Formative Assessment Visual Feedback in Computer Graded Essays

Robert Williams and Heinz Dreher
Curtin University of Technology, Perth, Western Australia

bob.williams@cbs.curtin.edu.au h.dreher@curtin.edu.au

Abstract

In this paper we discuss a simple but comprehensive form of feedback to essay authors, based on a thesaurus and computer graphics, which enables the essay authors to see where essay content is inadequate in terms of the discussion of the essay topic. Concepts which are inadequately covered are displayed for the information of the author so that the essay can be improved. The feedback is automatically produced by the MarkIT Automated Essay Grading system, being developed by Curtin University researchers.

Keywords: AEG, Automated Essay Grading, visualisation, automated assignment assessment, formative assessment, graphical representation.

Background

The motivation for developing computer supported techniques to assess or grade free text assignments or essays is rather obvious - increased speed, efficiency and consistency, and thus reduced costs and an amelioration of the onerous nature of (humans) marking large volumes of essays in a short time. Of course, this assumes effectiveness, reliability and user (student and teacher) acceptance of 'computer as assessor'. These three aspects have been reported on in the work of Williams & Dreher (2004) for example.

Automated Essay Grading (AEG) is an emerging phenomenon widely documented in the literature (Shermis & Burstein, 2003; Valenti, Neri & Cucchiarelli, 2003; Williams, 2001; Williams & Dreher, 2004). Many of the current AEG systems claim to produce various kinds of feedback regarding the knowledge deficit or other problems in the essays enabling the essay authors to learn, improve, and correct the errors for future submissions. However, much of the feedback is generic in form, for example "this section is inadequate" or "this section needs improvement". This sort of feedback is not very helpful to the learner, and if the truth be known, it is often provided as a justification for the mark, so that when a student queries the grade given, the assessor can offer some further 'soothing' words at least not inconsistent with the original feedback. Of course, the type of evaluation we are concerned with here is formative, and we appreciate that the case of summative evaluation needs to be treated separately – our interest is in the former.

Purpose of Assessment

In our work on grading and assessment we take the view that incremental improvement is an important goal for the learner and the teacher. This implies that when students are given assignments it is the teacher's role to evaluate the work against the stated assignment assessment criteria and provide the student with a grade and some reasons which explain why the particular grade was awarded. An example of such a scheme can be seen in Figure 1 for a course dealing with JavaScript programming and website development.

criterion	**mark**
1) Features	/10
Minimum of 10 features to be listed	
2) Functionality	/10
Implemented features must be purposeful and function correctly	
3) Navigation	/10
Website must be navigable with navigation support	
4) Usability	/10
Website must have good usability	
5) JavaScript code & explanation	/50
5 functions implemented from the suggested list – mark out of 10 for each of 5 functions (5 for code + 5 for explanation)	
6) Innovative aspects	/10
Anything new, different, & exciting; ***Zero*** *is the default mark; Nominate your candidate feature*	
Total score	**/100**

Figure 1: example of assignment assessment critera for an interactive website

Note: a third column headed "assessor's comments" is used to provide constructive feedback
Source: from the authors' coursework teaching

Naturally, the criteria given in Figure 1 must be distributed with the assignment specification; else the students' would have no goal. The assessment task for such assignments involves considering the assignment from the viewpoint of each of the six criteria and making some judgment and generating relevant comments.

Assignment tasks which can conveniently be subdivided into chunks, an extreme example being Multiple-Choice or True-False Tests, lend themselves to computer scoring. However the more essay-like the assignment task the greater the challenge for automated or semi-automated assessment. Nevertheless, there is a growing body of literature in the field of AEG – see below.

In an interesting case of formative evaluation in a course with well in excess of one hundred students, and the flexibility for the students to choose from a variety of topics or themes (Dreher, Scerbakov & Helic, 2004), the authors claim good support provided by the Learning Management System (WBT-Master), which permits individual and relevant formative evaluation comments to be efficiently generated. Figure 2 is a screenshot of an essay assignment being assessed and commented upon.

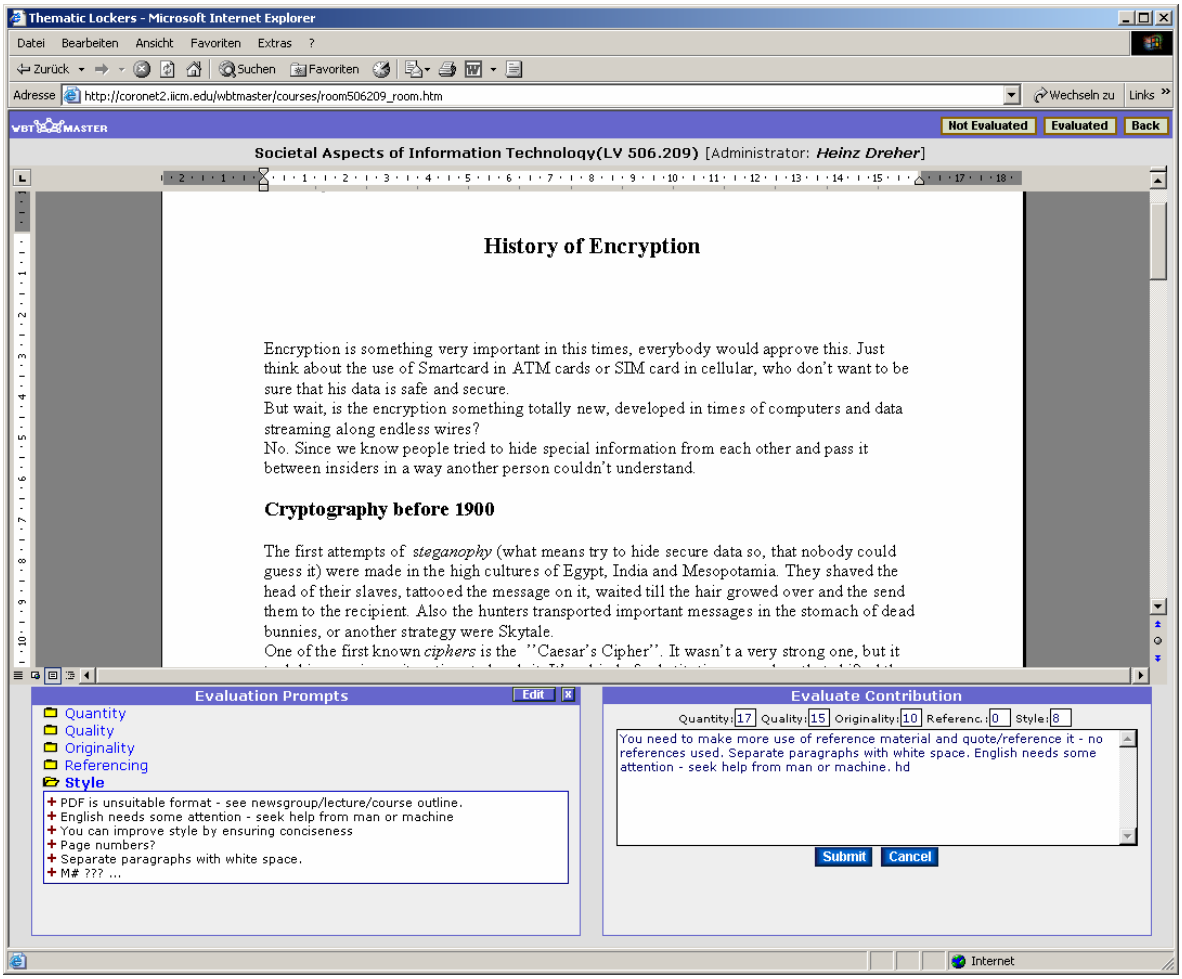


Figure 2 – semi-automated assessment feedback provision for essay assignments

Source: Dreher, Scerbakov & Helic (2004) – reproduced with permission

It should be clear that evaluating assignments and providing feedback to students for the purposes of improvement is on the one hand good education practice, and on the other is very 'expensive'. As we have been developing our AEG system (MarkIT) we have had a unique opportunity to ponder on the provision of meaningful, relevant, consistent feedback which students can use to reflect on their own performance in essay writing.

We now present a short section on the state of the art of AEG, making particular note of the nature and extent of feedback which is provided by these systems, and then take the opportunity to explain how our AEG system has been engineered in terms of feedback provision.

Automated Essay Grading Systems and Feedback Provision

AEG systems are now emerging from the research laboratories into primary, secondary and tertiary education systems around the world (Shermis & Burstein, 2003; Valenti, Neri & Cucchiarelli, 2003; Williams & Dreher, 2004). In the four systems mentioned below, which can be considered as representative of the various approaches to AEG, we consider the level and the form of feedback provided to students. We note that the emphasis is on the grade and not on feedback which may be used to guide improvement and thus further learning. Formative evaluation, including that of content, is considered to be an important aspect of assessment and hence we have worked at including such functionality in MarkIT.

One of the earliest systems for computer grading of essays in the literature was reported in an article by Page in which he described Project Essay Grade (PEG) (Page, 1966). With the rapid advancement in computing power and text processing technologies since the 1960's, more powerful essay grading systems have emerged, and we now discuss the most serious contenders in the field.

PEG

PEG has its origins in work begun in the 1960's by Page and his colleagues (Page, 1966). The idea behind PEG is to help reduce the enormous essay grading load in large educational testing programs, such as the Scholastic Aptitude Test (SAT) (College Board, 2002). When multiple graders are used, problems arise with consistency of grading. A larger number of judges are likely to produce a true rating for an essay. A sample of the essays to be graded is selected and marked by a number of human judges. Various linguistic features of these essays are then measured. A multiple regression equation is then developed from these measures. This equation is then used, along with the appropriate measures from each student essay to be graded, to predict the average score that a human judge would assign. It appears that the main form of this feedback is an essay score, which indicates the level achieved by the student who wrote the essay:

> "The feedback provided suggests whether or not students are on a trajectory to take college-level coursework and what remedial options the district offers for those who are not on that trajectory" (Shermis, Mzumara, Olson, & Harrington, 2001, p 248).

E-rater

E-rater uses a combination of statistical and Natural Language Processing (NLP) techniques to extract linguistic features of the essays to be graded. As in all the conceptual models discussed in this paper, E-rater student essays are evaluated against a benchmark set of human graded essays. E-rater has modules that extract essay vocabulary content, discourse structure information and syntactic information. Multiple linear regression techniques are then used to predict a score for the essay, based upon the features extracted. For each new essay question, the system is run to extract characteristic features from human scored essay responses. Fifty seven features of the benchmark essays, based upon six score points in an Educational Testing Services (ETS) scoring guide for manual grading, are initially used to build the regression model. Using stepwise regression techniques, the significant predictor variables are determined. The values derived for these variables from the student essays are then substituted into the particular regression equation to obtain the predicted score. One of the scoring guide criteria is essay syntactic variety. After parsing the essay with an NLP tool, the parse trees are analysed to determine clause or verb types that the essay writer used. Ratios are then calculated for each syntactic type on a per essay and per sentence basis. Another scoring guide criterion relates to having well-developed arguments in the essay. Discourse analysis techniques are used to examine the essay for discourse units by looking

for surface cue words and non-lexical cues. These cues are then used to break the essay up into partitions based upon individual content arguments. The system also compares the topical content of an essay with those of the reference texts by looking at word usage. Given that a detailed analysis of the essay is done it is possible to provide some detailed feedback. A commercial implementation of E-rater is known as Criterion. Criterion feedback gives details of errors in grammar, usage, and mechanics. Other comments about the essay style are also provided. Criterion also provides feedback relating to the essay background, thesis, main ideas, supporting ideas and conclusion (Attali & Burstein, 2004).

IEA

The Intelligent Essay Assessor (IEA) is a Latent Semantic Analysis (LSA) based system. LSA represents documents and their word contents in a large two dimensional matrix semantic space. Using a matrix algebra technique known as Singular Value Decomposition (SVD), new relationships between words and documents are uncovered, and existing relationships are modified to more accurately represent their true significance. The words and their contexts are represented by a matrix. Each word being considered for the analysis is represented as a row of a matrix, and the columns of the matrix represent the sentences, paragraphs, or other subdivisions of the contexts in which the words occur. The cells contain the frequencies of the words in each context. The SVD is then applied to the matrix. SVD breaks the original matrix into three component matrices that, when matrix multiplied, reproduce the original matrix. Using a reduced dimension of these three matrices in which the word-context associations can be represented, new relationships between words and contexts are induced when reconstructing a close approximation to the original matrix from the reduced dimension component SVD matrices. These new relationships are made manifest, whereas prior to the SVD, they were hidden or latent. Landauer, Foltz & Laham (1998) developed the Intelligent Essay Assessor, using the LSA model. To grade an essay, a matrix for the essay document is built, and then transformed by the SVD technique to approximately reproduce the matrix using the reduced dimensional matrices built for the essay topic domain semantic space. The semantic space typically consists of human graded essays. Vectors are then computed from a student's essay data. The vectors for the essay document, and all the documents in the semantic space are compared, and the mark for the graded essay with the lowest cosine value in relation to the essay to be graded is assigned. Such techniques would presumably permit detailed feedback provision - the system gives an estimated grade for the essay, and also details of subtopics that the student did not cover in the essay. (Foltz, Laham, & Landauer, 1999).

TCT

Larkey (1998) implemented an AEG approach based on text categorization techniques (TCT), text complexity features, and linear regression methods. The Information Retrieval literature discusses techniques for classifying documents as to their appropriateness of content for given document retrieval queries (van Rijsbergen, 1979). Larkey's approach

> ".. is to train binary classifiers to distinguish "good" from "bad" essays, and use the scores output by the classifiers to rank essays and assign grades to them." (Larkey, 1998, p90)

The technique firstly makes use of Bayesian independent classifiers (Maron, 1961) to assign probabilities to documents estimating the likelihood that they belong to a specified category of documents. The technique relies on an analysis of the occurrence of certain words in the documents. Secondly, a k-nearest neighbour technique is used to find the k essays closest to the student essay, where k is determined through training the system on a sample of human graded essays. The Inquery retrieval system (Callan, Croft, & Broglio, 1995) was used for this. Finally, eleven text complexity features are used, such as the number of characters in the document, the

number of different words in the document, the fourth root of the number of words in the document, and the average sentence length. Larkey conducted a number of regression trials, using different combinations of components. She also used a number of essay sets, including essays on Social Studies, where content was the primary interest, and essays on general opinion, where style was the main criterion for assessment. This system appears to only provide a discrete grade for each essay processed (Larkey, 1998).

The MarkIT Automated Essay Grading System

MarkIT is an AEG system that uses propriety technology based on NLP techniques, which has at its core an electronic thesaurus (Williams & Dreher, 2004). As with some other AEG systems, 50-200 human graded essays are used to build a scoring algorithm using multiple linear regression. Better performance is obtained if multiple humans grade the same essays and the scores averaged. An instructor prepares an electronic model answer on the essay topic. Typically this is done with reference to the assignment objectives and assessment criteria. In practice the model answer is often represented as 'the best' of the human graded essays, as instructors may not have developed as clear a formulation of 'good' answers as would be desirable. Students electronically submit their essays on the topic, via the web. The model answer is processed by the system to build up a propriety representation of the meaning of the content of the essay. Student answers are processed in the same manner. The student answers are then processed to ascertain how much of the model answer's content is contained in them. Grades are assigned accordingly.

The MarkIT system relies on building a propriety representation of the knowledge contained in the model answer. A student essay is processed using a combination of NLP techniques to build the corresponding propriety knowledge representation. Pattern matching techniques are then employed to ascertain the proportion of the model answer knowledge that is present in the student answer, and a grade assigned accordingly. An electronic version of a thesaurus is used to extract lexical information for the building of the document knowledge representation.

The technique allows a formal representation of free unseen text to be quickly and robustly built for further analysis by the MarkIT system. The approach used has a need for a semantic representation that does not need substantial hand coding of knowledge structures prior to use, and that can deal with unlimited unseen text. Many NLP systems use some kind of a parser to initially extract the syntax of sentences in a document as an initial step prior to further processing. Semantic analysis then follows. MarkIT uses a specially designed chunking algorithm to perform preliminary processing to extract noun phrases and verb clauses contained in essay sentences.

First experiences show good performance. Experiments have been conducted with a number of 1st year Information Systems student essays, and 2nd year Law student essays, both at university level, and also year 8 secondary school English essays. These essays were prepared by students using a word processor, and comprised some 300 to 500 words, or about one page of text. Expert human graders created the "Human" scores in the usual way by applying the model answer criteria to the essays presented to grading. The computer scoring was a rather simple process of compiling all student answers into text files and submitting them to the computer algorithm. Our technology takes less than 5 seconds per essay to deal with the types of inputs described above. Feeding the model answer which is derived from the course content to the computer is a slightly more involved task.

MarkIT Results for 20 Law Essays

The graph in Figure 3 - Human vs Computer-based scores in ascending order of Human scores represents results for a sample of 20 law essays (horizontal axis) in which the maximum possible assessment was 30 (vertical axis) and shows the comparison between expert human and computer

assessments. The data is (arbitrarily) ordered by increasing computer score. Assignment 1 is assessed by the human at 2 and by the computer at 10 (leftmost data item). Assignment 10 at assessed at 21 by both human and computer, whereas assignment number 20 (rightmost data point) is assessed by the human at 27, and by the computer at 32 – yes, we omitted to inform the computer about the maximum mark on this run! As can be seen the computer tracks the human reasonably well, but further scoring algorithm refinement is indicated. The correlation between the human and computer scores is 0.72.

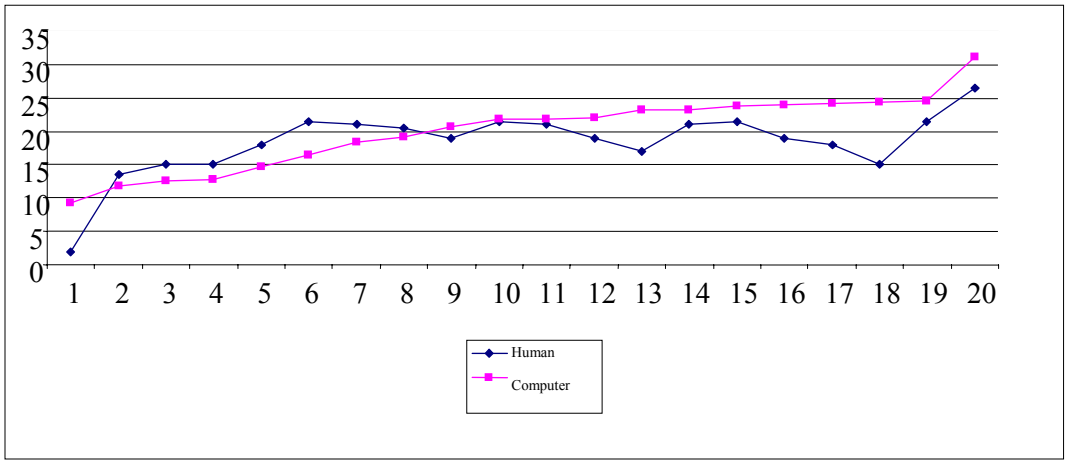


Figure 3 - Human vs Computer-based scores in ascending order of Human scores

Graphical Feedback

In Figure 4 – Concept frequencies: student answer and course content, we have presented another example of MarkIT output. In this case we have a graph showing the 'concepts' associated with both the model answer and the student answer. Naturally, the better the correspondence between the concept representation in both, the better the score. If we focus on the tallest bar (Concept_Number 31) we see that the student answer (dark bar) contains a concept_frequency of 6 (vertical axis) where the model answer called for no discussion on this topic or concept. We say the student has introduced irrelevancies into the answer; or perhaps this is what can be termed an error on the student's part. Concept_Number 26 has a better match between model and student answer, indicating the student has learned relevant material. There are three cases where the model answer concepts are not matched by a student contribution (3, 28, 30) – this we would call "ignorance" or a deficit in knowledge. Such visual feedback is rather informative to student and teacher alike. It is intended to further develop such visual feedback into a dynamic object which responds to inquiry for concept name (associated with Concept_Number), and the possibility of linking back to the sections of the student assignment which are good, and those needing improvement.

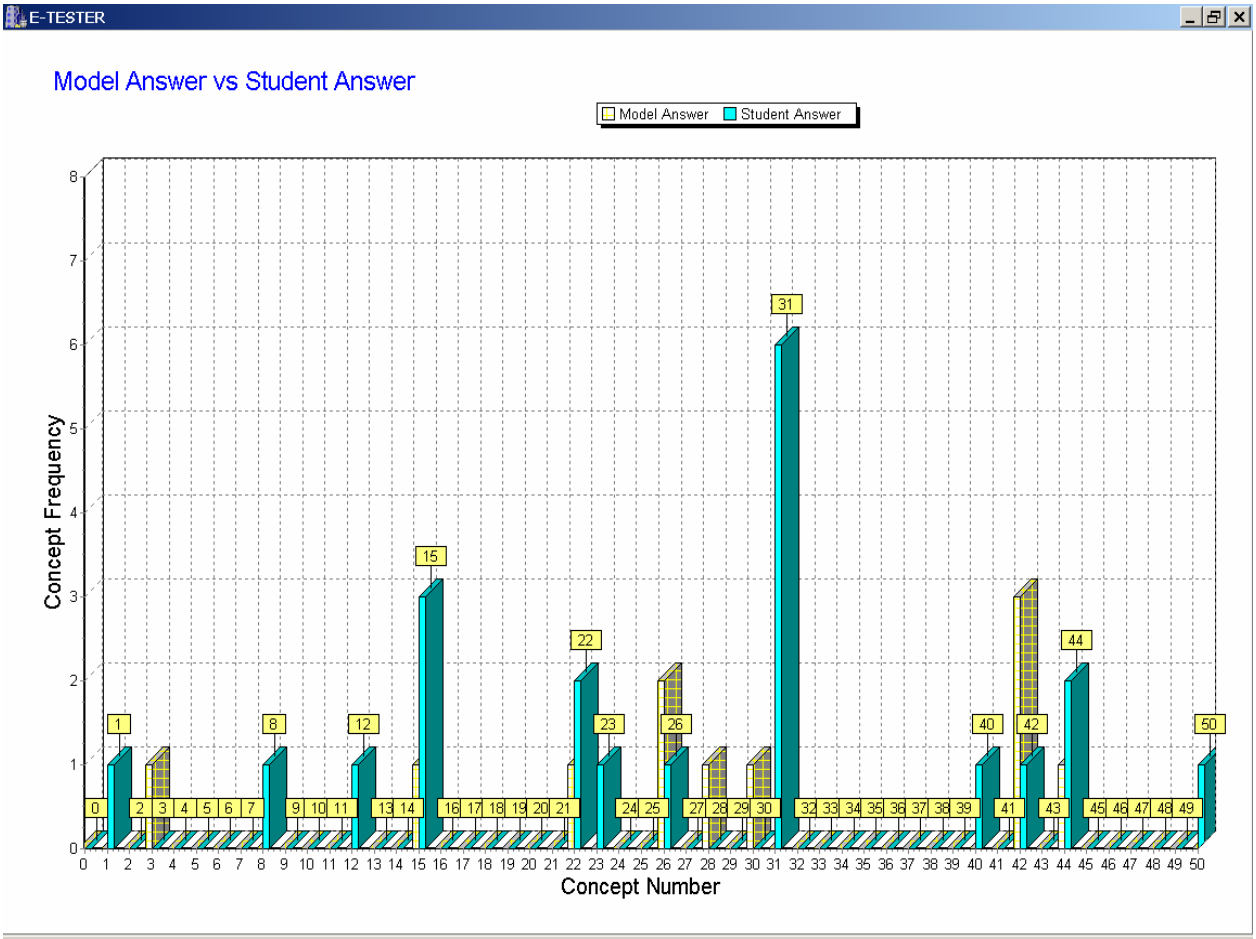


Figure 4 – Concept frequencies: student answer and course content

It is proposed to further develop MarkIT so that these graphs, in computer readable format, will form part of the feedback to the student and the teacher. The teacher will then be able to interactively explain to the student the strengths and weaknesses of the student's answer. If a bar in the graph is double clicked, the thesaurus text for the category represented by the bar will be displayed. The student can then see the amount of discussion that should have been devoted to the topic, and also get a good feel, from the many words in that thesaurus category, how to express that content. A percentage of the discussion above or below the expected amount of discussion will also be displayed.

Summary

MarkIT has been developed to provide automated grading of essay-type documents. Along with its peers in the AEG domain MarkIT performs as well as human graders under certain given conditions. Unlike many of its competitors, MarkIT is now endowed with the added feature of providing meaningful, relevant, and detailed feedback to assist learners improve their performance.

References

Attali, Y. & Burstein, J. (2004). Automated essay scoring with E-rater V.2.0. Paper presented at the *Conference of the International Association for Educational Assessment (IAEA)*, June 13-18, 2004, Philadelphia, USA. Retrieved from http://www.ets.org/research/dload/IAEA.pdf

Callan, J. P., Croft, W. B. & Broglio, J. (1995). TREC and TIPSTER experiments with INQUERY. *Information Processing and Management,* 327-343.

College Board (2002). The new SAT: Implemented for the class of '06. PowerPoint slides posted on http://www.collegeboard.com/prod_downloads/about/newsat/newsat_presentation.ppt

Dreher, H., Scerbakov, N., & Helic, D. (2004). Thematic driven learning. *Proceedings of E-Learn 2004 Conference*, Washington DC, USA, November 1-5. Retrieved from http://www.aace.org/conf/ELearn/

Foltz, P., Laham, D. & Landauer, T. (1999). Automated essay scoring: Applications to educational technology. Retrieved from http://www-psych.nmsu.edu/~pfoltz/reprints/Edmedia99.html

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes*, *25*, 259-284.

Larkey, L. S. (1998). Automatic essay grading using text categorization techniques, *Proceedings of the Twenty First Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Melbourne, Australia, 90-95.

Maron, M. E. (1961). Automatic indexing: An experimental inquiry. *Journal of the Association for Computing Machinery*, *8,* 404-417.

Page, E. B. (1966). The imminence of grading essays by computer. *Phi Delta Kappan*, January, 238-243.

Page, E. B. (1994). Computer grading of student prose, using modern concepts and software, *Journal of Experimental Education*, *62*, 127-142.

Shermis, M. & Burstein, J. (2003). *Automated essay scoring: A cross-disciplinary perspective 2003.* New Jersey, USA: Lawrence Erlbaum Associates.

Shermis, M., Mzumara, H., Olson, J. & Harrington, S. (2001). On-line grading of student essays: PEG goes on the World Wide Web. *Assessment and Evaluation in Higher Education*, *26,* 3. Retrieved from http://taylorandfrancis.metapress.com/media/804PYKUXVJC2076KLQFW/Contributions/H/E/8/4/HE84J5VPEDVRVL3T.pdf

Valenti, S., Neri F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, *2*, 319 – 330. Retrieved from http://jite.org/documents/Vol2/v2p319-330-30.pdf

van Rijsbergen, C. J. (1979). *Information retrieval* (2nd ed.). London: Butterworths.

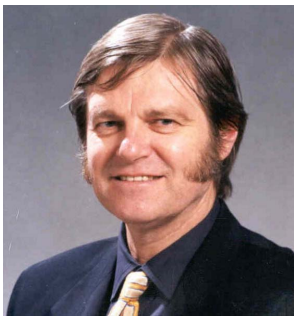
Williams, R. (2001). Automated essay grading: An evaluation of four conceptual models. In M. Kulski & A. Herrmann (Eds.), *New horizons in university teaching and learning: Responding to change.* Perth, Australia: Curtin University of Technology.

Williams, R. & Dreher, H. (2004). Automatically grading essays with Markit©. *Journal of Issues in Informing Science and Information Technology, 1*, pp693-700

Biographies

Robert Williams has over 25 year's experience in the Information Systems industry, as a practitioner, researcher and lecturer. He currently is a lecturer in the School of Information Systems at Curtin University of Technology in Perth, Western Australia. He has extensive experience in systems analysis and design, and programming, on a variety of mainframe, mini and personal computers, and a variety of operating systems and programming languages. Applications he has worked with include mathematical, statistical, bridge and road engineering, financial, corporate resource allocation, business simulation and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading

systems. In 2001 he led a team of researchers in the School of Information Systems at Curtin University of Technology which conducted what is believed to be the first trial in Australia of an Automated Essay Grading system. Robert holds a Bachelor of Arts degree with double majors in Mathematics and Economics from the University of Western Australia, a Graduate Diploma in Computing from the Western Australian Institute of Technology, and a Master of Information Systems degree from Curtin University of Technology.



Heinz Dreher is senior lecturer and research fellow in the School of Information Systems at Curtin University of Technology. He has published in the educational technology and information systems domain through conferences, journals, invited talks and seminars; is currently the holder of Australian National Competitive Grant funding for a 4 year e-Learning project; is participating in a Digital Library project with TU Graz / Austria; is collaborating on Automated Essay Grading technology development, trial usage and evaluation; has received numerous industry grants for investigating hypertext based systems in training and business scenarios; and is an experienced and accomplished teacher, receiving awards for his work in cross-cultural awareness and course design. In 2004 he was appointed Adjunct Professor for Computer Science at TU Graz, and continues to collaborate in teaching & learning, and research projects with local and overseas partners.