (number of times the question was answered incorrectly in the past) and Φ (difficulty accumulator) (Huang, 1996). Even though the bank of items in ADEPT covers only Instrumental English, several knowledge domains may be present. Therefore, the contents of the items had to be balanced (Huang, 1996), with the items being classified according to several sections. In ADEPT, the contents are split into the Modules 1 through 4 with 15%, 30%, 30% and 25%, respectively. As for the weight of each component and Module in the curriculum hierarchy (Huang, 1996), 1 was adopted for all levels. In ADEPT the student is the agent of calibration in real time of the test, with his/her success (failure) in the questions governing the classification of the items in the bank.

2.2.3 Estimate of the Student Ability

In order to estimate the ability θ of a given student, ADEPT uses the modified iterative Newton-Raphson method (Lord, 1980), using the following formulas:

---

[2] According to Weissberg and Buker (1990) the main components of an Introduction are Setting, Review of the Literature, Gap, Purpose, Methodology, Main Results, Value of the Work and Layout of the Article.

$$\theta_{n+1} = \theta_n + \frac{\sum_{i=1}^{n} S_i(\theta_n)}{\sum_{i=1}^{n} I_i(\theta_n)}$$

$$S_i(\theta) = [r_i - P_i(\theta)] \frac{P_i'(\theta)}{P_i(\theta)[1 - P_i(\theta)]}$$

where θ_n is the estimated ability after the nth question. $r_i = 1$ if the ith- answer was correct and $r_i = 0$ if the anwer was wrong.. For the initial ability $\theta_0 = 0.0$ was adopted.

2.2.4 Stopping Criteria

The criteria for stopping an automated test are crucial. In ADEPT two criteria were adopted:

- The number of questions per module of the test is between 3 (minimum) and 6 (maximum);
- θ should lie between -3.0 and 3.0 (Baker, 2001).

3. Task-based Environments

A task-based environment provides the student with tasks for a specific domain. The rationale of this kind of learning environment is that the student will learn by doing, in a real-world task related to the domain being taught. There is no assessment of the performance from the students while carrying out the tasks, but in some cases explanations on the tasks are provided.

3.1 CATESE

The *Computer-Aided Task Environment for Scientifc English* (CATESE) comprises tasks associated with the 4 modules of the Proficiency tests described in Section 2. The tasks are suggested to each student after performing the test of a specific module. This is done first for the "easier" Modules (1 and 4) and then for the most difficult (Modules 2 and 3), seeking also a balance for the reading of long and short chunks of text.

The four tasks are as follows:

- Task 1 (T1): **identification and classification** of discourse markers in sentences of the component *Gap* of an Introduction. Identification of verb tenses of the component *Purpose*;
- Task 2 (T2): **selection** of the components for an Introduction and retrieval of well-written related texts from a text base for subsequent **reading**;
- Task 3 (T3): **reading** of sentences with discourse markers for the student to establish relationships between the functions of the discourse and the markers;
- Task 4 (T4): **identification and classification** of writing strategies for the components *Background* and *Review of the Literature*.

The text base for Tasks 1, 3 and 4 of CATESE was extracted from the Support tool of AMADEUS (Aluisio & Oliveira, 1995), with the sample texts being displayed in XML. Task 2 is

an adaptation of CALESE[3] with filters for displaying the cases. Task 1 has 13 excerpts of papers with the components *Gap* and 40 for the *Purpose*, Task 2 has 51 Introductions of papers, Task 3 contains 46 excerpts from scientific texts and Task 4 has 34 excerpts from the component *Setting* and 38 for the component *Purpose*.

4. Integration of ADEPT and CATESE

The CALEAP-Web integrates two systems associated with learning and assessing tasks, as follows (Gonçalves, 2004):

- Module 1 (Mod1) – assessment of the student with ADEPT to determine his/her level of knowledge of Instrumental English.
- Module 2 (Mod2) – tasks are suggested to the student using CATESE, according to his/her estimated knowledge, particularly to address difficulties detected in the assessment stage.

Mod1 and Mod2 are integrated as illustrated in Figure 1. Information for modeling the user performance (L1) comes from the EPI module in which the student is deficient, θ and $P(\theta)$, normalized score of the student in the test, number of correct and incorrect answers and time taken for the test in the EPI module being assessed.

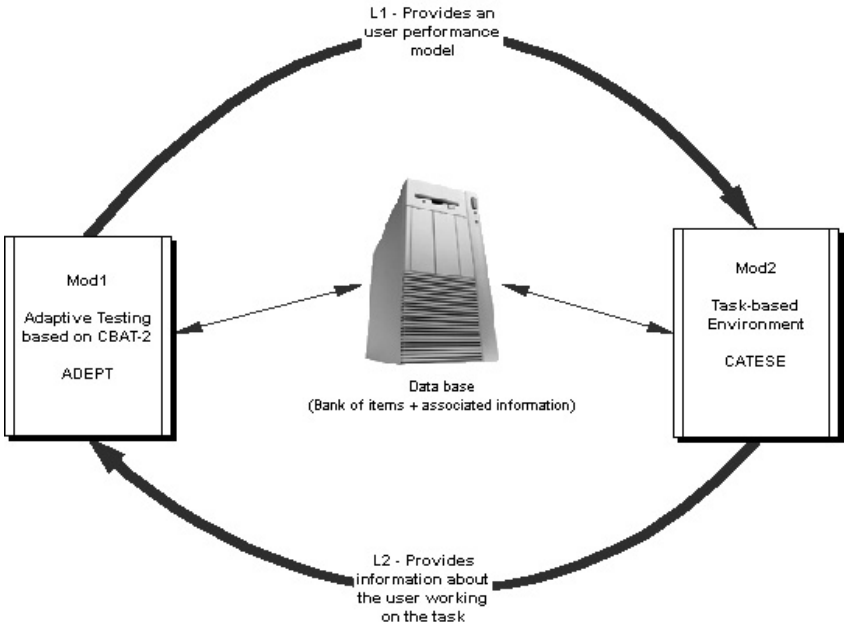

Figura 1. Integration Scheme in CALEAP-Web.

At the end of the test of each module of the EPI, the student will be directed to CATESE if his/her performance was below a certain level (if 2 or more answers are wrong in a given module). For example, if the student does not do well in Module 1 (involving *Gap* and Purpose) for questions associated with the component Gap, he/she will be asked to perform a task related to

[3] http://www.nilc.icmc.usp.br/calese/

Gap, but not *Purpose*. If the two wrong answers refer to *Gap* and *Purpose*, then two tasks will be offered, one for each component.

The information about the student (L2) includes the tasks recommended to the student and monitoring of how these tasks were performed. It is provided by CATESE to ADEPT, so that the student can take another EPI test in the module where deficiencies were noted. If the performance is now satisfactory, the student will be taken to the next test module. The sequence suggested by CALEAP-Web involves activities for Modules 1, 2, 4 and 3 of the EPI, presented in the following sections. In all tasks, chunks of text from well-written scientific papers are retrieved. The cases may be retrieved as many times as the student needs, and the selection is random. After the task, the student will be assessed by the CAT, which will check whether he/she is prepared to start the next module.

4.1 Task 1

Task 1 deals with the components *Gap* and *Purpose* of Module 1 from EPI, with the texts retrieved belonging to two classes for the *Gap* component:

1. *Class A: special words are commonly used to indicate the beginning of the Gap. Connectors such as **however** and **but** are used for this purpose. The connector is followed immediately by a gap statement in the present or present perfect tense, which often contains modifiers such as **few**, **little**, or **no**. Signal word + Gap (present or present perfect) + Research topic;*
2. *Class B: subordinating conjunctions like **while**, **although** and **though** can also be used to signal the gap. When such signals are used, the sentence will typically include modifiers such as **some**, **many**, or **much** in the first clause, with modifiers such as **little**, **few**, or **no** in the second clause. Signal word + Previous work (present or present perfect) + Gap + topic.*

In this classification two chunks of text are retrieved, where the task consists in the **identification** and **classification** of markers in the examples, two of which are shown below.

Class A

However, in spite of this rapid progress, many of the basic physics issues of x-ray lasers remain poorly understood.

Class B

Although the origin of the solitons has been established, some of their physical properties remained unexplained.

The texts retrieved for the *Purpose* component are classified as:

1. *Class A: the orientation of the statement of purpose may be towards the report itself . If you choose the report orientation you should use the present or future tense. Report orientation + Main Verb (present or future) + Research question;*
2. *Class B: the orientation of the statement of purpose may be towards the research activity. If you choose the research orientation you should use the past tense, because the research*

activity has already been completed. Research orientation + Main Verb (past) + Research question.

The Tasks consists in **identifying** and **classifying** the markers in the examples for each class, illustrated below.

Class A

In this paper we report a novel resonant-like behavior in the latter case of diffusion over a fluctuating barrier.

Class B

The present study used both methods to produce monolayers of C16MV on silver electrode surfaces. This study sought to clarify literature reports regarding the observation of splitting of the first reduction peak in the CV of C16MV during one-electron reductions.

4.2 Task 2

Task 2 is related to the *Introduction* of Module 2 of EPI, which provides information about the components of an Introduction of a scientific paper in English. The student **selects** the components and strategies so that the system retrieves the cases (well-written papers) that are consistent with the requisition and **reads** them. With this process, the student may **learn by examples** where and how the components and strategies should be used. The components are classified as follows:

- *In the **setting**, the writer establishes a context, or frame of reference to help readers understand how the research fits into a wider field of research;*
- *In the **review of the literature**, the writer reviews the findings of other researchers who have already published in his/her area of interest;*
- *In the **gap**, the writer indicates that previous literature described in the review of the literature is inadequate because an important aspect of the research area has been ignored by other authors; OR indicates that there is an unresolved conflict among authors of previous studies concerning the research topic; OR indicates that an analysis of the previous literature suggests an extension of the topic; OR raises a new research question not previously considered by other workers in the field;*
- *In the **purpose**, the writer formally announces the purpose(s) of the research;*
- *In the **methodology**, the writer describes either the steps followed in conducting the research or some method considered in the study;*
- *In the **main results**, the writer presents the main findings of the research;*
- *In the **value of the research**, the writer indicates possible benefits or application of the research;*
- *In the **structure of the article**, the writer presents the structure or layout of the paper, indicating the sections and briefly commenting on them.*

This task was created from the Support Tool of AMADEUS (Aluisio et al, 2003), which employs case-based reasoning (CBR) to model the three stages of the writing process: the user selects the intended characteristics of the Introduction of a scientific paper, the best cases are retrieved from the case base, and the case chosen is modified to cater for the user intentions. The

student may repeat this task and select new strategies (with the corresponding components). After the task, the student is assessed by the CAT, and if successful, he/she may go to Module 4 of EPI. Otherwise, he/she will have to perform Task 2 again.

4.3 Task 4

Task 4 deals with the *Setting* and *Review of the Literature* from Module 4 or EPI. The cases retrieved are classified into 3 classes:

1. *Class A: Arguing about the topic prominence: uses arguments;*
2. *Class B: Familiarizing terms or objects or processes: follows one of the three patterns: description, definition or classification;*
3. *Class C: Introducing the research topic from the research area: follows the general to particular ordering of details.*

Three examples are retrieved from the case base, and the student is asked to **identify and classify** the various writing strategies in the text for the *background* component of the Introduction. An example of retrieved text follows:

Class A

Si (111) is one of the most studied semiconductor surfaces. It has been established, both experimentally [1-5] and theoretically [6-8], that the surface has a 77 reconstruction of Takayanagi type [9] which disorders at around 870c to a high temperature phase commonly referred to as "11" [2,3].

Class B

Phycoerythrin is the outermost phycobiliprotein of the phycobilisome "light harvesting system" found in red algae. Individual phycoerythrin phycobiliproteins are highly and characteristically fluorescent, being twenty-fold more fluorescent than the chromophore fluorescein on a molar basis. These proteins also possess an unusually large Stokes' shift, 81 nm (495 nm excitation and 576 nm emission), which is approximately 2.7 times that of fluorescent [3, 4].

Class C

Noise cancellation is a well-established technique in acoustics [1] and electronic signal processing [2] whereby undesirable fluctuations are suppressed by destructive interference. This is achieved by the addition of an "antinoise" component equal in amplitude, but with opposite phase. In this Letter, we demonstrate that an optical analog of this phenomenon occurs naturally in the intensity-fluctuation noise spectrum of the laser.

For the *Review of the Literature*, there are three classes

1. *Class A: Citations grouped by approaches: better suited for reviews of the literature which encompass different approaches;*
2. *Class B: Citations ordered from general to specific (General to particular ordering for citations): citations are organized in order from those most distantly related to the study to those most closely related;*

3. *Class C: Citations ordered chronologically (Historical Review): used, for example, when describing the history of research in an area.*

An example of retrieved text follows:

Class A

Barnum et al. [5] used an indirect technique to determine ACmix, the excess specific heat due to mixing of a binary polymer blend. Ahn et al. [6] have studied blends of a low-molecular- weight liquid crystal (LC) with a polymer, using optical techniques and differential scanning calorimetry (DSC). They stated that the changes in heat quantities due to phase separation in such systems are too small to be detected using DSC.

Class B

Previous results suggested a role for the nature of the anionic species present in the supporting electrolyte. The electrochemical, spectroelectrochemical and surface enhanced Raman spectroscopy (SERS) of C16MV in monolayer form have been treated in an extended monograph [1], which also summarizes many previous studies. Observation of splitting of the first reduction peak in the cyclic voltammogram of LB monolayers of C16MV on glassy carbon (in this case the viologen is irreversibly bound) was independently reported [2]. Park et al. [3], after a study of a related surface active piperidinyloxy viologen, found that the reduction peak was sometimes ill-defined and sometimes showed two peaks.

Class C

In a paper entitled "significance of electromagnetic potentials in quantum theory" published in 1959, Aharonov and Bohm [1] proposed two types of actual electron interference experiments aimed at exhibiting these conclusions. The phenomena predicted came to be known as the Aharonov-Bohm (AB) effect, and have given rise to a literature of almost 400 journal articles over the last thirty-odd years. The essence of the AB experiments [2] is that electrons suffer phase shifts in passing through regions of space of zero fields but nonzero potentials. The effects are of two types, the usual magnetic (or vector) AB effect, and the less often cited electric (or scalar) AB effect which is conceptually quite simple. It concerns the phase shift caused by the scalar potential v = -eu in the Schrodinger equation: ... To exhibit the scalar AB effect, the potential of cylinder 2 alone is pulsed during a time when the wave packet is contained inside it.

4.4 Task 3

The last Task is related to *Comprehension* of Module 3 of EPI. Here a sequence of discourse markers are presented to the student, organized according to their function in the clause (or sentence). Also shown is an example of well-written text in English with annoted discourse markers. Task 3 therefore consists in **reading** and **verifying** examples of markers for each discourse function. The nine functions considered are: contrast/opposition, signalling of further information/addition, similarity, examplification, reformulation, consequence/result, conclusion, explanation, deduction/inference. The student may navigate through the cases and after finishing, he/she will be assessed by the CAT.

It is believed that after being successful in the four stages described above in the CALEAP-Web system, the student is prepared to undertake the official test at ICMC-USP.

5. Evaluating CALEAP-Web

CALEAP-Web has been assessed according to two main criteria: item exposure of the CAT module and robustness of the whole computational environment. With regard to robustness, we ensured that the environment works as specified in all stages, with no crash or error by simulating students using the 4 tasks presented in Section 4. The data from four students that evaluated ADEPT, graded as having intermediate level of proficiency[4], were selected as a starting point of the simulation. All the four tasks were performed and the environment was proven to be robust to be used by prospective students in preparation for the official exam in 2004 at ICMC-USP. The analysis of item exposure is crucial to ensure a quality assessment. Indeed, item exposure is critical because adaptive algorithms are designed to select optimal items, thus tending to choose those with high discriminating power (parameter a). As a result, these items are selected far more often than other ones, leading to both over-exposure of some parts of the item pool and under-utilization of others. The risk is that over-used items are often compromised as they create a security problem that could jeopardize a test, especially if it's a summative one. In our CAT parameters a and c were constant for all the items, and therefore item exposure depends solely on parameter b. To measure item exposure rate of the two types of item from our EPI (simple and testlet) we performed two experiments, the first with 12 students who failed the 2003 EPI and another with 9 students that passed it. From the 140 items only 66 were accessed and re-calibrated[5] after both experiments, where 30 of them were from testlets. Testlets are problematic because they impose application of questions as soon as selected. The 21 testlets of CAT involve 78 questions, with 48 remaining non re-calibrated. As for the EPI modules, most calibrated questions were from modules 1 and 4 because they include simple questions, allowing more variability in items choice. In experiment 1 questions 147 and 148 were accessed 9 times, with 16 questions being accessed only once and 89 were not accessed at all. In experiment 2, the most accessed questions were 138, 139 and 51 with 9 accesses each. On the other hand, 16 questions had only one access and 83 were not accessed at all. Taken together these results show the need to extend the studies with a larger number of students in order to achieve a more precise item calibration.

6. Related work

Particularly with the rapid expansion of open and distance-learning programs, fully-automated tests are being increasingly used to measure student performance as an important component in educational or training processes. This is illustrated by a computer-based large-scale evaluation using specifically adaptive testing to assess several knowledge types, viz. the *Test of English as a Foreign Language*[6] (TOEFL). Other examples of learning environments with an assessment module are the Project entitled Training of European Environmental trainers and technicians in order to disseminate multinational skills between European countries (TREE) (Conejo et. al,1999, 2000, 2001) and the Intelligent System for Personalized Instruction in a Remote Environment (INSPIRE) (Papanikolaou, 2001). TREE is aimed at developing an Intelligent Tutoring System

[4] θ in the range $-1.0 \leq \theta \leq 1.0$

[5] The second author has realized a pre-calibration of the parameter b of all the 140 items from the bank, using a 4-value table including difficult, medium, easy and very easy item category with respectively 2.5, 1.0, -1.0 and -2.5 value.

[6] http://www.toefl.org/

(ITS) for classification and identification of European vegetations. It comprises three main subsystems, namely, an Expert System, a Tutoring System and Test Generation System. The latter, referred to as Intelligent Evaluation System using Tests for Teleducation (SIETTE), assesses the student with a CAT implemented with the CBAT-2 algorithm, the same we have used in this work. The task module is the ITS. INSPIRE monitors the students' activities, adapting itself in real time to select lessons that are adequate to the level of knowledge of the student. It differs from CALEAP-Web, which is based in the learn by doing paradigm. In INSPIRE there is a module to assess the student with adaptive testing (Gouli et al., 2001), also using the CBAT-2 algorithm.

7. Conclusions and Further Work

The environment presented here, referred to as CALEAP-Web, is a first, important step in implementing adaptive assessment in relatively small institutions, as it offers a mechanism to escape from a pre-calibration of test items (Huang, 1996). It integrates a CAT system and a task-based system, which serves both to assess the performance of users and assist them with a handful of learning strategies. The ones implemented in CALEAP-Web were all associated with English for academic purposes, but the rationale and the tools developed can be extended to other domains. One major present limitation of CALEAP-Web is the small size of the bank of items; furthermore, increasing this size is costly in terms of man power due to the time-consuming corpus analysis to annotate the scientific papers used in both the adaptive testing and the task-based environment. With a reduced bank of items, at the moment we recommend the use of the adaptive test of CALEAP-Web only in formative tests and not in summative tests as we still have items with over-exposure and a number of them under-utilized.

Acknowledgments

This work was financially supported by FAPESP and CNPq.

References

Aluisio, S.M. & Oliveira Jr. O.N. A case-based approach for developing writing tools aimed at non-native English users. Lectures Notes in Artificial Intelligence 1010, 1995, pp 121-132.

Aluísio, S.M. & Gantenbein, R.E. Towards the application of systemic functional linguistics in writing tools. Proceedings of International Conference on Computers and their Applications, 1997, pp181-185.

Aluísio, S.M., Barcelos, I. Sampaio, J., Oliveira Jr., O N. How to learn the many unwritten "Rules of the Game" of the Academic Discourse: A hybrid Approach based on Critiques and Cases. In proceedings of the IEEE International Conference on Advanced Learning Technologies, Madison/Wisconsin, August 2001, pp 257-260.

Aluísio, S. M.; Aquino, V. T.; Pizzirani, R.; Oliveira JR, O. N. High Order Skills with Partial Knowledge Evaluation: Lessons learned from using a Computer-based Proficiency Test of English for Academic Purposes. Journal of Information Technology Education, Califórnia, USA, v. 2, n. 1, 2003, pp 185-201.

Baker, F. (2001) The Basics of Item Response. College Park, MD: ERIC Clearinghouse, University of Maryland.

Conejo, R.; E. Millán; J. L. P. Cruz; M. Trella. An empirical approach to online learning in siette. ITS-Intelligent Tutorial Systems, June 2000, pp 604–615.

Conejo, R.; E. Millán; J. L. P. Cruz; M. Trella. Modelado del alumno: um enfoque bayesiano. Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial (12), 2001, pp 50–58. URL http://tornado.dia.fi.upm.es/caepia/numeros/12/Conejo.pdf, (20/nov/2002).

Conejo, R.; A. Rios; M. T. Eva Millán; J. L. P. Cruz. Internet based evaluation system. AIED-International Conference Artificial Intelligence in Education, IOS Press, 1999. URL http://www.lcc.uma.es/~eva/investigacion/papers/aied99a.ps, (20/nov/2002).

Gonçalves, J. P. A integração de Testes Adaptativos Informatizados e Ambientes Computacionais de Tarefas para o aprendizado do inglês instrumental. (Portuguese). Dissertação de mestrado, ICMC-USP, São Carlos, Brasil, 2004.

Gouli, E; H. Kornilakis; K. Papanikolaou; M. Grigoriadou (2001, December). Adaptive assessment improving interaction in an educational hypermedia system. PC-HCI Conference., December 2001. URL http://hermes.di.uoa.gr/lab/CVs/papers/gouli/F51.pdf, (17/feb/2003).

Huang, S. X. On content-balanced adaptive testing algorithm for computer-based training systems. ITS-Intelligent Tutorial Systems, June 1996, pp 12–14.

Lord, F. M. (1980). Application of Item Response Theory to Practical Testing Problems. Hilsdale, New Jersey, EUA: Lawrence Erlbaum Associates.

McKenna, C. & J. Bull. Design effective objective test questions: an introductory workshop. Proceedings of the Conference at Loughborough University, Flexible Learning (Third), June 1999, pp 253–257.

Olea, J.; V. Ponsoda; G. Prieto (1999). Tests Informatizados Fundamentos y Aplicaciones. Ediciones Pirámede.

Oliveira, L. H. M. Testes adaptativos sensíveis ao conteúdo do banco de itens: uma aplicação em exames de proficiência em inglês para programas de pós-graduação. (Portuguese). Dissertação de mestrado, ICMC-USP, São Carlos, Brasil, 2002.

Papanikolaou, K.; M. Grigoriadou; H. Kornilakis; G. D. Magoulas. Inspire: An intelligent system for personalized instruction in a remote enviroment. Third Workshop on Adaptive Hypertext and Hypermedia, paper 3, July 2001. URL http://wwwis.win.tue.nl/ah2001/papers/papanikolaou.pdf, (17/feb/2003).

Schank, R. (2002). Engines For Education (Hyperbook ed.). Chicago, USA: ILS-Institute for the Learning Sciences, Northwestern University, 2003. URL http://www.engines4ed.org/hyperbook/index.html, (23/sep/2003).

Weissberg, R. & Buker, S. (1990).Writing Up Research - Experimental Research Report Writing for Students Of English. Prentice Hall Regents.