

## Authoring of Adaptive Computer Assisted Assessment of Free-text Answers

**Enrique Alfonseca, Rosa M. Carro, Manuel Freire, Alvaro Ortigosa, Diana Pérez and  
Pilar Rodríguez**

Computer Science Department, Universidad Autonoma de Madrid  
Carretera de Colmenar Viejo, Km. 14'5, 28049 Madrid, Spain  
Enrique.Alfonseca@uam.es  
Rosa.Carro@uam.es  
Manuel.Freire@uam.es  
Alvaro.Ortigosa@uam.es  
Diana.Pérez@uam.es  
Pilar.Rodriguez@uam.es

### ABSTRACT

Adaptation techniques can be applied not only to the multimedia contents or navigational possibilities of a course, but also to the assessment. In order to facilitate the authoring of adaptive free-text assessment and its integration within adaptive web-based courses, Adaptive Hypermedia techniques and Free-text Computer Assisted Assessment are combined in what could be called Adaptive Computer Assisted Assessment of Free-text Answers. This paper focuses on the integration of this type of assessment within adaptive courses, as well as on the presentation of an authoring tool able to manage the insertion and modification of different question statements and reference answers for open-ended questions. The results of the evaluation of this tool with course authors show the feasibility of proposing and evaluating open-ended questions adapted to each student, as well as that of getting a better model of the student's progress.

### Keywords

Authoring, Adaptive hypermedia, Computer-assisted assessment, E-learning, Free-text answers

### Introduction

Adaptive hypermedia has been widely used for the development of adaptive Web-based courses, in which each student is individually guided during the learning process (Brusilovsky, 2001). Most of these systems obtain feedback from the student from two sources: their behaviour when browsing the course (e.g. pages visited, time spent in each page, or navigational path followed) and the result obtained when answering test questions (e.g. true-false, multiple-choice or fill-in-the-blank). Some authors have expressed their concern that this limited way of assessment may not be really measuring the depth of the student learning (Whittington & Hunt, 1999). This fact has been the motivation of the field known as Computer-Assisted Assessment (CAA) of student essays. This is a long-standing problem that has received the attention of the Natural Language Processing research community. There are many possible ways to approach this problem, including: a study of the organization, sentence structure and content of the student essay such as in E-rater (Burstein et al., 2001); pattern-matching techniques such as in the Intelligent Essay Marking System (IEMS) (Ming et al., 2000); or Latent Semantic Analysis such as in the Intelligent Essay Assessor (IEA) (Laham et al., 2000). In Valenti et al. (2003) a state-of-art of CAA systems is described.

In order to support adaptive Web-based teaching and learning, we have developed the TANGOW system, which supports the specification and dynamic generation of adaptive web-based courses, so that the course components are tailored to each student at runtime (Carro et al., 1999; Carro et al., 2003). We have also developed, independently, a CAA system called Atenea (Pérez et al., 2004). It is based on n-gram co-occurrence metrics (Papineni et al., 2001), which allow the system to perform a vocabulary analysis and to study how similar student and teacher answers are. In Atenea, these metrics are combined with shallow natural language processing techniques, such as removing meaningless words, identifying the sense of polysemous words, or looking for synonyms in order to cover as much paraphrasing as possible. TANGOW and Atenea can work individually, but their potential can be much higher if they are integrated. TANGOW-based courses, as the majority of AH-based courses, rely on objective testing to evaluate the student knowledge, and this might not evaluate the higher cognitive skills. Therefore, by adding the possibility of evaluating open-ended questions we intend, on the one hand, to improve the courses generated by TANGOW and, in general, the quality of evaluation processes in distance learning. On the other hand, information about the user can be used by Atenea for adaptation purposes.

After the integration of both systems, Atenea has access to the information about the users, kept by TANGOW in the student model. It uses this information to adapt the assessment offered to each student. Both static and dynamic parameters of the user model can be used for the adaptation. For example, the web pages can be presented in the student's language, and the next question statement, as well as its specific version (style) can be dynamically chosen depending on both the student's features and how well the previous questions were answered, so that the level of the training is not too difficult (that could cause rejection from students that are not able to answer anything) or too easy (that could cause boredom to students who do not really learn anything new). TANGOW, on its part, uses the feedback from the open-ended question activity done by the student to perform the subsequent adaptation during the rest of the student learning process.

Up to our knowledge, there are no previous systems that support this type of integration. However, this work could represent the integrated evolution of related fields such as Computer Adaptive Testing (CAT) that relies on statistical measures to modify the order in which the test items are presented to the students according to their performance during the test (Linden and Glass, 2000). One example of CAT system is SIETTE (Guzmán and Conejo, 2002) that is currently being used in the University of Málaga in Spain. Another example of adapting the assessment of objective testing is presented in AthenaQTI (Tzanavari et al., 2004), based on the use of the QTI standard. Other approaches consist in adapting the course and the MCQ evaluation section, so that a course would be no longer a simple set of learning items but a complex structure with several branches able to recommend the optimal one for a user or for a class of users (Cristea and Tuduce, 2004); in adapting the feedback provided to the students (Lutticke, 2004), the problem selection (Mitrovic and Martin, 2004; Chou, 2000), the order of the problems (Gutiérrez et al., 2004; Sosnovsky, 2004); and, even in some cases, to re-adapt the adaptive assessment according to the student objectives (Panos et al., 2003).

The paper is structured as follows: in the section titled "TANGOW", we briefly review the main features of this system; in the "Atenea" section, we describe the architecture and performance of this system; the integration of both systems is explained in the "Integration" section; the authoring tool is presented in the next section; and, in the last section, some conclusions are drawn and future work is shown.

## TANGOW

The TANGOW (Task-based Adaptive learner Guidance on the Web) system delivers adaptive websites and, particularly, adaptive web-based courses, and has evolved significantly since (Carro et al., 1999). Courses delivered by TANGOW are composed of several activities or tasks that can be accomplished by the students. A task can correspond to a theoretical explanation, an example, an exercise to be done individually, or an activity to be performed collaboratively (problems to be solved, discussions, etc). The set of available tasks is constantly updated, tracking changes in the student's profile (static features and dynamic actions). Once a task is chosen by a student, the system generates the corresponding web pages by selecting, among the content fragments related to the task (or the set of problem statements or collaborative tools available, in the case of practical or collaborative tasks), those that provide the best possible fit to the student profile.

The whole processes for both developing adaptive web-based courses and automatically delivering them are described in (Carro et al., 2002). The rule-based formalism that facilitates the description of a course, including the specification of the activities, their alternative organisation for different students, diverse teaching strategies and adaptation capabilities, has been extended to support collaboration activities and group management (Carro et al., 2003).

Figure 1 shows two pages generated by the TANGOW system. The first one corresponds to an individual activity (an example to complement the theory), while the second one is a collaborative workspace dynamically generated to support the interaction of a group of students with visual learning style while working together in the resolution of the problem presented.

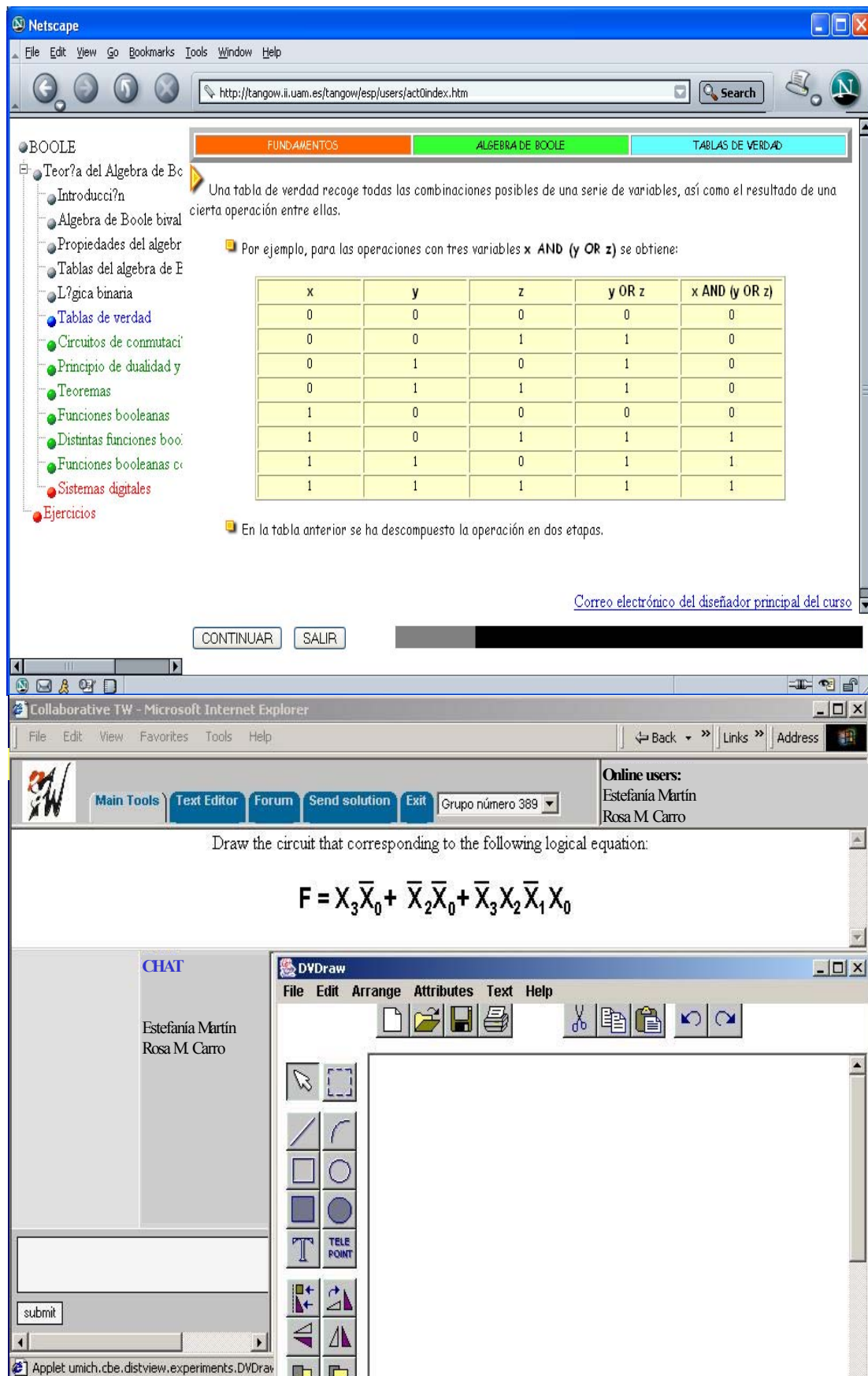


Figure 1. Examples of TANGOW interface

## Atenea

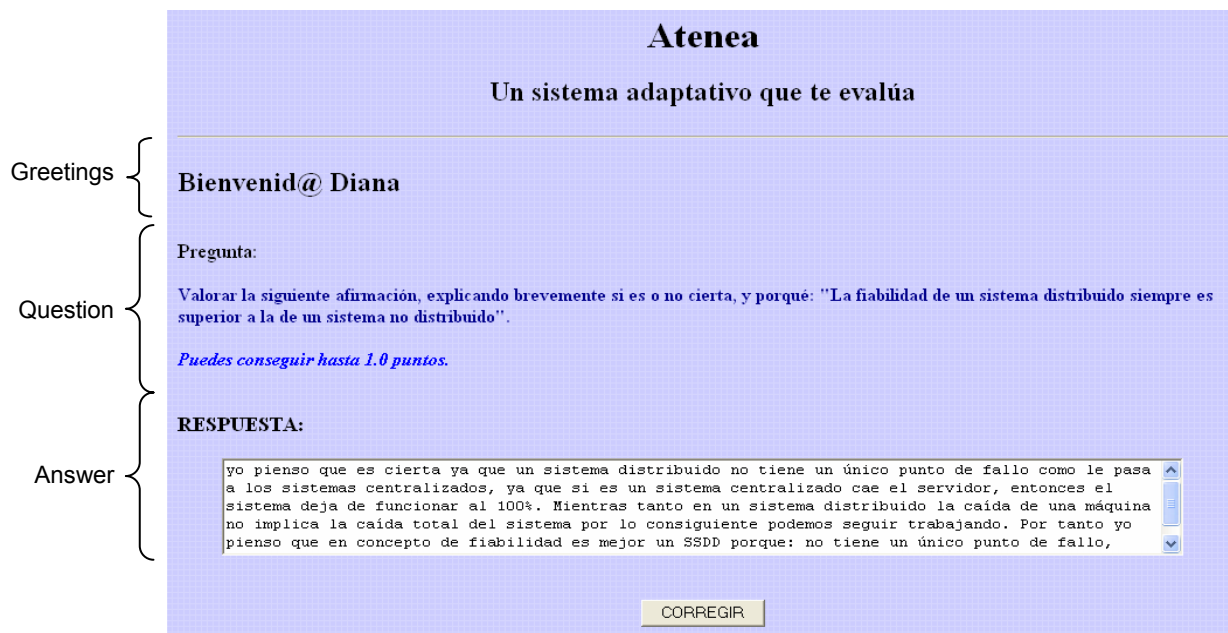


Figure 2. Interface of Atenea

Atenea (Pérez et al., 2004a; Pérez et al., 2004b) is a Computer-Assisted Assessment system for automatically scoring students' short answers. Its name is the Spanish translation of Athene or Athena, the Greek goddess of wisdom. Atenea relies on the combination of shallow natural language processing (NLP) modules (Alfonseca et al., 2004) and statistically based evaluation procedures. It has been coded in Java as a stand-alone application, but it has also an on-line version that can be accessed from any web browser connected or not to the Internet. Figure 2 shows a snapshot of the interface of the on-line version of Atenea.

As can be seen, the input Atenea expects is the answer typed by the student in order to compare it with a set of reference answers (ideal answers) written by the teachers, which have previously been stored in its database. There should be at least three teachers' references per each student answer, and it is advisable that different teachers write them, in order to cover as much paraphrasing as possible. Furthermore, these references can also be taken from the answers of the best students, in order to have more alternatives (Pérez et al., 2004a).

The internal architecture of Atenea is composed of a statistical module, called ERB, and several Natural Language Processing (NLP) modules:

1. ERB relies on the BiLingual Evaluation Understudy (Bleu) algorithm (Papineni et al., 2001). This is the reason why we have called it ERB (Evaluating Responses with Bleu). Bleu was created by (Papineni et al., 2001) as a procedure to rank systems according to how well they translate texts from one language to another. It is based on an n-gram co-occurrence scoring procedure that has been successfully employed to accomplish its aim (Papineni et al., 2001; Doddington, 2002). The core idea of Bleu is that a system-made translation will be better when it is closer to a translation written by a human expert. Therefore, to evaluate a system, it is necessary to have a set of human-made reference translations, and a numerical similarity metric between the system's translations and the manual ones. Besides, this procedure has also been applied to evaluate text summarization systems (Lin and Hovy, 2003). This is because the core idea remains: the more similar a computer-made summary is to a human-made reference, the better it is. In fact, this idea can also be applied for automatically grading students' texts (Pérez et al., 2004a). However, in this case, it is equally important to measure the precision and the recall of the student answer, to make sure both that all that is said is correct, and that it is complete. Therefore, Bleu has been transformed into ERB in order to incorporate the recall by calculating the percentage of each of the reference texts that is covered by the candidate text. Another change is that in an educational environment, it is usually necessary to use a standard scale for the scores, such as between 0 and 10, or between 0 and 100. ERB's scores are always between 0 and 1, hence it is necessary to scale the result (Alfonseca et al., 2004). Tested on a corpus of students' and teachers' answers from real exams and from the Internet, ERB has attained correlation values

with hand-made scores as high as 82%, a state-of-the-art result. Pérez et al. (2004a) and Pérez et al. (2004b) describe the evaluation in more detail.

2. It is clear that the simple use of Bleu algorithm is not enough to build a completely new system for CAA of free text answers, because it lacks the necessary level of robustness in order to face spelling or grammar mistakes, to deal with synonyms or to distinguish the students' word sense. Thus, we have built several Natural Language Processing modules (Alfonseca et al., 2004) that are based on the *wraetic* tools (Alfonseca et al., 2003), available at <http://www.ii.uam.es/~ealfon/eng/download.html>. These modules add the following possibilities to the system: stemming, removal of closed-class words, Word-Sense Disambiguation, synonyms treatment and parsing to translate the text into an intermediary logical form (Alfonseca et al., 2004).

The feedback that the students get from the system is a numerical score and, optionally, an annotated copy of their answer following this colour code:

- If a single *word* (a unigram) is found in any reference text, its background is coloured in *light green*.
- If a *block of two words* (a bigram) is found in any reference text, its background is coloured in *medium green*.
- If a *block of three words* (a trigram) is found in any reference text, its background is coloured in *dark green*.
- We have not considered n-grams larger than trigrams, since it is unusual to find more than three consecutive words between a candidate student text and the human reference text.
- Finally, all words that are not found in any reference text are written over a grey background.

From this output, students can easily discern which portions of their answers are correct and have contributed in incrementing their score, and which their weak points are. Figure 3 shows an example of feedback page. In the user profile, students may also indicate whether they just want the score and are not interested in receiving this feedback.

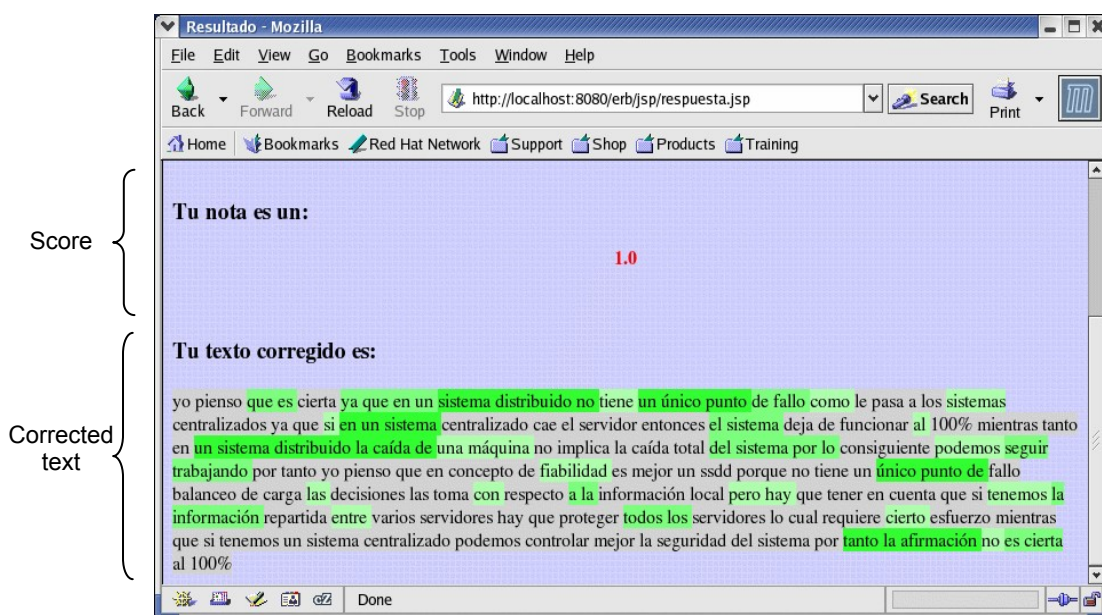


Figure 3. Feedback for student answer

## The integration of Atenea and TANGOW

The integration supports the inclusion of free-text CAA exercises within adaptive courses, as a new type of TANGOW task. The process is as follows: Atenea is launched from TANGOW and, after presenting the corresponding questions to the students and automatically evaluating their answers, it returns the results to TANGOW so that this information can be used to update the user model and to continue with the adaptation process in the rest of course. The author of the course has to provide TANGOW with information about the open-ended questions: for each question, its statement and at least three answer references (ideally, each one would be written by a different teacher) are required. Currently, the adaptation performed by Atenea in our example course uses information such as the user's language (English or Spanish), age (young or adult) and previous knowledge about the subject (novice or advanced).

An initial step in the integration process was to decide which features from the user model currently managed by TANGOW would be used in Atenea in this first experience. We decided to use the *student name* as the login input in order to address the student by his or her name; *age*, because questions should be formulated in a simpler fashion for youngsters than for adults, and different writing styles are expected from them; *experience*, because the assessing process should be different for advanced students than for novice ones; *feedback type*, because when formative assessment is used, the feedback should be more detailed than for summative assessment (where the score is the most relevant result); and *language*, because Atenea is multi-lingual and, therefore, it can deal with students and teachers from different nationalities. In fact, the authors of a course would simply need to write the reference texts in their own language (e.g. Spanish) and the student (e.g. an English speaker) would see the question translated automatically into English, write the answer, and the system would automatically translate it into Spanish and score it against the teacher's references. It has been proven that this does not affect the accuracy of the automatic evaluation (Alfonseca et al., 2004)

Concerning the order in which questions are presented, it is possible to take into account the student experience so that advanced students are not asked questions that are too easy for their level or those they have already solved. Moreover, the higher the level of knowledge the student has, the stricter the system should be when assessing his/her answers. The protocol for communicating TANGOW and Atenea is the following (see Figure 4):

- TANGOW proposes different types of activities to the students, depending on both the adaptation capabilities of the course and the information stored in each user model, and gathers information about their behaviour and performance when accomplishing these activities. When it is considered appropriate for the student to accomplish a task corresponding to Atenea-based assessment, TANGOW sends Atenea the user id, the task id, the type of feedback desired, and all requested user-model attributes.
- Atenea randomly chooses a question that has not already been solved by the student (that is, not yet graded or graded less than half of the maximum score) from the dataset corresponding to the given id. The question is chosen taking into consideration the information stored in the student model. The answer submitted by the student is then evaluated by Atenea, and the resulting score and feedback is presented to the student. This process is repeated until the student has answered the required number of exercises. Finally, once the stop condition is satisfied, Atenea returns a holistic student score for the task to TANGOW.

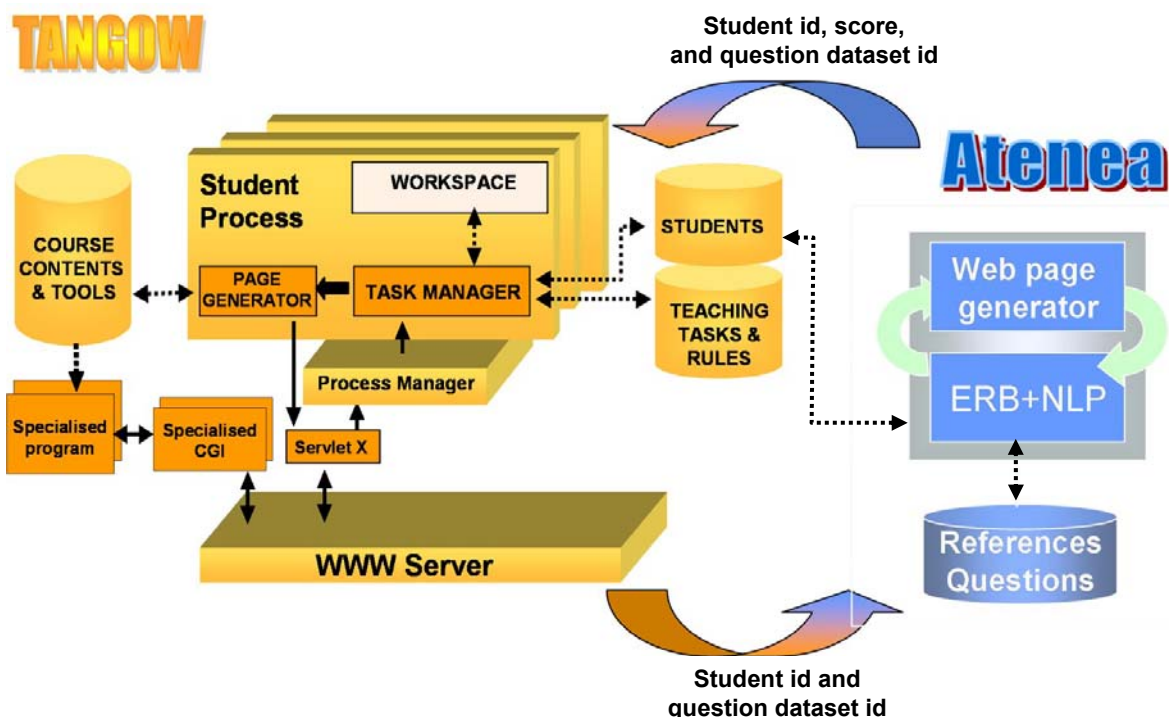


Figure 4. Architecture of the integration between TANGOW and Atenea

We can illustrate this protocol with an example: Peter, a student, logs into TANGOW to study the lesson about concurrency in a course about Operating Systems. Peter does not know anything about concurrency and wants to

learn as much as possible. The first time he logs in TANGOW, he is asked to fill in a form with information about his profile. He logs in as Peter Smith, age: 24, language: English, level (previous knowledge): novice, and desired level of detail: high. Then, TANGOW proposes the activities and presents the corresponding contents adapted to him, according to the course's adaptation rules (Carro et al., 2003) specified for this type of student.

As the course author has included open-ended questions as a practical task to be accomplished at a certain point of the course for students like Peter, TANGOW asks Peter to answer these questions at the corresponding time, and sends the information about Peter to Atenea. Then, Atenea chooses the most adequate question for this topic among the existing ones, in order to ask Peter the question according to his profile. Thus, what Peter sees on the screen of his computer is the Atenea interface presenting him the question: *If you are working with the Unix operating system and you need to run several applications so that they can share information, what are the Unix resources available to accomplish this task?*, and a text area to write the answer in. When Peter pushes the "Go" button, he receives instantaneously the score in the scale indicated by the course author and, as feedback, his answer with the aforementioned colour-code (the darker the background the better the sentence, and grey background for 'useless' information in Peter's answer). One issue to highlight here is that the reference answers Atenea uses for this student are the ones written by the teachers for English novice students, which means that Peter will be less strictly corrected than a more advanced student (since, in this case, the references provided by the teacher for novice students are not so complex as the ones for advanced students).

Atenea continues choosing up to the number of open-ended questions, all of them according to the student's profile (language and expertise level), and iteratively proceeds as before until this number of questions is fulfilled. Finally, Atenea returns the score achieved by Peter for the open-ended questions task to TANGOW, which uses this information to update Peter's user model and to go on with the adaptation procedure in order to decide which the most suitable tasks are to be proposed next.

## An authoring tool for adaptive open-ended questions

A web-based wizard has been developed to facilitate the task of creating adaptive open-ended questions, with the corresponding reference answers, and that of managing the question datasets. It allows:

- Creating new datasets or augmenting an existing one.
- For the selected questions dataset, modifying an existing question or adding a new one.
- For the selected question, modifying the existing question statement, the maximum score or the reference answers.

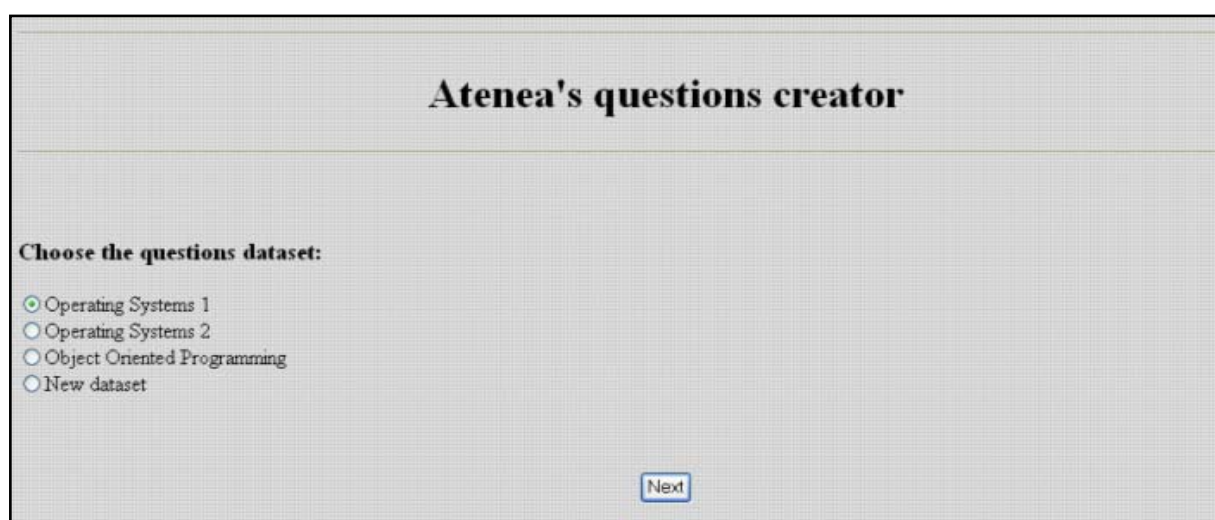


Figure 5. Example of the interface for managing questions dataset

For instance, if an author wants to include a new question for the set of exercises about "Operating Systems I" in the database, with different versions of the same question for English and Spanish students and also for novice and advanced students, the first step is to choose the "Operating Systems I" question dataset as shown in Figure 5. After that, the author will be asked to write the score for this question, and its statement in both English and Spanish languages, using different vocabulary or demanding different details to novice students than to more

advanced ones (see Figure 6). It is also possible to use the “translation” button to automatically translate the statement and the reference answers (currently, only from English to Spanish and vice-versa). In this way, authors only have to write in their preferred language.

The second step for the author is to write the reference answers for each question. (S)he can decide how many references to write (obeying the lower limit of three). Figure 7 shows an example of this step.

The screenshot displays a vertical list of four question entries. Each entry consists of a text box for the question statement, a 'Translate' button, and an 'Answers' button. The questions are as follows:

- Type the statement of the question in English for advanced students:**  
Describe in detail the Unix's interprocess communication possibilities.
- Type the statement of the question in Spanish for advanced students:**  
Explique detalladamente las formas de comunicación interproceso en Unix.
- Type the statement of the question in English for novice students:**  
If we are working with the Unix operating system, how can be accomplished that several programs share information among them?
- Type the statement of the question in Spanish for novice students:**  
Si estamos trabajando con el sistema operativo Unix, ¿cómo podemos conseguir que al ejecutar varios procesos compartan información entre ellos?

Figure 6. Example of adding questions through the authoring tool

The screenshot shows the 'fourth step' of the interface, titled 'Atenea's questions creator - fourth step'. It features two main text input areas:

- Statement of the question:** Contains the English question: 'If we are working with the Unix operating system, how can be accomplished that several programs share information among them?'
- Answer:** Contains the reference answer: 'It could be done by creating a file that could be read by all the programs or with shared memory.'

Below the answer text is a button labeled 'Add another possible answer'.

Figure 7. Example of the interface for adding references

This tool has been tested by six different authors whose familiarity with authoring tools is represented in Table 1.

Table 1. Degree of familiarity of the authors with authoring tools

	Author1	Author2	Author3	Author4	Author5	Author6
<b>Very familiarized</b>		×			×	
<b>Familiarized</b>						×
<b>Medium-level of familiarity</b>			×	×		
<b>Little familiarized</b>						
<b>Not familiarized at all</b>	×					

The procedure followed to perform this test consists in asking each author to complete three tasks with the authoring tool (to insert a new question in one of the question data set, to create a new question dataset and to update the information about a question in one of the question dataset) and next, to interview them according to the survey shown in Figure 8. The results are gathered in Graphs 1 to 6.

1. Rate your familiarity with authoring tools:
  - a) Very familiar
  - b) Familiar
  - c) Medium-familiarity
  - d) Little familiar
  - e) Not familiar at all
  - f) Not known/not answered
  
2. How difficult did you find it to perform these tasks?
  - a) Very easy
  - b) Easy
  - c) Medium-easiness
  - d) Difficult
  - e) Very difficult
  - f) Not known/not answered
  
3. Rate the usefulness of an authoring tool such as this one for adaptive open-ended questions:
  - a) Very useful
  - b) Useful
  - c) Medium-usefulness
  - d) Little useful
  - e) Very little useful
  - f) Not known/not answered
  
4. How intuitive did you find the authoring tool's interface?
  - a) Very intuitive
  - b) Intuitive
  - c) Medium-intuitiveness
  - d) Little intuitive
  - e) It is not intuitive at all
  - f) Not known/not answered
  
5. Which task(s) did you find most difficult?
  - a) None of them
  - b) The first: to insert a new question
  - c) The second: to create a new collection
  - d) The third: to modify an already existing question
  - e) All of them
  
6. In general, you would rather work with the adaptive open-ended questions:
  - a) Using the authoring tool
  - b) Accessing the database through its manager
  
7. How worthwhile do you think that the effort of building different versions of the same statement and answer references for each question is?
  - a) Very worthwhile

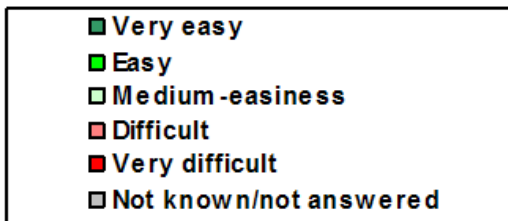
- b) Worthwhile
- c) Medium-worthwhileness
- d) Little worthwhile
- e) Not worthwhile at all
- f) Not known/not answered

8. The thing that I have enjoyed the most of this authoring tool is:

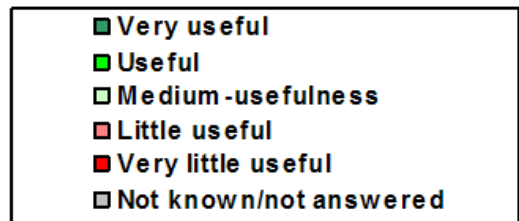
9. The thing that I have liked the least is:

10. I think that this should be improved:

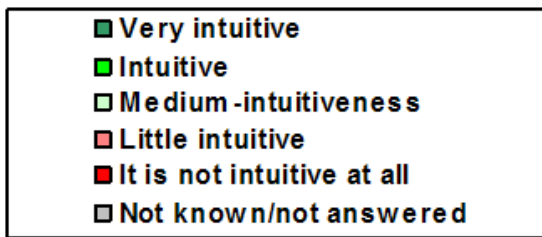
Figure 8. Survey for the authors



Graph 1. Ease of use

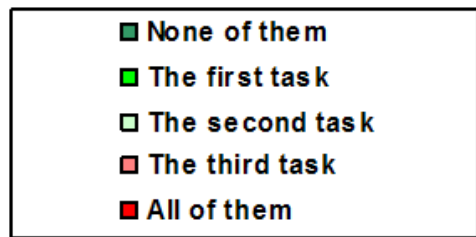
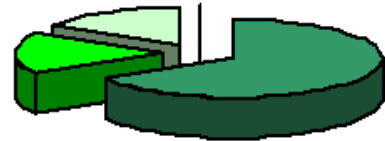


Graph 2. Usefulness

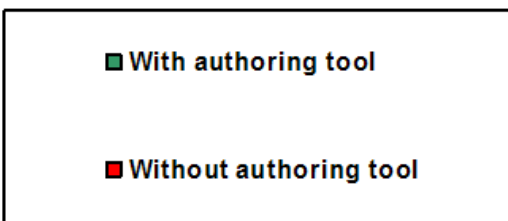
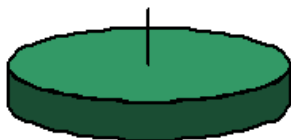


Graph 3. Intuitiveness

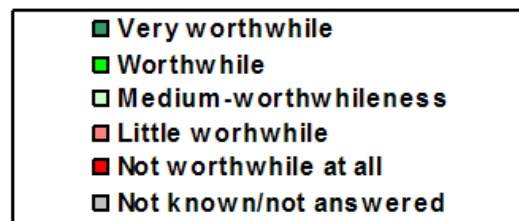
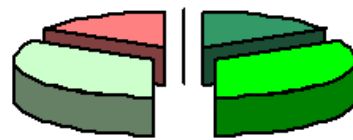
723x258



Graph 4. Task complexity



Graph 5. Preferred development



Graph 6. Worthwhileness

It can be seen that 100% of the interviewed authors, irrespective of their degree of familiarity with authoring tools, have stated that they would rather use the authoring tool than not use it. Besides, they consider it very easy (67%) or easy (33%) to use. Concerning how useful they think it is, more than 80% think that it is very useful. Half of the interviewed authors regard it as intuitive, and the other half says it is intuitive (33%) or more or less intuitive (17%). Most of them (67%) claimed that none of the proposed tasks was difficult to complete. Regarding the items 7-10 of the survey, most of the authors commented that the tool is very simple and that this allows them to gain more control over the task to perform. The greatest inconvenience detected by the authors was the necessity of writing the same question statement and answer references in several ways according to the stereotypes defined. It should be noted here that the experiment was done with the same predefined stereotypes for all the authors. In a real use of the tool, each author is the responsible of establishing the stereotypes to be used for the adaptation. Therefore, writing different versions for distinct stereotypes would be exactly what they were planning to do. One of the authors claimed that he would consider the effort as worthwhile if the time devoted to this task were shorter than the time devoted to the traditional manual assessment of the students' answers. Moreover, he would regard it as exponentially more useful as the number of student answers to assess increased. Another author highlighted as one of the main benefits that he only has to write the information (statements and answer references) once and then it can be used by many different students for several years. Finally, the option of automatically translating from English to Spanish and vice-versa was very well considered since it allowed them to write the texts in their mother tongue.

## Conclusions and future work

Up to date, attempts to develop adaptive CAA systems have been limited to Computer Adaptive Testing (CAT) (Linden and Glass, 2000; Guzmán and Conejo, 2002; Cristea and Tuduce, 2004). That is, to modify the order in which the test items are presented to the students according to their performance during the test. In this paper, a new possibility is presented: to adapt the assessment of free-text answers by taking into account different student's features and to integrate it with different learning activities in adaptive and collaborative web-based courses.

The integration of TANGOW, a system for the dynamic generation of adaptive web-based courses, with Atenea, a program for the automatic assessment of student answers, is based on the following protocol:

- Atenea uses the information stored in TANGOW user models, which includes personal features, preferences, learning styles, knowledge about the subject to be studied and all the student actions and the scores obtained by him/her during the course evolution. A richer profile allows better adaptation of question statements and datasets to individual students.
- The adaptation engine from TANGOW decides when each student should be assessed, depending on his/her profile and achievements, and Atenea chooses the most adequate set of questions for the student, resulting in a fairer evaluation.
- TANGOW benefits not only from the possibility of automatically evaluating free-text answers, but also from the feedback from those questions, which can be used to guide the students during the rest of the course.

An authoring tool has been designed to facilitate the management of adaptive open-ended questions. This tool has been evaluated by several authors that have highlighted its easiness of use and the importance of having tools like that. Moreover, the evaluation of this tool has successfully proven one of the expectations from the integration of TANGOW and Atenea: the possibility of extending the adaptation not only to the contents or the navigation of the course but also to the assessment by choosing the questions to be asked and the reference answers according to the information about each student (stored in the user model). Other conclusions are: the ease of adding the new open-ended questions type of task to TANGOW and, therefore, of including a richer set of activities in TANGOW-based courses, as well as the interest of using the TANGOW formalism for letting the course authors specify different teaching strategies by incorporating CAA activities at different points of the course, depending on the evolution of each student.

The combination of the techniques from two prosperous fields such as Adaptive Hypermedia and Computer Assisted Assessment of open-ended questions could give birth to a new field that could be called Adaptive Computer Assisted Assessment of Open-ended Questions. Computer-assisted learning can be useful for all types of students, and it is particularly well suited to those which, because of any reason (e.g. being physically impaired) cannot attend traditional lectures. Providing teaching materials and activities adapted to the student's specific profile, as well as immediate and detailed feedback for the student's answers, has a special interest in these cases.

As future work, firstly the authoring tool will be used to add more adaptive open-ended questions to the datasets. Secondly, the integration of TANGOW and Atenea will be used as an additional support to traditional lectures about Operating Systems in the studies of Computer Science in our university. And thirdly, the application of different methods and techniques of adaptation to the assessment of open-ended questions will be studied.

## Acknowledgments

This work has been funded by the Spanish Ministry of Science and Technology, project number TIN2004-03140. Many thanks also to the authors that tried our tool, answered the survey for its evaluation, and gave us very useful feedback.

## References

- Alfonseca, E. (2003). *Wraetlic user guide version 1.0*, retrieved July 15, 2005, from <http://www.ii.uam.es/~ealfon/pubs/wraetlic-1.0.tar.bz2>.
- Alfonseca, E., & Pérez, D. (2004). Automatic Assessment of Short Questions with a Bleu-inspired Algorithm and shallow NLP. *Lecture Notes in Computer Science*, 3230, 25-35.
- Brusilovsky, P. (2001). Adaptive Hypermedia. *User Modeling and User-Adapted Interaction*, 11, 87-110.
- Burstein, J., Leacock, C., & Swartz, R. (2001). Automated evaluation of essay and short answers. In Danson, M. (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.
- Carro, R. M., Pulido, E., & Rodríguez, P. (1999). Dynamic generation of adaptive internet-based courses. *Journal of Network and Computer Applications*, 22 (4), 249-257.
- Carro, R. M., Pulido, E., & Rodríguez, P. (2002). Developing and accessing adaptive Internet-based courses. In L. C. Jain, R. J. Howlett, N. S. Ichalkaranje, & G. Tonfoni (Eds.), *Virtual Environments for Teaching and Learning*, World Scientific Publishing Company, 111-149.
- Carro, R. M., Ortigosa, A., & Schlichter, J. (2003). A rule-based formalism for describing collaborative adaptive courses. *Lecture Notes in Artificial Intelligence*, 2774, 252-259.
- Chou, C. (2000). Constructing a computer-assisted testing and evaluation system on the World Wide Web- the CATES experience. *IEEE Transactions on Education*, 3 (43), 266-272.
- Cristea, D., & Tuduce, R. (2004). Test authoring for intelligent e-learning environments. *Paper presented at the 1<sup>st</sup> International Workshop Authoring of adaptive and adaptable educational hypermedia at the Web-Based Education Conference (WBE)*, February 16-18, 2004, Innsbruck, Austria.
- Doddington, G. (2002). Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. *Paper presented at the Human Language Technology Conference*, March 24-27, 2002, San Diego, CA, USA.
- Gutiérrez, S., Pardo, A., & Delgado, C. (2004). An Adaptive Tutoring System Based on Hierarchical Graphs. *Lecture Notes in Computer Science*, 3137, 401-404.
- Guzmán, E., & Conejo, R. (2002). An adaptive assessment tool integrable into Internet-based learning systems. *Paper presented at the International Conference on Information and Communication Technologies in Education*, November 13-16, 2002, Badajoz, Spain.
- Laham, D. (2000). Automated content assessment of text using Latent Semantic Analysis to simulate human cognition. *PhD Dissertation*, University of Colorado, Boulder.
- Lilley, M., & Barker, T. (2003). An evaluation of a computer adaptive test in a UK University Context. In Danson, M. (Ed.), *Proceedings of the Seventh Computer Assisted Assessment Conference*, Loughborough, UK.

- Lin, C. Y., & Hovy, E. H. (2003). Automatic evaluation of summaries using n-gram co-occurrence statistics. *Paper presented at the Human Language Technology Conference*, May 30-31, 2003, Edmonton, Alberta, Canada.
- Linden, W. J. van der, & Glas, C. A. W. (2000). *Computerized Adaptive Testing: Theory and Practice*, Norwell, MA: Kluwer Academic Publishers.
- Lutticke, R. (2004). Problem solving with Adaptive Feedback. *Lecture Notes in Computer Science*, 3137, 417-420.
- Mason, O., & Grove-Stephenson, I. (2002). Automated free text marking with paperless school. In Danson, M. (Ed.), *Proceedings of the 6th International Computer Assisted Assessment Conference*, Loughborough, UK.
- Ming, Y., Mikhailov, A., & Kuan, T. L. (2000). Intelligent essay marking system. In C. Cheers (Ed.), *Learners Together, February*, Singapore: NgccANN Polytechnic.
- Mitrovic, A., & Martin, B. (2004). Evaluating Adaptive Problem Selection. *Lecture Notes in Computer Science*, 3137, 185-194.
- Panos, S., George, P., & Theofanis, D. (2003). Re-Adapting an Adaptive Assessment. Is this possible? *Paper presented at the Third IEEE International Conference on Advanced Learning Technologies*, July 9-11, 2003, Athens, Greece.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W. (2001). BLEU: a method for automatic evaluation of machine translation. *Technical Report RC22176 (W0109-022)*, IBM Research Division, Thomas J. Watson Research Center.
- Pérez, D., Alfonseca, E., & Rodríguez, P. (2004a). Application of the BLEU method for evaluating free-text answers in an e-learning environment. *Paper presented at the Language Resources and Evaluation Conference (LREC-2004)*, May 26-28, 2004, Lisbon, Portugal.
- Pérez, D., Alfonseca, E., & Rodríguez, P. (2004b). Upper bounds and extension of the BLEU algorithm applied to assessing student essays. *Paper presented at the International Association for Educational Assessment Conference (IAEA-2004)*, June 13-18, 2004, Philadelphia, USA.
- Salton, G., Wong, A., & Yang, C. S. (1975). A Vector Space Model for Automatic Indexing. *Communications of the ACM*, 11 (18), 613-620.
- Sosnovsky, S. (2004). Adaptive Navigation for Self-assessment quizzes. *Lecture Notes in Computer Science*, 3137, 365-371.
- Tzanavari, A., Retalis, S., & Pastellis, P. (2004). Adaptive Navigation for Self-assessment quizzes. *Lecture Notes in Computer Science*, 3137, 340-343.
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2, 319-330.
- Whittington, D., & Hunt, H. (1999). Approaches to the computerized assessment of free text responses. In Danson, M. (Ed.), *Proceedings of the Sixth International Computer Assisted Assessment Conference*, Loughborough, UK.