A USABILITY STUDY ON HOW NON-TECHNICAL STUDENTS INTERACT WITH A FREE-TEXT COMPUTER ASSISTED ASSESSMENT SYSTEM

Ismael Pascual-Nieto, Diana Pérez-Marín and Pilar Rodríguez
Computer Science Department, Universidad Autónoma de Madrid
Francisco Tomás y Valiente, 11, 28049, Madrid, Spain

Keywords: Free-text Computer Assisted Assessment, Usability Evaluation, Blended Learning, Web-Based Formative Assessment, Open Learner Modelling, Human-Computer Interaction.

Abstract: Willow is a free-text Computer Assisted Assessment system, which can automatically assess students' short written answers in Spanish or in English. Willow is based on the combination of techniques from Natural Language Processing and User Modelling to generate students' conceptual models (i.e. a set of interconnected concepts of a certain area-of-knowledge associated with an estimated value that indicates how well each concept has been assimilated by the student) from the students' free-text answers. In the past, the system was used by a group of students of an Operating Systems course within an Informatics degree. The results of that study suggested that the system was useful for these students. Nevertheless, our hypothesis was that the procedure implemented in Willow is also suitable for non-technical domains and, that students without computer training are able to use Willow without any technical difficulty. Therefore, we asked a group of voluntary students of a Pragmatics course within an English Studies program to use the system. The results achieved support our hypothesis that Willow can successfully be applied to a non-technical domain, and it can be used by non-technical students.

1 INTRODUCTION

Computer Assisted Assessment (CAA) is the field that studies how computers can effectively be used for evaluating students' work. In early work, CAA tools were only used to score Multiple Choice Questions (MCQs) or fill-in-the-blank exercises. This can be explained because these types of items are easier to automatically evaluate with computers. However, according to the general opinion of the field, other types of assessment are necessary to cover higher cognitive skills (Sigel, 1999).

On the other hand, the automated assessment of students' free-text answers has been regarded by many as the Holy Grail of CAA. Regardless, several factors have supported the increasing interest in this field including i) advances in Natural Language Processing (NLP), ii) teachers not having sufficient time to give students appropriate feedback (despite the general assumption of its importance), and iii) the conviction that assessment should not be based only on MCQs.

Currently, there are many different free-text CAA programs, used both in academic and commercial environments, and which are able to process many European and Oriental languages. Moreover, they have been applied both to technical and non-technical domains. For instance, the Automark system (Mitchell et al., 2002) uses Information Extraction techniques to automatically score Science essays in English; the Japanese Essay Scoring System (Jess) (Ishioka and Kameda, 2004) automatically assesses Japanese students' general topic essays using LSA. Table 1 gathers a representative list of free-text CAA systems together with the technique and domain applied.

For the evaluation, the metric reported by the author is the one used: Corr, correlation; Agr, Agreement; EAgr, Exact Agreement; CAcc, Classification accuracy; f-S, f-Score; and, - for not available. When the authors have presented several values for the evaluation, the average value has been taken.

Pascual-Nieto I., Pérez-Marín D. and Rodríguez P. (2009).
A USABILITY STUDY ON HOW NON-TECHNICAL STUDENTS INTERACT WITH A FREE-TEXT COMPUTER ASSISTED ASSESSMENT SYSTEM.
In *Proceedings of the First International Conference on Computer Supported Education*, pages 127-132

Table 1: Domains to which the current existing CAA of free text answers systems have been applied, the technique that they use and their evaluation (Pérez-Marín, 2007).

SYSTEM	DOMAIN	TECHNIQUE	EVAL.
AEA	Marketing, engineering	LSA	Corr:.75
Apex Assessor	Sociology of education	LSA	Corr:.59
ATM	Factual disciplines	Pattern matching	---
Automark	Science	Information Extraction	Corr:.95
Auto-marking	Biology	Pattern matching	EAgr:.85
BETSY	Text classification tasks	Bayesian networks	CAcc:.77
CarmelTC	Physic	Machine learning	f-S:.85
C-rater	Comprehension, algebra	NLP	Agr:.83
EGAL	Opinion and factual texts	NLP	---
E-rater	GMAT exam	NLP	Agr:.97
IEA	Psychology and military	LSA	Agr:.85
IEMS	Non-mathematical texts	Pattern matching	Corr:.80
IntelliMetric	K-12 and creative writing	NLP	Agr:.98
Jess	General topic essays	Pattern matching	Corr:.71
Larkey's system	Social and opinion	TCT	EAgr:.55
MarkIT	General topic essays	NLP	Corr:.75
MRW	Semantic networks	Logical inference	---
PEG	Non-factual disciplines	Linguistic features	Corr:.87
PS-ME	NCA or GCSE exam	NLP	---
RMT	Research on Psychology	LSA	---
SEAR	History	Pattern matching	Corr:.45

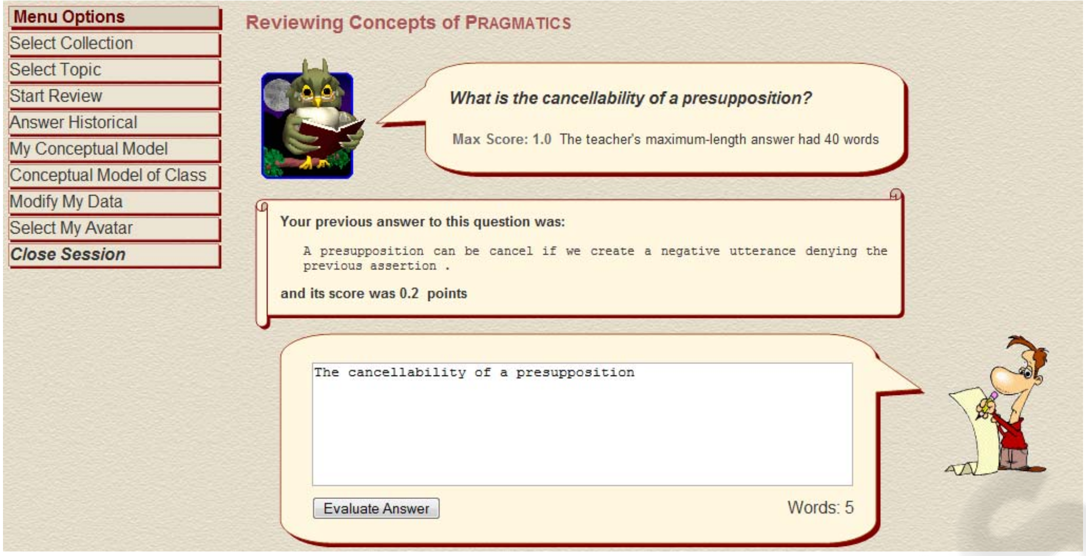
As can be seen, free-text CAA systems have been applied to many different domains, and there is no a clear trend of using certain techniques for certain domains. For instance, according to their authors, the best correlation between the automatic and the teachers' scores (95%) is achieved by Automark, which uses Information Extraction in a technical domain. On the other hand, the highest Agreement value (98% measured as the percentage of times that the automatic and the teacher scores only differed by a certain small margin) is achieved by Intellimetric, which uses full Natural Language Processing techniques in a non-technical domain.

In previous work, we implemented Willow, a free-text CAA system. Willow is based on the synergic combination of NLP and User Modelling techniques to automatically assess students' short answers written both in Spanish and English.

The core idea of the system is that the student's answer should be similar to the teachers' correct answers (reference answers).

During the 2005-2006 and 2006-2007 academic years, students of the Informatics degree at our university were given the possibility of using Willow to review their Operating Systems course (Pérez-Marín, 2007). We used this course for initial trials of the system for two reasons. Firstly, our algorithm for grading free-text answers depends on comparing student answers to the reference answers of teachers, and thus the more restricted the correct answers are, the better the system works. In technical domains, correct answers are reasonably restricted. Secondly, students of Informatics can be expected to have more ability to handle innovative software.

However, in the 2007-2008 course, we wanted to test that Willow can also successfully be used in non-technical domains (i.e. non-Informatics domains) with students without Informatics training. Therefore, we asked teachers of other faculties to collaborate with us. The English Studies Faculty took notice of our petition. In particular, the teachers



Reviewing Concepts of PRAGMATICS

What is the cancellability of a presupposition?

Max Score: 1.0 The teacher's maximum-length answer had 40 words

Your previous answer to this question was:

```
A presupposition can be cancel if we create a negative utterance denying the
previous assertion .
```

and its score was 0.2 points

```
The cancellability of a presupposition
```

Figure 1: Sample snapshot of the interface of Willow.

of the Pragmatics subject were willing to use Willow as a complement of their lessons. 22 Pragmatics students volunteered to use Willow during one semester.

The results achieved support our hypothesis that Willow not only can be successfully applied to a non-technical domain, but also that Willow can be used by students without Informatics training with very little introduction to the system.

The paper is organized as follows: Section 2 provides a description of Willow; Section 3 describes the settings and results of the usability study conducted; and, finally, Section 4 ends with the main conclusions.

2 WILLOW

Willow is a web-based application able to assess students' short answers written both in Spanish and English in an automatic way. The goal is not to replace the teacher but to help him or her by providing students with an alternative mean of reviewing the course material. Willow scores the student's answer in terms of its similarity to a set of correct answers provided by the teacher. The more similar the answer is to the teachers' answers, the higher the score assigned is.

Figure 1 shows a snapshot of the Willow's interface. As can be seen, the interface tries to emulate a dialogue between two animated agents: an owl representing the system, and a character chosen by the student from a gallery representing the student.

The rationale to choose an owl as the animated agent to represent Willow is because in most Western cultures an owl is usually regarded as a bird of wisdom. The reason for this can be found in the Greek mythology. In fact, Athena, the Greek goddess of wisdom is often depicted holding an owl. On the other hand, the rationale to let the student choose his or her own animated agent is to increase the possibilities that the student feels that the character chosen represents him or her.

Willow is intended for formative assessment rather than summative assessment. That is, the goal of Willow is to provide more training to the students before their final exam. Thus, the system does not only provide a numerical score as feedback, but also the student's processed answer and the correct answers as provided by the teachers.

Therefore, Willow is similar to the other free-text CAA systems reviewed in the previous section as its goal and core idea is the same. The goal is to automatically assess free-text students’ answers to provide the students with immediate feedback. The core idea is that the more similar the student’s answer is to the correct answers provided by the teachers, the higher the score the student should achieve.

Willow has usually been applied to the Informatics domain because we are teachers of this subject. Thus, it is easier for us to do the first experiments of the system with technical students that are both more used to new software and, more approachable as they attend lessons in our faculty.

The NLP techniques used by Willow are different for Spanish and English languages. In fact, the experiments performed highlight that there is a different optimum combination of NLP techniques for each language used in Willow (Pérez-Marín, 2007).

In particular, for Spanish the optimum combination found is NLP+LSA+Genetic Algorithms reaching up to 63% Pearson correlation between the automatic and the teachers' scores (Pérez-Marín, 2007). For English, the Genetic Algorithms could not be applied as Willow has not been used by English students and thus, we do not have the information needed to run the algorithms. Hence, the optimum combination found is NLP+LSA reaching up to 56% Pearson correlation between the automatic and the teachers' scores (Pérez-Marín, 2007).

Willow has a unique feature: it is able to automatically generate students' conceptual models from the students' free-text answers. A student conceptual model can be defined as a set of concepts and their relationships for a certain area-of-knowledge. Each concept is associated an Estimate of Learner Level of Competence (ESLOC) value by the system. The ESLOC value of a concept can be between 0 and 1.

The procedure to automatically generate the student model will not be described here, as it is out the scope of this paper and, it has already been published (Pérez-Marín, 2007).

3 USABILITY STUDY

3.1 Setting Up the Experiment

After using Willow in a technical domain during two years, we wanted to test our hypothesis that the system could also be applied to non-technical domains and used by non-technical students. Therefore, we asked the rest of faculties of our home university to collaborate with us. The English Studies faculty took notice of our petition and, the Pragmatics teachers told us that they were willing to use Willow with their students.

Hence, we asked the Pragmatics teachers to provide us with the questions they usually ask their students in order to check whether it was really possible to apply Willow's core idea to Pragmatics. That is, to measure if it is possible to write a set of correct answers to the Pragmatics questions and, to automatically compare these correct answers to the students' answers.

We observed from the sample questions and correct answers provided by the Pragmatics teachers, that there is indeed more openness in what can be answered in Pragmatics than in Operating systems. On the other hand, we also realized that there were a defined set of concepts that should be reviewed. Furthermore, we decided that given that providing definitions for concepts was more difficult in Pragmatics, it was not necessary that all questions were in the form of requesting a definition.

Therefore, we asked the Pragmatics teachers to introduce non-ambiguous questions using Willow's authoring tool (Willed), or a text editor.

The two teachers of the subject agreed that they would rather use the text editor. Although they knew the application was easy to use, they already have information in text documents and they considered it would be easier for them to prepare the documentation in plain text.

After one month of non-full time work, the Pragmatics teachers came up with 49 questions, with three different correct answers per question, and covering the first four topics of the subject.

3.2 The Experiment

Once the domain has been established and the information introduced in Willow, we asked the Pragmatics teachers to allow us to go one day at their class to present Willow to the students who voluntarily wanted to use the system.

That way, we could immediately solve any problem or doubt the students may have, and at the same time, we could start observing how the students interact with the system.

Moreover, given that it was the first time that we had non-technical students using the system, we wanted to know the students' opinion before using Willow (to find out if they were somehow prejudiced against automatic free-text scoring) and thus, we asked them to fill in a questionnaire before starting the experiment.

The questionnaire consisted of three closed-answer items and two open-answer items. The closed-answer items asked the students about their degree of familiarity with computers, on-line applications and concept maps. The open-answer items asked the students whether they thought they would prefer to view just their conceptual model, or that for the class as a whole, and which representation format they would prefer: concept map, conceptual diagram, table, bar chart or textual summary.

22 students out of the 45 students enrolled in the Pragmatics course (i.e. 49%) volunteered to take part in the experiment. From them, 19 students filled in the questionnaire on a voluntary and anonymous basis. The analysis of these questionnaires revealed us that they are not prejudiced at all with on-line scoring systems.

On the other hand, these students, albeit they did not have any computer training, were quite familiar with computers: 47% of the students claimed that they were familiar with on-line applications and none of the students stated that s/he did not know how to use an on-line application.

The questionnaire also suggested that the students' knowledge of concept maps was low with 58% of the students answering that they were little familiarized with concept maps. Some students even asked what concept maps are. Nevertheless, when we explained what concept maps are, it turned out that 37% of the students prefer this form of representation to view conceptual models over the other formats available in Willow. Most of the students also stated that they would prefer just to look at their the individual conceptual model (74%), giving reasons such as that they are more interested to find out which concepts they do not yet fully understand, than in looking at the general picture for the class as a whole.

According to the results of the questionnaire, the first day the students use Willow in class with us, we could observe that none of the non-technical students had any technical difficulty in using any of the system's features. On the contrary, all of the students were able to answer several systems' questions with very little explanation (just a 5-minute Powerpoint presentation of the interface).

In fact, 123 students' answers were recorded, and as can be seen in Figure 2: 95% of the students modified which lessons they wanted to be asked, 77% students modified the animated agent used to represent them, 77% looked at the history of questions, 27% changed their personal data, 79% looked at the model and even 18% of the students try to cheat the system by copying the correct answers of the teachers as if they were their answers.

None of the students complaint about the interface of Willow and, they thought that the owl was a quite friendly animated agent.

Figure 3 shows a graph displaying the average number of questions answered by each student since November 16th 2007 (the first day the experiment started in class) till December 15th 2007 (i.e. the first month of experiment). And, again since February 5th 2008 till February 7th 2008, the next time the students started using Willow after Christmas holidays (from the end of December till the beginning of January) and the other exams in January, to review before the final exam on February 8th 2008.

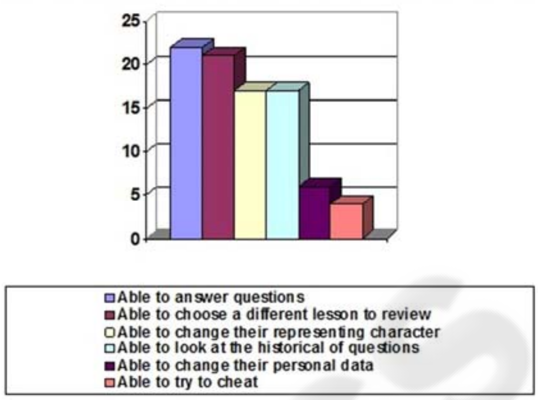


Figure 2: Number of non-technical students who have used some Willow's features.

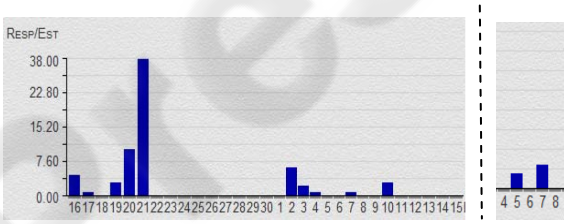


Figure 3: Average number of questions answered by the non-technical students.

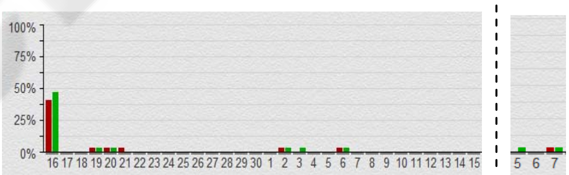


Figure 4: Percentage of students who have looked at their generated conceptual models.

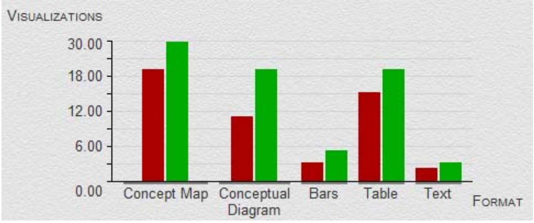


Figure 5: Number of times the students have looked at different representation formats.

As can be seen, the students have not regularly used Willow, although they have valued this possibility by using Willow again in the days previous to the exam. Some comments that the students have emailed us about the use of Willow have been to thank us and their teachers for giving them this opportunity. Additionally, and despite students have not complaint about the interface and have regarded it as friendly, some of them (specially the ones who have used the system longer, even in 2-hour sessions, while the average assessment session was half an hour) have expressed their wish of being presented a higher variety of exercises. That is, not only open-ended questions, but also interactive games of choosing a solution or directly relating concepts.

It is also interesting to highlight that one of the student who has used more Willow has been the woman in her fifties. Contrarily to what could be thought giving the digital gap between young people and adult people, she has completed 93% of the questions of the course whereas the average percentage of completion of the course has been 17% (22% standard deviation).

Regarding the use of the generated conceptual models, 32% of the 19 students have looked again at them. Even, sometimes the students have entered the system just to look at their concept map representation and the class concept map representation without answering any questions.

Figure 4 shows the percentage of students who have entered Willow to look at their particular conceptual model (painted in green, light colour) or, to look at the class conceptual model (painted in red, dark colour). As can be seen, not only the students have valued the possibility of getting more training before the exam with Willow, but also of looking at the generated conceptual model. In fact, the logs revealed how, in the days previous to the final exam, some students have also looked again at their individual and class conceptual models.

Regarding whether they prefer the individual or the class conceptual model, according to their answers in the questionnaire, most students thought they would prefer the individual conceptual model (74%), as stated before, and the logs confirmed this preference. Finally, concerning which form of the representation formats available they prefer (concept map, conceptual diagram, bar chart, table and textual summary), it can be seen in Figure 5 how although the students have looked at all of them, the one they have inspected more is the concept map.

4 CONCLUSIONS

In this paper, the hypothesis that Willow can also be applied to non-technical domains and be used by students without Informatics training has been proved. Willow has, in the past, been used to review the Operating systems subject of an Informatics degree. However, in the 2007-2008 academic year we thought that it could also be applied to non-technical domains. Our belief was based on the fact that free-text scorers have been used both for technical and non-technical domains and, that the core idea of Willow (i.e. that the student's answer should be similar to the teachers' answers) is applicable to non-technical domains too, provided that a fairly limited and non-ambiguous set of correct answers can be written for each question.

Therefore, we carried out an experiment in the English Studies faculty, in which 22 students without English training have been able to interact with the system without problem.

In the future, we would like to do a more systematic experiment to collect more data about the differences in using free-text scoring systems such as Willow by non-technical or technical students.

ACKNOWLEDGEMENTS

This work has been sponsored by Spanish Ministry of Science and Technology, project TIN2007-64718.

REFERENCES

Ishioka, T. and Kameda, M. (2004), 'Automated Japanese Essay Scoring System: JESS', *Proceedings of the 15th International Workshop on Database and Expert Systems Applications*, 4-8.

Mitchell, T.; Russell, T.; Broomhead, P. and Aldridge, N. (2002), Towards Robust Computerised Marking of Free-Text Responses, *in* 'Proceedings of the 6th Computer Assisted Assessment Conference'.

Pérez-Marín, D. 2007. *Adaptive Computer Assisted Assessment of free-text students' answers: an approach to automatically generate students' conceptual models*, PhD thesis, Computer Science Department, Universidad Autónoma de Madrid, Spain.

Sigel, I., ed. (1999), *Development of mental representations: Theories and Applications*, Lawrence Erlbaum Associates, New Jersey, U.S.A.