Assessing High-Order Skills with Partial Knowledge Evaluation: Lessons Learned from Using a Computer-based Proficiency Test of English for Academic Purposes

Sandra Maria Aluísio
Universidade de São Paulo & ICMC-USP, São Carlos, Brazil

sandra@icmc.usp.br

Valéria Tomas de Aquino
Centro Universitário Barão de Mauá, Ribeirão Preto, Brazil

valeriata@baraodemaua.br

Rafael Pizzirani
Universidade de São Paulo, São Carlos, Brazil

rafaelp@grad.icmc.usp.br

Osvaldo Novaes de Oliveira Jr.
Universidade de São Paulo & ICMC-USP, São Carlos, Brazil

chu@ifsc.sc.usp.br

Executive Summary

English proficiency tests (EPT) are required from students enrolled in Master and PhD programs in many non-English speaking countries, as the students must demonstrate their ability to understand and produce technical literature that is predominantly in English. Some universities and funding agencies in Brazil require the student to pass a general-purpose exam, such as TOEFL or IELTS. However, these exams do not evaluate students' competence in terms of the demands of highly standardized research articles written in English, which will be needed for the student to review the literature of his/her research area. In order to overcome this limitation we implemented a web-based, fully automated system for English Proficiency tests, named CAPTEAP, which checks the students' ability in the use of English for academic purposes. In this paper, we describe how CAPTEAP assesses high-order skills of Bloom's taxonomy, where students are assessed in their reception and familiarity with the schematic structure of scientific papers. CAPTEAP employs the resources of a case-based reasoning tool to assist non-English speakers in producing scientific papers. It comprises 4 modules, through which the students are evaluated with regard to their ability to: i) analyze the structure of a given section of a paper; ii) analyze relationships among clauses signalled by discourse markers; iii) recognize conventions and rhetorical functions in English from scientific texts; iv) identify writing strategies for each component of a section. CAPTEAP has been applied in the official proficiency test required from MSc. students at the Computer Science Department at University of São Paulo, Brazil. In order to be fully automated, CAPTEAP contains multiple-choice, objective questions but allows a refined assessment of abilities by using scoring procedures based on the Admissible Probability Measurement (APM). The lessons learned from introducing this fully automated test are presented in connection with the following topics: the web-based system interface, the modules and cognitive levels of the tasks in the tests, and the scoring procedure. As for the EPT interface, we learned that in formal exams users are unlikely to tolerate system faults as the latter may affect their performance. Evaluation of interfaces by real users

Special Series Editor: Salvatore Valenti

combined with less costly usability assessment procedures, such heuristic evaluation, has proven important to guide the development of a user-friendly interface. With regard to the evaluating strategies, we believe that Bloom's taxonomy allows one to select questions in a balanced way to test different abilities. Also, it is important to provide students with the opportunity to learn beforehand how the test works and to practice in simulated exams. Following experimentation with scoring procedures, we realized that APM is helpful for classifying students' performance and in choosing the passing score for a given purpose. The passing score must be judiciously chosen, though, and this usually requires several pilot tests to be performed.

Keywords: Computer Assessment System, English Proficiency Tests, Bloom Taxonomy, Admissible Probability Measurement.

Introduction

In non-English speaking countries, students enrolled in Master and PhD degree programs are usually required to undertake an English proficiency test (EPT), aimed at assessing their ability to understand and produce technical literature in English as it is the language used in most scientific journals. In practice, such EPT consists in asking students to translate a text passage from a scientific paper or a technical book from English (L2) into their mother tongue (L1), and in some cases to produce a version in L2 of another piece of technical text originally in L1. It is usual, as was the case of the Institute of Mathematics and Computer Science at the University of São Paulo (ICMC), Brazil, until April 1998, that the test is prepared each year by a different member of the academic staff. This is disadvantageous since distinct exams may be highly non-uniform due to the high degree of subjectiveness in the evaluation process. The alternative approach used by some Universities and funding agencies is to require the student to pass a general-purpose exam such as TOEFL (http://www.toefl.com) or IELTS (http://www.ielts.org). None of these exams, however, evaluate students' competence in terms of the demands of highly standardized research articles written in English. Furthermore, the need of "genre-consciousness" is not aroused, which is essential for a novice researcher to better and faster perform the reading and writing tasks for his/her own research. In order to obviate these limitations, the first author of this paper proposed a new type of proficiency exam for graduate students, in which students' competences would be evaluated according to four modules: M1: dealing with the analysis of the structure of a given section of a paper, where the students had to identify the section components; M2: dealing with the analysis of relationships among clauses which are signalled by discourse markers; M3: involving knowledge of conventions in the English language for scientific texts; M4: involving knowledge of writing strategies for each component of the sections' paper.

The four modules include questions that assess two cognitive levels from the well-known Bloom´s Taxonomy (Bloom, 1956), which divide cognitive objectives into six subdivisions ranging from the simplest behavior to the most complex: knowledge, comprehension, application, analysis, synthesis and evaluation. In addition to the knowledge category, normally used in multiple-choice tests, we included questions of two modules (M1 and M2) to assess competences in the fourth level of difficulty (analysis). In order to evaluate such abilities, particularly those for a higher cognitive level, we employed an information theoretical model of knowledge assessment to measure the student knowledge base. According to this model, a student may have total information, almost total information, partial information, partial misinformation, misinformation or total lack of information about a given topic being assessed. This approach makes use of a scoring procedure referred to as Admissible Probability Measurement (APM) developed by Shuford & Brown (1974) and employed in computerized exams by Bruno (1986, 1987) and Bruno, Holland & Ward (1988) and in manually marked exams by Klinger (1997).

As we shall comment upon later, this new proficiency exam was successful but relied on the work of a lecturer who is knowledgeable of a number of issues in scientific writing. This severe limitation could only be overcome if the exam was fully automated. This prompted us to develop a computer-based as-

sessment system in which the students' ability to read and write scientific literature in English was evaluated with objective questions and the APM scoring procedure. This computer-based system possesses a test management subsystem and a test delivery subsystem that are integrated into a single application. It is web-based and allows access to 5 different types of users: test administrator, instructor, students, master program secretary, and general public. In this paper we present the rationale to build this computer based system, coined CAPTEAP (http://www.nilc.icmc.usp.br/capteap/), which, in addition to fulfilling the requirements of conventional proficiency tests, takes the student one step further by testing his/her genre-consciousness, helping him/her to develop the schemata for the academic discourse, to spot a specific information in a paper and to understand the relations signalled by discourse markers. Also, we present the lessons learned from using this new computer-assisted English proficiency test that has been delivered since 2001 to students of a Master degree program of University of São Paulo, Brazil.

In Section 2, we describe how the system evolved from a manually-marked exam to a fully-automated one. Because CAPTEAP is based on our previous experience in developing writing tools for non-native English writers (Aluísio, Barcelos, Sampaio, & Oliveira Jr, 2001; Aluísio & Gantenbein, 1997; Aluísio & Oliveira Jr., 1995; Aluísio & Oliveira Jr., 1996; Aluísio Caldeira, De Oliveira, Fontana, Nacamatsu, & Oliveira Jr., 1992; De Oliveira, Aluísio Caldeira, Masiero, & Oliveira Jr., 1992; Fontana, Aluísio Caldeira, De Oliveira, & Oliveira Jr., 1993; Oliveira Jr., Aluísio Caldeira, & Fontana, 1992;), in Section 3 we present the writing tool that provided the framework for CAPTEAP. The decisions related to the assessment system which involves multiple choices and the CAPTEAP system itself are presented in Section 4. In Section 5 we discuss the lessons learned, which are grouped into 3 categories: those related with the content of the new test (its modules and the cognitive levels of the tasks to be accomplished); the ones associated with the interface of the web-based assessment system; and those associated with the scoring procedure employed. Section 6 concludes the paper indicating further steps in the development of CAPTEAP.

Recent English Proficient Tests at ICMC-USP

We strongly advocate that English proficiency tests (EPT) for graduate students should be based on technical writing studies rather than only on tests of vocabulary, reading comprehension and grammar. A considerable bulk of literature exists for helping students in technical writing and reading, e.g. (Gosden, 1995; Swales, 1990; Trimble, 1985; Weissberg & Buker, 1990). This literature was used in developing several writing tools for the AMADEUS suite (Aluísio et al., 2001; Aluísio & Gantenbein, 1997; Aluísio & Oliveira Jr. 1995; Fontana et al. 1993) on which the EPT tests to be described here was based. In order to carry out a test that considered the above issue, in 1998 the first author of this paper implemented a paper-and-pencil version of an EPT at the Computer Science Department of USP, São Carlos. In the test, one of the exercises was designed to assess the students' consciousness about the conventions in English usage in research papers. Guided (cued) exercises were designed, since they can help the student to get familiarized with the scientific discourse. The setup for the test was the following:

Goal to be met: how functional parts of a section of a research paper are arranged and how the linguistic expressions match to rhetorical purposes.
Instructional material presented: a description of a schematic structure of an Introduction (i.e. pieces of information and order they are usually sequenced in research papers).
Support material: the title, Abstract and the Introduction section of a paper in the students' respective research area: Computer Science, Statistics or Computational Mathematics.
The task: to identify which sentences correspond to the parts/stages of the schematic structure

The results were extremely encouraging because the students acknowledged that the test represented their first opportunity to analyze sections of a scientific paper from the perspective of the rhetorical divisions and their functions within the paper. Even though 24 out of 37 students passed the exam, the over-

all performance in some exercises was poor especially because the students lacked experience in technical writing and in reading instructions carefully. For instance, several students included material from the Abstract in this exercise, despite the clear instruction to treat only the schematic structure of the Introduction. Most students failed to realize how much background information is usually provided in an Introduction before the Purpose of the paper is stated, particularly in interdisciplinary fields. Segments identified by the students were normally much longer than expected, for they failed to relate the various stages of an Introduction. These observations have stimulated us to try and implement a type of test where detailed information is provided to the student on rhetorical divisions and their functions, writing strategies, language conventions used in scientific papers and discourse markers. For that use was made of writing tools we developed over the last 10 years, as explained in Section 3.

In order to apply the test on a regular basis with supervision of different lecturers, we replaced the paper version of the EPT by a computer-based system. This required the implementation of a bank of items and patterns made of available parts from the four modules mentioned in the Introduction. For the first prototype two computer-assisted assessment systems were evaluated - Question Mark for Windows (http://www.questionmark.com/), a popular commercial product, and a freeware package developed specifically for delivering language learning activities, the Hot Potatoes Suite (http://web.uvic.ca/hrd/halfbaked/). Although Question Mark is a well-recognized system for online assessment, we chose Hot Potatoes because it is a slim package with several question types and open code, ideal features for prototyping activities. Our results with Hot Potatoes are reported in Aluísio and Oliveira Jr. (1999). Difficulties with the interface and the need to provide detailed instructions to the students made us implement several changes. A brief history of the tests from April 1998 until September 2001 is given in Table 1, with the specific goals of each version, its format and the results.

Date	Objects	Format	Results
April 1998 (official exam)	Change the contents and focus of the EPT.	Traditional paper-and-pencil delivery and manual right-wrong scoring.	Students welcomed the opportunity to learn and analyze rhetorical divisions of scientific papers and their functions.
September 1999 (pilot test)	Investigate existing software for fully automated exams (Hot Potatoes).	Traditional paper-and-pencil delivery using manual right-wrong scoring and the Hot Potatoes questions type.	The need for better clarifying the tasks proposed by the questions. The use of off-shelf tools was ruled out due to security problems and need for customized scoring systems.
September 2000 (official exam)	Assess the APM scoring procedure.	APM in the traditional paper-and-pencil delivery and manual scoring.	Good reception of the method on the part of the students. The passing score, however, was the one used traditionally in our Institution, and therefore the real benefits of the APM method were not taken advantage of.
February 2001 (pilot test)	Investigate the impact of a fully-automated exam.	APM in a web-based exam: automatic delivery and marking.	The user interface was not sufficiently friendly, and appeared to interfere in the performance of the students.
April 2001 (official exam)	Investigate the remodelling of the interface.	APM in a web-based exam: automatic delivery and marking.	Several changes made at once caused anxiety on the students. Changes included: new contents, new scoring system in a fully-automated exam. The need of several improvements on the system.
September 2001 (official exam)	Ensure that the last changes made were sufficient.	APM in a web-based exam: automatic delivery and marking.	Students performed well with no interference from interface or difficulties with the novel procedures.

Table 1: History of EPTs at ICMC-USP.

Since April 2001, 171 students undertook the exams. Each student has 3 chances to pass the exam; if he/she fails after 3 chances he/she automatically fails the MSc Program at ICMC-USP. The present CAPTEAP version is the same since September 2001, which is the version to be described in Section 4, after we present the Support Tool that provides the background for CAPTEAP.

An Overview of the Case-based Support Tool

The Support tool is part of a learning environment for scientific writing named AMADEUS (http://www.nilc.icmc.usp.br/nilc/projects/amadeus.htm). AMADEUS targets the understanding of the processes involved in the creation of successful scientific papers, and encompasses advisory and tutoring tools. It was motivated by research (Fontana & Oliveira Jr., 1991) involving error categorization and analysing error gravity in a corpus comprising theses, articles and research reports written in English by Brazilian graduate students. The latter work suggested that localized errors do not interfere with communication as much as those which affect the global meaning of a text passage. By localized errors we mean basically spelling and grammatical mistakes. It was also concluded that providing users with input material from naturally occurring examples is more efficient in dealing with problems with inter-sentence relationships or with functional meanings in discourse, for the linguistic input material offers microelements within a context. Indeed, the reuse of linguistic material from a corpus has been employed in several systems, including report generators (Kukich, 1983; Smadja, 1991), hypertext-based support systems for software documentation (Born, 1992), case-based letter generators (Pautler, 1994), and document drafting tools (Branting, 1996; Paris, 1996). Within the AMADEUS Support tool, linguistic material is reused to improve the cohesion and coherence of Introductions of a scientific paper. Case-based reasoning (CBR) (Mantaras, 1995) is used to model the various stages of the writing process:

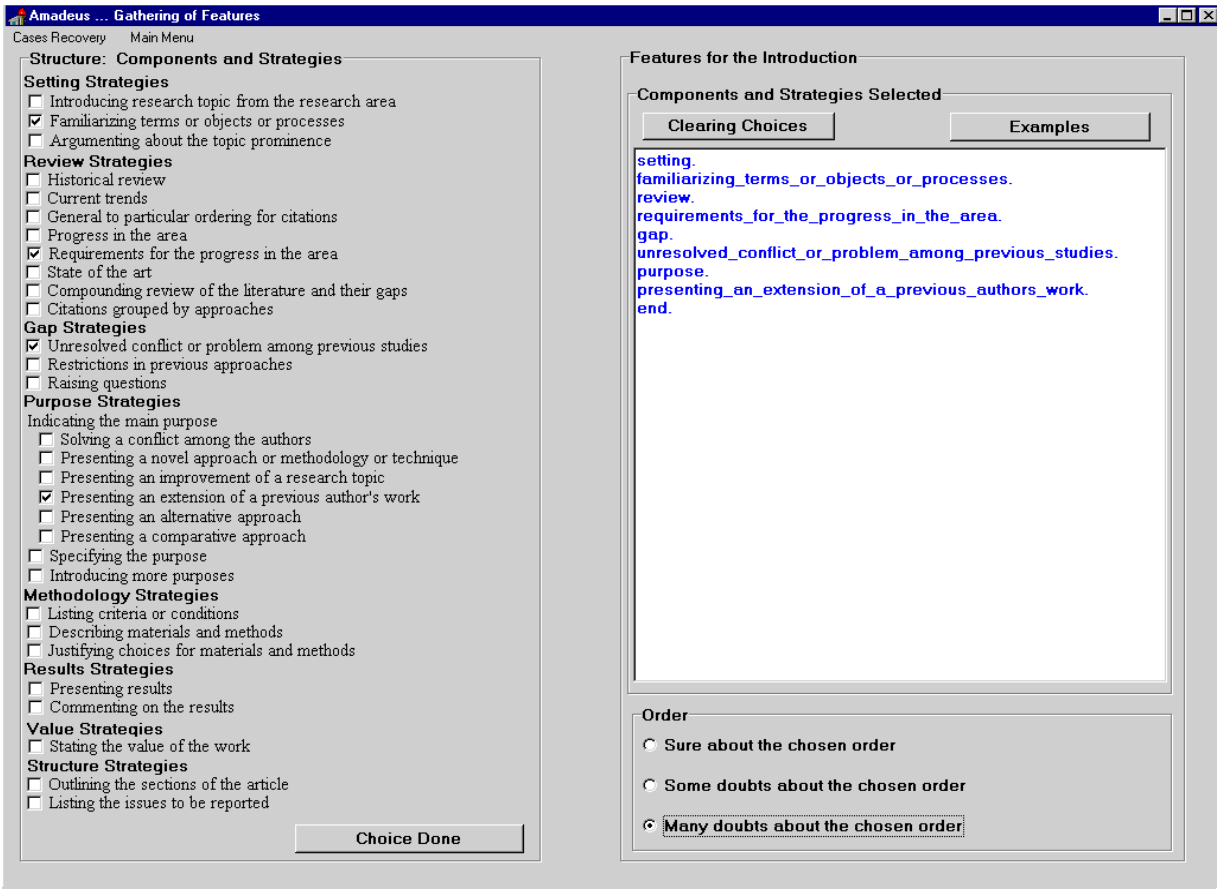


Figure 1: Gathering of structural features.

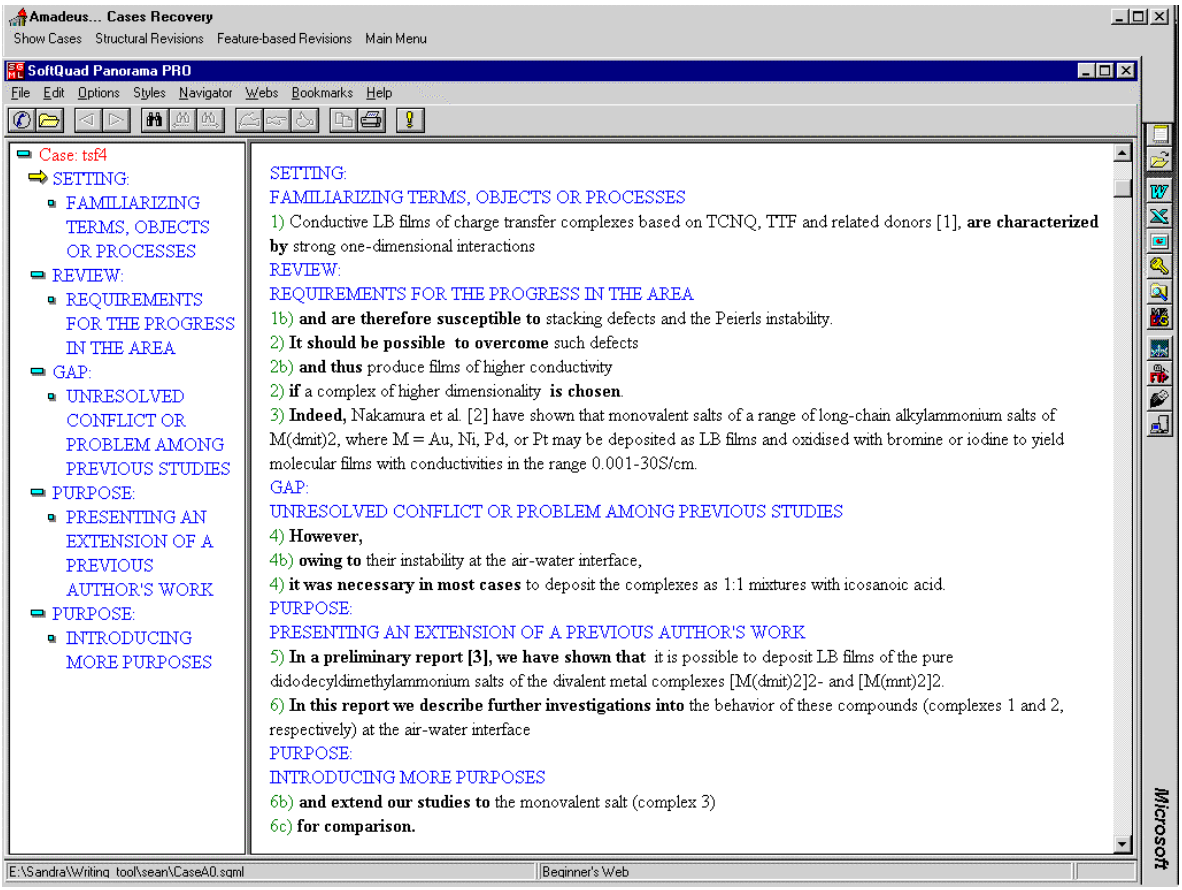


Figure 2: Case Recovery.

planning, drafting and self-reviewing (Hayes, 1980). Figure 1 shows the detailed schematic structure while Figure 2 displays an example of the rhetorical structures for the Introduction. In the Support Tool, these knowledge sources were applied to the three stages of the writing process. Accordingly, the user follows a three-step procedure: i) gathering of features, in which the user selects the features intended for his/her Introduction; ii) selection of the best-match case, following case recovery by the system; iii) revision/adaptation on the selected case in order to meet the user's purposes.

The screen dumps presented in Figures 1 and 2 illustrate the gathering and case recovery phases of the writing tool using a case base consisting of Physics papers. In Figure 1, the writer chooses the main components - setting, literature review, gap, purpose, methodology, main results, value of the research, and layout of the paper - and strategies - introducing research topic from the research area, historical review, listing criteria or conditions, for example - to be included in the introduction along with the writer's degree of certainty about their relative order in the text - sure about the order, some doubts, and many doubts about the order. Three types of pattern matching between the selected features and the available cases are used to recover the best cases: perfect match (equal lists), proper undermatch (sub-list) and non-proper undermatch (intersection). The tool selects cases to be returned to the user by evaluating various combinations of these three similarity metrics, with the particular combination used depending on the degree of certainty of the user about the order of the components and strategies. One of the cases recovered from the components and strategies selected in Figure 1 is shown in Figure 2.

In the Support Tool, revision/adaptation of the selected case is based on the method of Gosden for analysis of textual revisions in scientific texts using Systemic Functional Linguistics (SFL) (Gosden, 1995). The four categories of Gosden's revisions are:

1. addition of technical detail or statements [+TD]
2. deletion of technical detail or statements [-TD]
3. reshuffling of statements [R]
4. rhetorical machining [RM]

Rhetorical machining is subdivided according to three basic orientations:

1. rhetorical machining of discourse structure (by means of thematic (theme-rheme) and information (given-new) structures and cohesive relations) [RMd]
2. changes which relate to writers' claims and to writers' own research hypotheses and limitations work [RMc] (this also includes their research position to other published works)
3. rhetorical machining which relates to the writers' purpose, reasons for, results of research actions taken, and conclusions reached [RMp]

Figure 3 shows the revision names used in the Support Tool.

1. Addition of technical detail to illustrate ideas.
2. Deletion of technical detail to be used when the paper addresses knowledgeable audiences or in short papers.
3. Promotion of technical detail to allow syntactic variety and differences in emphasis or importance.
4. Demotion of technical detail to allow syntactic variety and differences in emphasis or importance.
5. Addition of text organisation markers (discourse markers) to link components together, improving the information flow.
6. Addition of discourse markers to make explicit the writer's claims about his/her research.
7. Use of attitudinal disjuncts to comment on the content of the communication.
8. Additions of discourse markers to make explicit the writer's research hypotheses or limitations.
9. Additions of discourse markers to make explicit the writer's research position in relation to other published work.
10. Addition of markers to make explicit the writer's purpose, the expression of reasons for, and the results of, research actions.

Figure 3: Revisions available in the Support Tool.

The potential revisions on a text may be shown by using markers from Gosden's framework. Figure 4 illustrates one of the possible ways to show revisions using the text previously presented in Figure 2. In this example, bold text represents reusable material, SFL-based revisions are highlighted in italics, while bold italics represent reusable material in the revisions.

CAPTEAP System: Partial Knowledge Evaluation in a Proficiency Test of English

The four modules of CAPTEAP and the resources taken from the Support Tool are shown in Table 2.

Currently, Module 1 uses the structure for the introduction section, which presents the 8 components shown in the left side of Figure 1, and the structure for abstracts presented in Weissberg and Buker (1990). As the Support Tool uses a case base consisting of Physics papers and we make use of Computer Science, Computational Mathematics and Statistics papers in the ETP test we only take the revisions shown in Figure 3 as examples to prepare questions for Module 2. The discourse markers used in the EPT are Contrastive (e.g. but, however, although), Elaborative signalling further information (e.g. also, in addition, moreover), an example (e.g. for example, for instance), a reformulation (e.g. that is, rather), Inferential signalling a consequence (as a consequence, as a result, therefore, thus), or a conclusion (then, as a conclusion), and Explanatory (e.g. because, since). For Modules 3 and 4 we employ messages

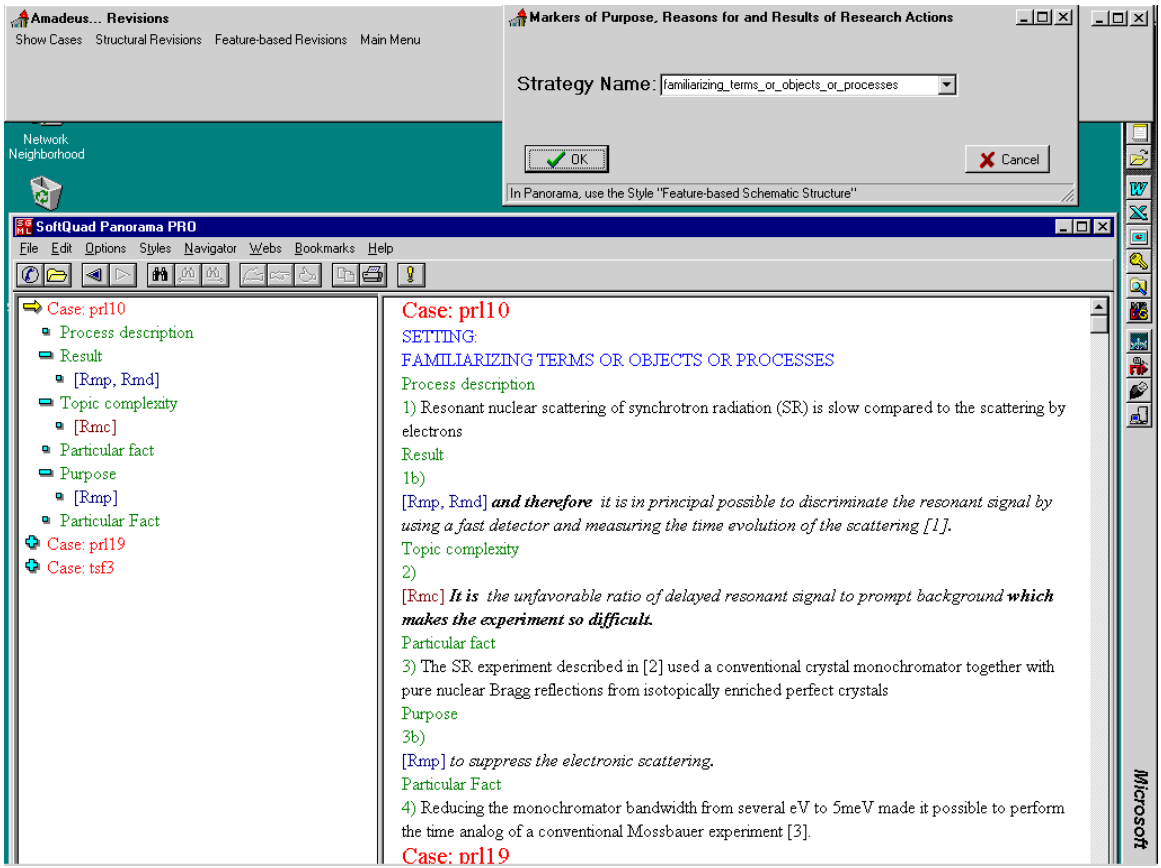


Figure 4: Displaying the SFL-revisions on an Introduction.

Module 1: analysis of the structure of a section in a scientific paper, with identification of its components	Schematic structure for research paper sections.
Module 2: analysis of relationships between clauses that are signaled by discourse markers	Examples from cases presenting the revisions 5, 6, 7, 8, 9, 10 (Figure 3)
Module 3: knowledge of conventions from the English language which are used in scientific texts	Messages used in the writing strategies
Module 4: knowledge of writing strategies employed in each component of a section	Writing strategies used in the components of the schematic structure.

Table 2: Resources from the Support Tool used in creating questions for EPT.

and writing strategies used in the Introduction Section of papers. For this section we have 30 strategies, also shown in Figure 1, and 45 different types of messages to compose strategies. The following writing strategies from the Review component illustrate the notation for the description in terms of messages or other strategies:

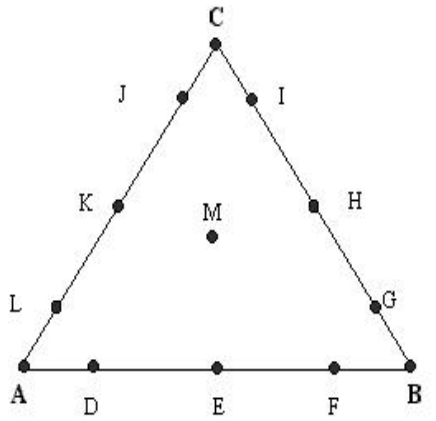
Strategy: **Requirements for progress in the area**
[*Topic Complexity*]
Requirements for progress
Support*

This can be interpreted as follows: the strategy Requirements for progress in the area comprises the optional message Topic Complexity, followed by the message Requirements for progress and by the Support strategy appearing zero or more times. The Support strategy is defined as:

(*Motivation / Cause / Result / Purpose / Evidence / Particularisation / Exemplification)*

The formation rule for strategies makes the user aware of the kind of information a specific genre is supposed to contain. Moreover, optional messages and messages appearing zero or more times may be deleted or included if necessary, since they usually carry details. Just to show the power to generate different types of exercises for modules 3 and 4 using the resources of this writing tool, and considering only the schematic structure for the introduction section, we can make 75 different types of question since the 8 components of the Introduction are made up of 30 writing strategies made with 45 different messages.

A crucial issue in developing effective computerized tests is to understand the potential and limitations of objective tests, which require the user to provide answers to questions with pre-defined answers. This type of test is easy to correct computationally, but promotes guessing. However, strategies can be used to reduce the probability of the student being successful by only guessing instead of employing his/her knowledge on the topic. Such strategies may stimulate students to consider their choices more critically and improve critical thinking skills. For example, Admissible Probability Measurements use ideas of risk and loss which guide students to answer with their true beliefs. The method consists in distributing the options of answers (3 options are offered, A, B and C) in an equilateral triangle (Figure 5). The triangle presents also intermediate points that may be chosen by the user in case he/she is in doubt between 2 options. In the total, there are 13 possible answers (A-M). The student is classified according to the space of the selected answers in relation to the correct answer. Each of the 13 answers can readily be understood as subjective probability statements, and the practical effect is that 13 responses enable converting 3 alternative questions into a means to evaluate whether complex material has been learned (Klinger, 1997). This is precisely the mechanism we were looking for to assess the higher levels of cognition in Bloom's taxonomy.



To answer the questions you should choose one of the thirteen options presented in the triangle at the left side. In the vertexes are the three options for the answers (A, B, C). The remaining letters (D to M) are used to express uncertainty about the options A, B, or C.
If totally sure use a definite **A, B or C** response. If totally unsure use an **M** response. If you think that one option is definitely wrong (**A, B or C)**, choose among the remaining five on the opposite line. To indicate preference between two values, choose **D, F, G, I, J** or **L**. To indicate uncertainty between two values choose **E, H,** or **K**.

Figure 5: Instructions of an EPI using CAPTEAP.

To motivate students to associate their answers with probabilities instead of using wild guesses, rewards were given to students for exact associations with the correct answer while inconsistent associations were penalized. Note, in Figure 5, that a student may choose M if he/she thinks that the probability of the options A, B and C is the same. The penalties tend to be larger than the rewards to prevent guessing. According to Shuford & Brown (1974), the students only give up guessing when the penalty associated with a wrong answer is larger than k-1, where k is the number of alternatives. As students are tested fre-

quently with the APM method, they get used to expressing their beliefs in terms of probabilities, and their scoring reflects their actual knowledge on the topic (Klinger, 1997).

The APM method uses the terms totally informed, nearly informed, partially informed, misinformed, partially misinformed and uninformed to describe the students' knowledge on a particular subject (Figure 6). According to Bruno et al. (1988) the instructional strategies associated with the information states for a student are: re-education if misinformed; advancement if totally informed; review if nearly informed and partially informed; and instruction if partially misinformed and uninformed. If one student has a misconception, a larger effort is required on the part of the student and of the Institution, because the wrong concept must be made explicit and then the right concept be learned by the student.

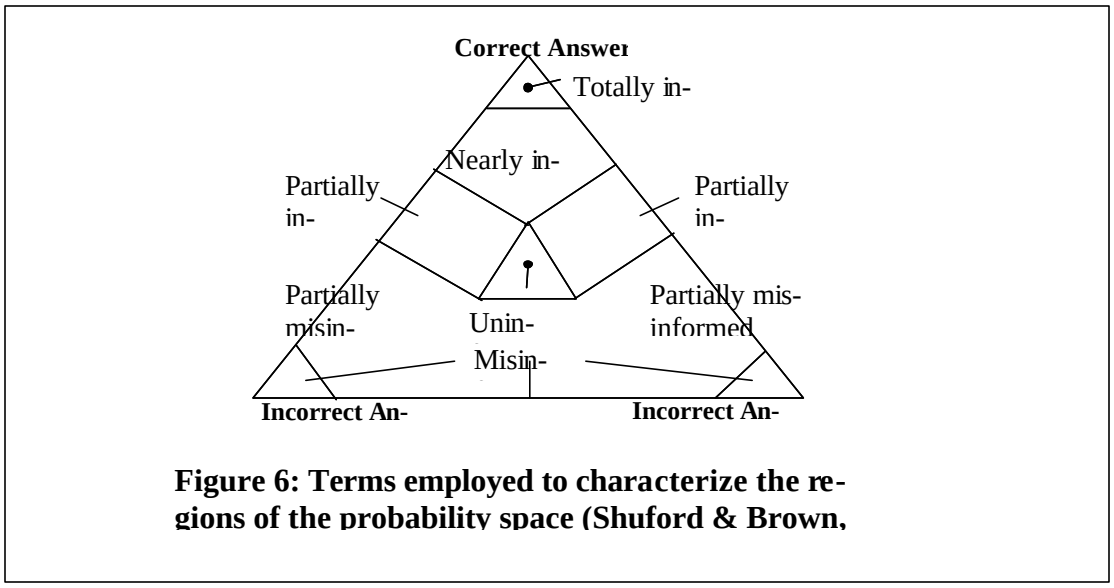


Figure 6: Terms employed to characterize the regions of the probability space (Shuford & Brown,

Figure 5 shows the triangle used in CAPTEAP and the instructions for the students given in the beginning of the exam. Figure 7 shows two screen dumps of a test containing 7 parts. On the left side appears the text related to Question 1 of Part 1 and on the right side is seen one of the four questions from Part 1. The student has direct access to each and every question of the test, i.e. he/she can jump from one question to the other or consult a table of contents to check unanswered questions.

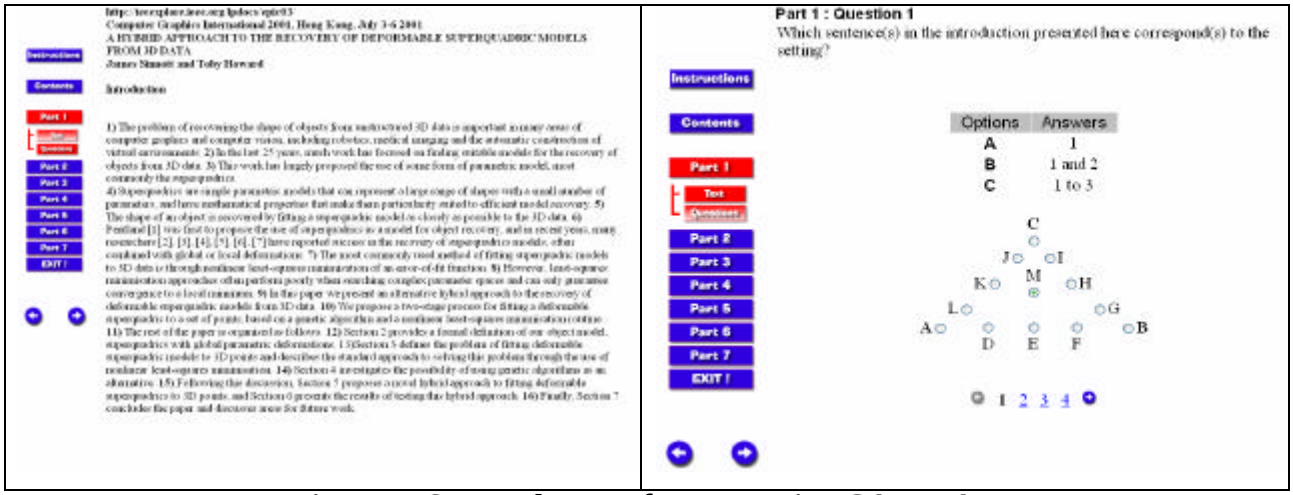


Figure 7: Screen dumps of an EPI using CAPTEAP.

Lessons Learned

The main purpose of this section is to report several problems and solutions employed to solve them while applying EPT tests since 1998, particularly after February 2001 when the problems associated with automated tests could be identified (see Table 1). It is hoped that by sharing our experience we may contribute to those intending to introduce a computer based assessment system in an Institution that only uses traditional paper-and-pencil assessment methods. In this context, we took advantage of McKenna & Bull's (1999) work and the online tutorial found in http://www.caacentre.ac.uk/resources/ on the design of effective objective test questions and of (Valenti, Cucchiarelli, & Panti, 2002) work that identifies a set of quality factors to evaluate Computer based assessment systems. The several lessons learned are divided below according to the following topics: interface of the web-based system, the modules and cognitive levels of the tasks of the tests, and scoring procedure.

Interface

The first web-based exam was carried out in February 2001, where the interface followed the design used in the Hot Potatoes package, i.e. different frames to deliver questions, reading material and instructions. This interface was evaluated by collaborators (graduated students and three experts in evaluating human-computer interfaces) and several problems were identified, particularly those related to overloading of information on the screen and the large number of clicks required to answer one question. The identification of these problems was made using the heuristic evaluation method (Nielsen, 1993), a kind of inspection-based evaluation and pilot testing with real users. Fifty-eight problems were detected in the usability of the interface, 4 of which were frequent:

1. There was no detailed report on the performance of students - the only piece of information given to the student was the passing score
2. Sequential access to the questions - no direct access to those question not yet answered
3. Difficulty in reading large chunks of text on the computer screen
4. Difficulties with questions that used small chunks of text that did not appear with the questions on the same screen.

With the present interface, shown in Figure 7, all the problems identified have been addressed. For example, Instructions and Support Material to answer the questions have been put together and the large chunks of text are presented alone in the screen; there is a content table allowing direct access to unanswered questions and the exam parts shown at left in Figure 7 also allow direct access to the questions. As we will see later, at the end of the exam the student receives a report on its performance and, in addition, there are links for the correct answers of the exam and for the criteria of evaluation. Besides, the instructions are presented in Portuguese in order to make sure the students understand the task to be carried out.

Lessons learned

Evaluation of interfaces by real users combined with less costly usability assessment procedures, such heuristic evaluation, has proven important to guide the development of a user-friendly interface. In formal exams such as those carried out with CAPTEAP, users are unlikely to tolerate system faults as the latter may affect their performance in an official test.

Contents of the Test

The initial impact of the EPT exams on the students was considerable due to the several changes introduced at once. Changes were made in the focus of the test contents, on the scoring system and on the very nature of the test that became a computer-based exam. The performance of the students in the first test was below the expectations and students complained about the anxiety brought by the unprece-

dented procedures for an English exam. We have therefore decided to provide work material for the students to practice before applying for the exam, which was done through a website that contains detailed instructions and simulated tests. With regard to the contents of the test, the Exam in April 2001 contained a question of the analysis category in Bloom's taxonomy, which was considered the most difficult one by the students. It consisted in rearranging passages of a given text that had been deliberately scrambled by the computer system. Such rearrangement required detailed knowledge of the structure of one of the sections of a paper. Another question in the evaluation category required students to provide a summary of a given text passage by applying summarizing strategies. The category evaluation in Bloom's taxonomy requires the student to judge the value of a material for a given purpose; in our setting we used this category requiring the student to judge whether a piece of text is to be discarded or not in order to summarize the given text. We have eliminated these two tasks and therefore in the latest exams CAPTEAP concentrates on the reception of English texts, which is the most essential activity of a novice scientist. The present modules of the test have two category types in Bloom's taxonomy: the first category is related to the first of the 3 lowest levels (knowledge) and the second one is associated with the first of the 3 highest levels (analysis).

Lessons learned

Bloom's taxonomy allows one to select questions in a balanced way to test different abilities of the students. It is important to provide students with the opportunity to learn how the test works beforehand and also to practice in simulated exams.

Scoring Procedure

We first analyzed the scoring procedure adopted by Klinger (1997), in which the range of points [-100, 30] to be awarded corresponding to the 13 possible answers is transformed into fractions, as shown in Table 3. For correct answers, which could be A, B or C with associated 30 points, the score will receive 1 for each answer. If the student is uninformed (M) the score receives 0.769 for each question, while if the student chooses the wrong alternative it receives zero. Note that if, for example, A is the correct answer, wrong alternatives are B, C and the ones between A and B, i.e. J, K, L. The fractions for the choices "Near right", "Between two" and "Near wrong" are also given in Table 3.

According to Bruno et al. (1988) those who have achieved mastery (APM score of 95 or, on the average, at least 75% confident and correct) can be assigned to enrichment or advanced topic sessions; those who scored between 85 and 95, or on average between 50% and 75% confident and correct, can be assigned to review and instruction sessions. Those who score below 85 need remedial sessions. It should be noted that a student who admits knowing nothing about the topic will have a score of 76.9%.

In the exam in April 2001, the passing score followed the mastery criteria mentioned above. This turned to be too severe, with only 3 out 63 students passing the exam. The result was not surprising though, because in our University tests are designed to assess students with a passing score of 50% with standard

Choice	Letter	Points	Fractions
Correct	A B C	30	1.000
Wrong	-------	-100	0.000
Uninformed	M	0	0.769
Near right	D F G I J L	20	0.923
Between two	E H K	10	0.846
Near wrong	D F G I J L	-10	0.692

Table 3: Values between A and M indicate belief.

questions (not multiple choices). We then adopted other criteria based on a more global analysis of the results, as considerable information could be taken into account in addition to the APM score information. For example, Bruno et al. (1988) presents to students their performance reports considering the following pieces of information: a) the information use score which corresponds to the percent time correct when using an A, B, C response; b) the percent overall certainty in the exam which corresponds to the percent A, B, C responses used in test of the total possible responses; c) the percent correct certainty in the exam which corresponds to the percent A, B, C responses that were correct of the total possible responses; d) the percent uncertainty which corresponds to the percent D to M responses of the total possible; e) the percent lack of information which corresponds to the percent M of total reflecting no information or lack of time to finish exam. Klinger (1997) established a classification based on the sum of the correct and near correct answers (C + NC) and on the sum of the wrong answers with the answers involving no preference among the three alternatives (W + M values). Three classes of students were identified with such criteria: i) those who achieved the objectives, ii) those with under-achievement performance and iii) those at the borderlines according to standard deviations. We employed a similar procedure (see results in Figure 8), with the passing score as follows. In terms of a global assessment, students pass the exam if they have:

a) 50% or above of their answers in the class “totally informed” and 25% or less in the class “misinformed”.

Or

b) 90% or above of answers in the classes “totally informed”, “nearly informed” and “partially informed” and 10% (or less) of answers in the class “misinformed”.

For the students that were very close to the passing score in the global assessment, i.e. 1 to 2 questions differing from the expected score a reassessment was carried out as follows. The student would pass if in the module dealing with “the structure of a scientific paper” he/she was eligible to pass according to the criteria adopted for the global assessment. This gives a heavier weight to one of the 4 modules (M1). In April 2001, 63 students undertook the exam, 13 of which passed based on the criterion a), 1 based on criterion b) and 11 based on the reassessment.

Figure 8 shows a screen dump with the number of questions per module (first column of the table at the top) and category of APM (second to seventh columns of the table at the top). The output screen for the simulated tests also displays a table of relative performance in comparison with previous exams done by the user. There are also links for the correct answers of the exam (in the link “Clique para ver o Gabarito”) and for the criteria of evaluation (in the link “Clique para ver o Critério de Aprovação”). The student passed the exam according to the criterion a), as she obtained 16 out of 27 answers (59%) in the class totally informed and 6 questions (22%) in the class misinformed.

After April 2001, the rate of students approved has been 85%, which means that the students are increasingly better prepared for the test. It is likely that the test has become easier than one could wish, though the approval rate has fallen to 58% in the latest EPI in April 2003. We are therefore reviewing the criteria for classifying students. We may reintroduce the task with scrambled text considering small texts like abstracts instead of large sections as before, and also consider increasing the penalty for wrong answers as advocated in the mastery criteria by Bruno et al. (1988). The changes will only occur after we have more data to be analyzed.

Lessons learned

APM is extremely helpful for classifying students’ performance and in choosing the passing score for a given purpose. The passing score must be judiciously chosen and this usually requires several pilot tests to be performed. As consecutive tests are applied and the students get used to the new procedures, APM is widely accepted. It is important to spell out all the criteria for evaluation, including passing score, to

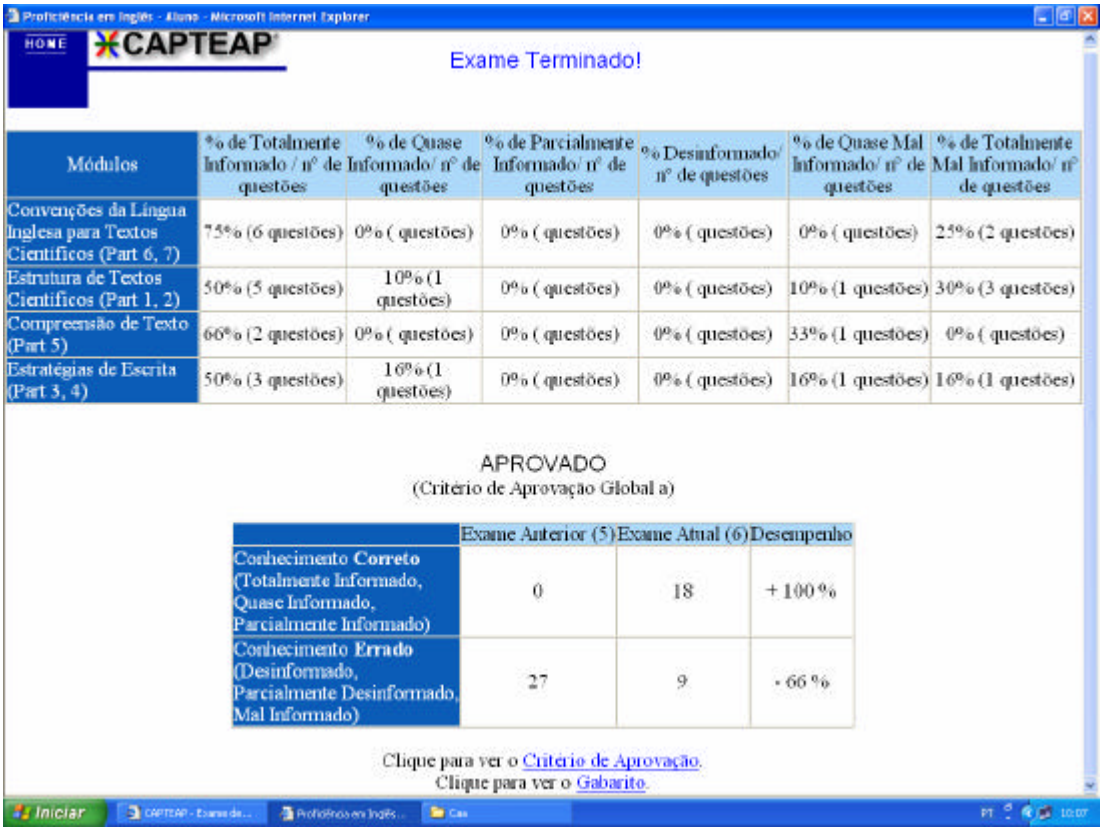


Figure 8: Performance report of a test by a student practicing for the official exam.

the student before the test. Reports of students' performance should be made available for students and instructors.

Conclusions and Further Work

We hope to have demonstrated that focusing on academic writing is possible for an English Proficiency Test, though caution is necessary in the introduction of new concepts. The several changes made over the 5-year period, since the first paper-and-pencil version of EPT, illustrate the need to search for a user-friendly interface and to tune scoring procedures to match the required abilities from the students. Of particular importance was the availability of resources from the Support Tool, which was specifically developed for non-English speakers, precisely the target users of CAPTEAP. After two years of stability of CAPTEAP in terms of procedures and tasks, with ample students' satisfaction who are able to practice and learn the topics to be evaluated in simulated tests, we are now reviewing the test contents and scoring procedures. In two independent endeavors we will a) increase the bank of questions by applying automatic rhetorical classification for components of scientific papers following Teufel & Moens (1998) and Teufel, Carletta, and Moens (1999), and b) create an integrated system composed by a diagnostic assessment module and the support writing tool (presented in Section 3) to assist users in identifying and overcoming their weak points.

Acknowledgements

The authors acknowledge the financial support from CNPq.

References

Aluísio, S.M., Barcelos, I., Sampaio, J., & Oliveira Jr., O.N. (2001). 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, 257-260.

Aluísio, S.M. & Gantenbein, R.E. (1997). Towards the application of systemic functional linguistics in writing tools. *Proceedings of International Conference on Computers and their Applications*, 181-185.

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

Aluísio, S.M. & Oliveira Jr,. O.N. (1996). A detailed schematic structure of research papers introductions: an application in support-writing tools. *Revista de la Sociedad Espanyola para el Procesamiento del Lenguaje Natural, 19*, 141-147.

Aluísio, S.M. & Oliveira Jr., O.N. (1999). An Innovative computer assisted proficiency test of English for academic purposes. In: Proceedings of the Third Annual Computer Assisted Assessment Conference, Loughborough University, 1999, 21-36.

Aluísio Caldeira, S.M., De Oliveira, M.C.F., Fontana, N., Nacamatsu, C.Y., & Oliveira Jr., O.N. (1992). Writing tools for non-native users of English. *Proceedings of the XVIII Conference LatinoAmericana de Informatica*, 224-231.

Bloom, B.S. (Ed.) (1956). *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*. New York; Toronto: Longmans, Green.

Born, G. A. (1992). Hypertext-based support aid for writing software documentation. In P. O'Brian-Holt and N. Williams (eds), *Computers and Writing - State of the Art* (pp. 266-277). Dordrecht: Kluwer Academic.

Branting, L.K. & Lester, J. (1996). A framework for self-explaining legal documents. In *Proceedings of the Ninth International Conference on Legal Knowledge-Based Systems (JURIX-96)*, Tilburg University, the Netherlands, 77-90.

Bruno, J.E. (1986). Assessing the knowledge base of students: An information theoretic approach to testing. *Journal of Measurement and Evaluation in Counseling and Development, 19* (3), 116-130.

Bruno, J.E. (1987). Admissible probability measures in instructional management. *Journal of Computer-Based Instruction, 14* (1), 23-30.

Bruno, J.E., Holland, J.R. & Ward, J.W. (1988). Enhancing academic services for special action students: An application of information referenced testing. *Journal of Measurement and Evaluation in Counseling and Development, 21* (1), 5-15.

De Oliveira, M.C.F., Aluísio Caldeira, S.M., Masiero, P.C., & Oliveira Jr., O.N. (1992). A discussion on human computer interfaces for writing support tools. *Proceedings of the XII International Conference of the Chilean Computer Science Society*, Chile, 223-233.

Fontana, N., Aluísio Caldeira, S.M., De Oliveira, M.C.F., & Oliveira Jr., O.N. (1993). Computer assisted writing--Applications to English as a foreign language. *CALL, 6* (2), 145-161.

Fontana, N. & Oliveira Jr., O.N. (1991). O texto acadêmico em Inglês como língua estrangeira - dificuldades e perspectivas. *Atas do IX Simpósio Nacional de Ensino de Física*, São Carlos, SP, 571-576.

Gosden, H. (1995). Success in research article writing and revision: A social-constructionist perspective. *English for Specific Purposes, 14* (1), 37-57.

Hayes, J.R. & Flower, L.S. (1980). Identifying the organization of writing processes. In L.W. Gregg & E. R. Steinberg (eds.), *Cognitive Processes in Writing*, (pp. 3-30). Hillsdale, N.J: Erlbaum.

Klinger, A. (1997). Experimental validation of learning accomplishment. In: *1997 ASEE/IEEE Frontiers in Education Conference*. Pittsburgh, Pennsylvania. Retrieved May 22, 2003 from the World Wide Web http://fie.engrng.pitt.edu/fie97/papers/1271.pdf

Kukich, K. (1983). *Knowledge-based report generation: A knowledge engineering approach to natural language report generation*. PhD Thesis, University of Pittsburg.

Mantaras, R.L. & Plaza, E. (1995). Case-based reasoning. *The Newsletter of the European Network of Excellence in ML*, Special Issue, September, 29-37.

McKenna, C. & Bull, J. (1999). Designing effective objective test questions: an introductory workshop. In *Proceedings of the Third Annual Computer Assisted Assessment Conference*, Loughborough University, 253-257.

Nielsen, J. (1993). *Usability engineering*. Academic Press.

Oliveira Jr., O.N., Aluísio Caldeira, S.M. & Fontana, N. (1992). Chusaurus: A writing tool resource for non-native users of English. In Ricardo Baeza-Yates and Udi Manber (eds), *Computer Science: Research and Application* (pp. 63-72). New York: Plenum Press.

Paris, C.L. & Van der Linden, K. (1996). Drafter: An interactive support tool for writing multilingual instructions. *IEEE Computer*, Special Issue on Interactive Language Processing.

Pautler, D. (1994). Planning and learning in domains providing little feedback. In *AAAI Fall Symposium on Planning and Learning Notes'94*, Technical Report FS-94-01, 126-131.

Smadja, F. (1991). *Retrieving collocational knowledge from textual corpora. An application: Language generation*. PhD Thesis, Computer Science Department, Columbia University.

Shuford, E. H. & Brown, T. A. (1974). Rationale of computer-administered admissible probability measurement. *Technical Report R-1371-ARPA*. Rand - Santa Monica. CA.

Swales, J. (1990). *Genre analysis - English in academic and research settings*. Cambridge, UK: Cambridge University.

Teufel, S., Carletta, J. & Moens, M. (1999). An annotation scheme for discourse-level argumentation in research articles. In *Proceedings of the Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL'99)*, 110-117.

Teufel, S. & Moens, M. (1998). Sentence extraction and rhetorical classification for flexible abstracts. *AAAI Spring Symposium on Intelligent Text Summarization*, Stanford, March 1998, Technical Report SS-98-06, 16-25.

Trimble, L. (1985). *English for science and technology: A discourse approach*. Cambridge University Press.

Valenti, S., Cucchiarelli, A., & Panti, M. (2002). Computer based assessment systems evaluation via the ISO9126 quality model. *Journal of Information Technology Education, 1* (3), 157-175.

Weissberg, R. & Buker, S. (1990). *Writing up research - Experimental research report writing for students of English*. Prentice Hall Regents.

Biographies

Sandra Maria Aluísio is an Assistant Professor at the Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, Brazil. She received a BSc. in Computer Science from the Federal University of São Carlos in Brazil, a MSc. degree in Computer Science from the University of São Paulo in Brazil and, in 1995, a Ph.D. degree in Computer Science also from the latter university. Her main interests are natural language processing, technical writing, computer assisted assessment (CAA), computer aided writing (CAW), knowledge acquisition from large corpora and information retrieval systems. She has been involved in research projects supported by Brazilian agencies such as the National Research Council (CNPq) and the Research Council of the São Paulo state (FAPESP) at the Interinstitutional Center for Research and Development in Computational Linguistics (NILC).

Valéria Tomas de Aquino is a Lecturer at the Centro Universitário Barão de Mauá, Ribeirão Preto, Brazil. She received a BSc. in Computer Science at UNOESTE, in Presidente Prudente (Brazil) and a MSc.degree in Computer Science at the Instituto de Ciências Matemática e de Computação, Universidade de São Paulo, in São Carlos, Brazil.

Rafael Pizzirani is an undergraduate student of Computer Science at the University of São Paulo, in São Carlos, Brazil.

Osvaldo N. de Oliveira Jr. is an Associate Professor at the Instituto de Física de São Carlos, Universidade de São Paulo, Brazil. He got a PhD at the School of Electronic and Engineering Science at the University of Wales, Bangor (UK) in 1990. At the Interinstitutional Center for Research and Development in Computational Linguistics (NILC), he has worked on software writing tools and natural language processing. He has also worked in nanoscience and nanotechnology, with development of novel materials in the form of ultrathin films.