Interactive On-line Formative Evaluation of Student Assignments

Heinz Dreher
School of Information Systems, Curtin University of Technology
Perth, Western Australia

h.dreher@curtin.edu.au

Abstract

Automated Essay Grading (AEG) technology has been maturing over the final decades of the last century to the point where it is now poised to permit a transition in 'assessment-thinking'. The administrative convenience of using objective testing to attempt to assess deep learning, learning at the conceptual level, has now been obviated by efficient and effective automated means to assess student learning. Further, the new generation AEG systems such as MarkIT deliver an unprecedented interactive formative assessment feedback capability, which is set to transform individualized learning and instruction as implemented in existing Learning Management Systems (LMS).

Keywords: AEG; Automated Essay Grading; automated formative assessment; conceptual learning; deep learning; formative evaluation; interactive assessment feedback; Latent Semantic Analysis, MarkIT; Normalised Word Vector; NWV.

Introduction

Assessment of student assignments has been an important part of the teaching and learning process long before the advent of desktop computers, Learning Management Systems, and the Internet. Whether the education is traditional format or facilitated in on-line mode, typically, assignments are submitted, assessed or graded, and returned, with a mark or grade, and some comments explaining the possible improvements and applauding the assignment's positive aspects. At times the assignment is returned with an assessment pro-forma or template containing the assessment criteria and weightings, with details of component marks and specific comments in addition to a summary. Creating such feedback is time-consuming and laborious. On the other hand, detailed feedback can be valuable for students wishing to improve, tracking student progress, or for multi-stage assessments in which the successful completion of earlier stages are required for later stages. When there are many students and time is short, feedback detail is reduced, assessment quality may be compromised, and in extreme cases, a 'tick and flick' approach may seem a tantalizing option.

Numerous techniques are available to support carrying out the assessment task to a high standard, but the greatest gain by far is afforded by utilizing computer based augmentation of the assessment process. Often it is the sheer volume of assignments that poses the greatest challenge to the assessment task. In recent years we have seen the development and

refinement of Automated Essay Grading Systems originating with the work of Ellis B. Page (1966) which he carried forward through the 1990s (Page 1994) and which spawned numerous competitor systems such as the Intelligent Essay Assessor (IEA) (Foltz, Laham, & Landauer, 1999), e-rater® (Attali & Burstein, 2004), and MarkIT (Williams & Dreher, 2004).

The Intelligent Essay Assessor's developers describe it as "a set of software tools for scoring the quality of essay content" (Foltz et al. 1999). The IEA has been commercialized by Knowledge Analysis Technologies, which company has recently been acquired by global publisher Pearson Education. Another competitive system, e-rater® is being used by the Educational Testing Service (http://www.ets.org). Whilst IEA and e-rater® are now clearly a commercially competitive products, on the other side of the globe in Perth, Western Australia, a new generation of AEG system has been taking shape since the turn of the century, and is now at what may be termed a pre-competitive stage.

The developers of MarkIT have been successful in devising a semantic representation method which has been termed *Normalised Word Vector* (NWV) (Williams, 2006) – it is interesting from the view of deployment in Learning Management Systems because it is computationally more efficient than prior systems, while still performing at the level of human graders. Another innovative aspect of MarkIT is the integral interactive formative assessment component that provides immediate feedback to students and teachers. For an evaluation of various conceptual models used in AEG technology see Williams (2001). Valenti, Neri, and Cucchiarelli (2003) provide a description of AEG research up to the turn of the century. Williams (2006) analyses the NWV technology relative to Latent Semantic Analysis (LSA). The interactive formative assessment feedback aspect of this technology is the focus of this paper.

In summary, we can see that there is growing interest in finding some way to support the educator in assessment of deep, conceptual learning as distinct from the relatively more superficial learning so easily checked with computer based objective testing mechanisms. Despite the interest, the educational community has in main remained ignorant of the all important advances in the Automated Essay Grading domain which Ellis B. Page (1966) introduced us to some 40 years ago.

Contemporary LMS Assessment Support Capability

When Learning Management Systems are compared and evaluated, scrutinized criteria include: usability and navigation; content development; discussion space; group interaction; communication; calendar and other study tools; multimedia; site administration; and of course monitoring student participation and progress (Lewis et al., 2005). However, even this last mentioned criterion, monitoring, is bereft of learning appraisal in a qualitative sense, never mind the all important formative evaluation aspect with its feedback cycle to promote further learning. Are we to conclude formative evaluation is unimportant? Serious educators know that immediate feedback on student learning is central to success. What we may conclude is that computer based formative evaluation of deep learning, as demonstrated in the writing of essays, has been too difficult and demanding relative to the leverage obtained by supporting the other mentioned LMS evaluation criteria. The problem has been that the required technological breakthrough is just now emerging as this new century begins. And even as the technology emerges, the issue of communicating the news and winning over the skeptics cannot be ignored.

In their article on supporting formative assessment with e-Learning tools, Heinrich and Lawn (2004) admit defeat when it comes to automating formative assessment, but are highly motivated to investigate support for human markers performing such assessments. Their MarkTool system provides assistance in document management, and very importantly some provision for feedback annotations and marks, although these are pixel-level objects and cannot be subjected to semantic manipulation. Whilst their motivation is clearly consistent with good educational theory, prac-

tice, and intentions, one might say, sadly their implementation efforts have been progressed in ignorance of the AEG technology in general and MarkIT in particular.

Paul Blayney (2005), was also ignorant of AEG when he developed his spreadsheet-based formative and summative assessment system, but as he works in a quantitative discipline, the need for automated support for free-text grading is not needed. His system had its roots in the same decade – a learning technology 20th century fin de siècle as it were – that AEG technology was being proven. What makes his system so dramatically effective is that accounting student assignments have numbers or formulae as answers, and these are of course simply checked by automatic means. Actually, the Blayney system is far from simple. It is very sophisticated in its design intent, which is to provide timely, and interactive constructive formative feedback via iterations of individualized assignments. The system is spreadsheet based and able to infer deep understanding of accounting concepts by tracking formulae that must be entered and used in a particular logical sequence to complete the non-trivial questions he sets his students. Out of the entire literature that has been accumulated on the matter of "interactive assessment and feedback" his system (apart from MarkIT) is the only one which automatically provides truly interactive real-time formative assessment feedback to students.

The considerable interest in deploying computer based formative assessment systems has produced some curious solutions. For example, Yasuko (2004) describes his system of "formative assessments" which relies on objective testing, but adds a contextual and temporal analysis of learning object accesses as stored in activity logs. The advantage claimed is that by analyzing logs of e-Learning sessions and comparing them with test results "it is possible to know the transition of the level of comprehension of each student" (Yasuko, 2004). Presumably, this indirect but automated approach permits one to appraise the deep learning. Some insight into deep learning may be derived from such a convoluted analysis and tenuous link (at best) between "transition of the level of comprehension" (Yasuko, 2004) and deep conceptual learning. One may contemplate whether Yasuko would persist with such a scheme if essay-type assignments, which contain the evidence of deep learning, could be automatically assessed. There is no degree of interactivity in the Yasuko system.

In their article "Thematic driven Learning", Dreher, Scerbakov, and Helic (2004) explain how their Learning Management System, WBT-Master, was endowed with assignment assessment and feedback support functions, but this still required the evaluator to read through hundreds of assignments. The system did however place prior assessments (assignment, marks and feedback) in an onscreen window juxtaposed with the current assignment being assessed. This permitted the current assignment to be assessed in the light of previous feedback in addition to the usual considerations, consistent with the incremental improvement goal of formative evaluation.

Advances in technology support embedded in alternate paradigms as compared with those adopted by the majority of LMS are beginning to emerge. For example, D'Mello et al. (2005) at the University of Memphis have developed "a fully automated computer tutor that simulates human tutors and holds conversations with students in natural language". This of course requires an ability to 'understand' or assess free-text or essay-type student responses, albeit in the form of short dialogue components. AutoTutor, as their system is known, uses the LSA technology mentioned earlier, and since the textual elements are restricted to tens of words, a paragraph or two, computational efficiency is less of an issue as in the case of grading thousand-word essays in real time, as is possible with MarkIT. Another important feature of the AutoTutor system is its interactivity, which its authors remark was inspired by the constructivist theories of learning and the commonsense knowledge that collaboration and feedback iterations are observable in the work of good teachers.

The Transition in 'Assessment-Thinking'

The so called objective testing schemes – short answer, true/false, and multiple-choice, were devised entirely as a matter of convenience. It is acknowledged they cannot readily assess the degree of deep learning that is evident from an essay-type assessment item, and yet objective testing is deployed in most educational settings, and especially where large numbers of students are being catered for. This is not to say that objective testing is not useful, but rather that its use has often been determined by administrative convenience, in contradistinction to educational imperative. Few Learning Management Systems (e.g. WebCT, and Blackboard, being the most popular) offer any technological support for essay grading, whilst they offer excellent support for objective testing. Large educational book publishers and their prominent textbook authors, orient their products to objective testing.

The assessment of written assignments or essay-like answers to exam questions is such a laborious task that for large student cohorts, educators have been forced into choosing computer-scored assessments. The argument has been that speed, accuracy, and consistency of such assessment outweighs the major drawback – the difficulty of assessing deep learning as opposed to knowledge recall and recognition. Even prior to the prevalence of desktop computers and online systems it was relatively straightforward to arrange a computer-scored test, even for thousands of students. Users of such systems and products are prepared to trade the quality of feedback against convenience. The result is inferior learning outcomes relative to what is possible by utilizing more educational resource or by deploying superior educational technology.

When students learn, they need to know how they have progressed with respect to some standard or benchmark. Teachers also need to know about their students' performance so they may devise progression or remedial strategies as necessary, or certify performance – the former being formative and the latter summative in intent. Our concern here is the formative evaluation of learning. That is, we would like to be able to provide relevant feedback so as to permit further learning, or re-learning as necessary. Since learning is individual – with reference to content, process, and time, it would be useful to have an assessment system that is tailored to individual students. Such a system is called "teacher", actually, one would need to qualify that perhaps and specify "good teacher". There are more qualifications required however in today's educational setting – minimizing cost, and maximizing efficiency, student numbers, teacher performance, and return on educational investment are at least as important as quality of educational outcomes. The problem with the system known as "teacher" is that it is expensive.

The skeptic will need to imagine, whilst others may take it on faith (at least for the moment), that AEG systems can now perform as good as humans (Palmer, Williams, & Dreher, 2002) when it comes to assessing or grading essays, in addition to providing feedback which is superior to that which humans can provide within realistic resource constraints. Williams and Dreher (2005) have given an insight in how the MarkIT system provides formative assessment feedback at the semantic and conceptual level. Guetl, Dreher, and Williams, (2005) have described the significance of the advantage of deploying auto-generated essay-type question and answer assessment.

The New Wave – Automated Formative Evaluation

Clearly, the old thinking is giving way to the new. As explained above, the new technology known as Automated Essay Grading (AEG) is now available commercially and is beginning to find its way into educational practice. In the years ahead one may witness the adoption of AEG to an extent similar to that of the now relatively predominant use of technology to create assignments – Word-processing technology has replaced the pen and paper method. The collegiate question "what are you doing this weekend?" will no longer induce a response of "assignment grading!", but rather some response which indicates that the onerous educational imperative of

formative assessment has been supported by AEG leaving the teacher to relax in preparation for meaningful educational activities the new week may bring.

One may well ask what are these “meaningful educational activities” if technology can now do so much of the teacher’s work. A typical scenario with the use of the Western Australian system MarkIT (www.essaygrading.com) follows:

Students write their essays as usual. Naturally they must be in computer readable form, and at this juncture in English (plans for a European languages version of MarkIT are well advanced). Preparatory arrangements need to have been made so that the MarkIT web-site is configured to cater to the particular essay assignment. Clearly, considerable work goes into generating a model answer (a teacher activity, although this may also be supported with technology, and is a MarkIT project agenda item for 2006), administrative and workflow functionality to support assignment upload, and feedback access by both student and teacher, and other necessary arrangements relating to ‘return-on-investment matters’.

Essays are submitted according to an agreed schedule, assessed on-the-fly or in batch mode, and results made available via the same interface, a standard Java enabled browser, as was used for assignment upload. The results comprise the usual quantitative parameters (assignment score according to pre-supplied criteria and weighting scheme – model answer; and via generally available technology, length, readability, spelling, grammar, and similar characteristics).

However, far more important for formative evaluation, MarkIT produces an interactive graph comparing the assignment concepts as represented in the model answer with that of an individual student. Students and teachers may see at a glance why the particular assessment or grade has been suggested or assigned by MarkIT. More, they may interact with the bars in the graph probing into the detail relating to each concept represented by the graph, which interaction yields the corresponding structured thesaurus entry. Interested readers are encouraged to visit www.essaygrading.com and explore the demonstration under menu <MY ACCOUNT> then <Guest Account> and experience the interactivity at the level of assignment and model answer concept representation and explanation.

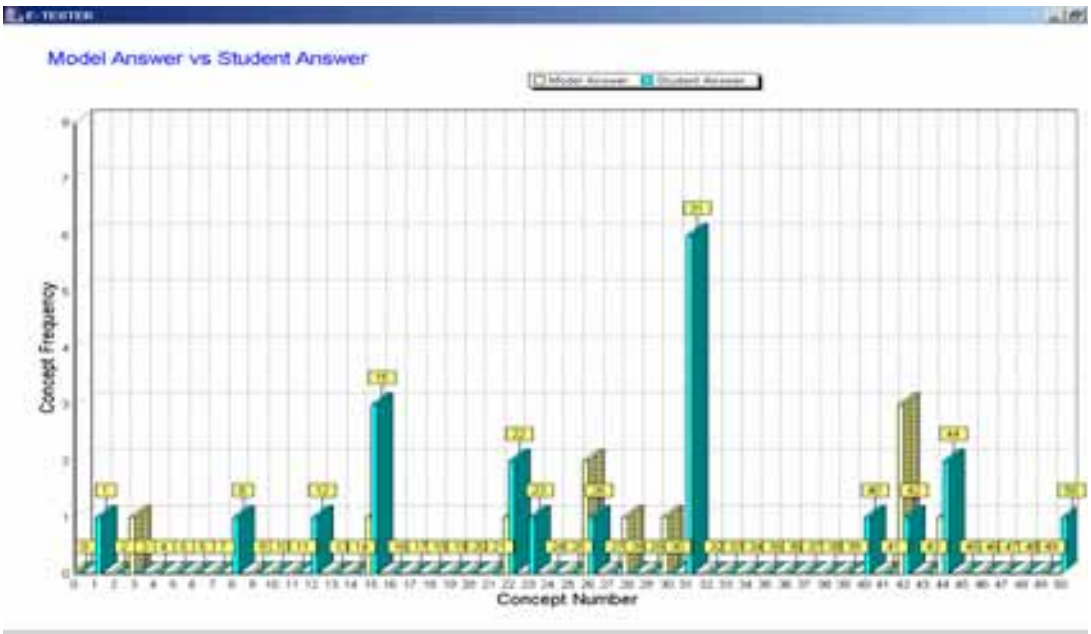

Figure 1: Concept frequencies: student answer and **course content**

In ***Error! Reference source not found.***, a graphical feedback element as designed into MarkIT, juxtaposes the 'concepts' associated with model answer and student answer. Naturally, the better the correspondence between the concept representation in both, the better the score. For example, the tallest bar (Concept_Number 31) reveals a concept_frequency of 6 (vertical axis) for the student answer (dark bar), where the model answer called for no discussion on this topic or concept. One may conclude that the student has introduced irrelevancies into the answer; or perhaps this is what can be termed an error on the student's part. MarkIT concludes likewise. Concept_Number 26 has a better match between model and student answer, indicating the student has learned relevant material. It is important to understand that MarkIT is not performing simple word matching, but rather employs the Normalised Word Vector technology devised by Williams (2006) to make comparisons at the semantic level, and thus allows one to assign a grade to conceptual learning.

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 that 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.

Error! Reference source not found. shows the graph produced by MarkIT when a user is exploring the reason for the mismatch between concept representation in student and model answers. In this case the concept labeled "Working Class" is being explored (see the shaded pair of bars in

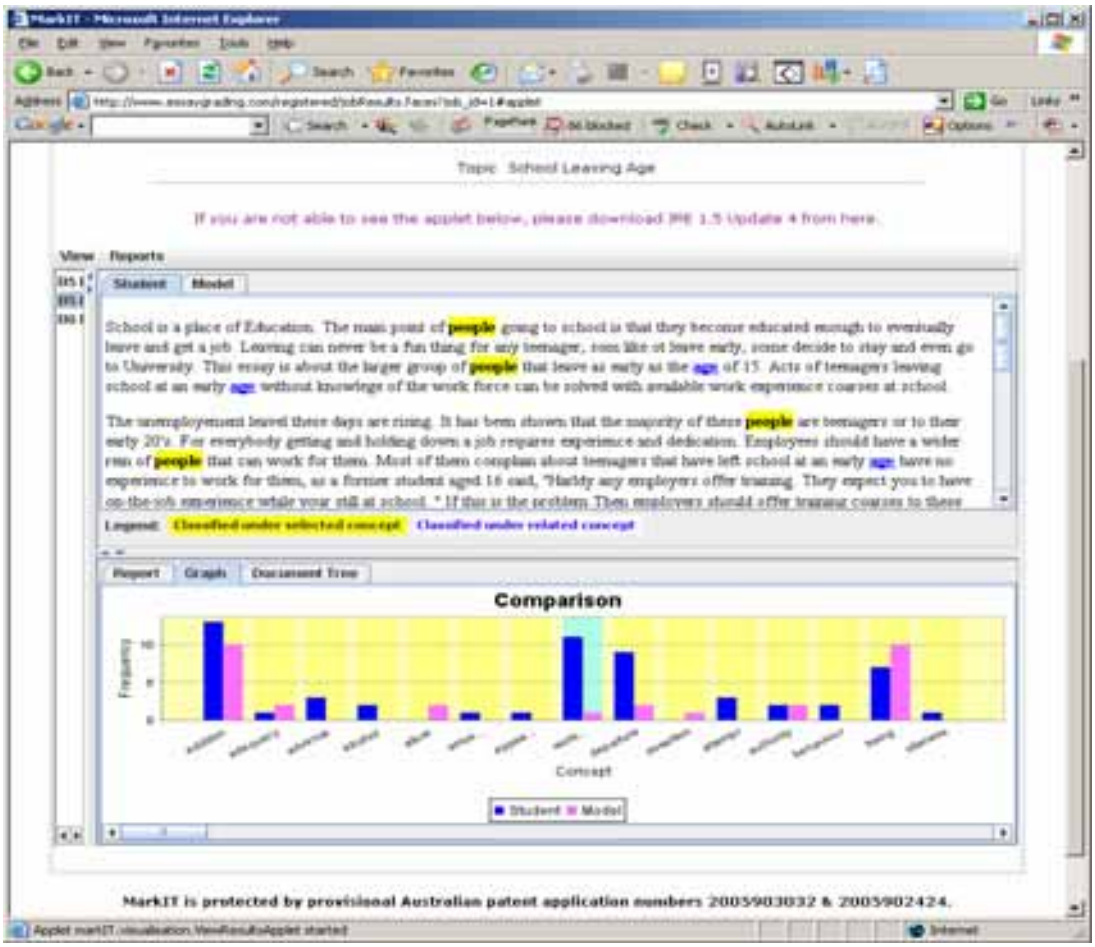

Figure 2: Interactive concept comparison (student vs model)

the middle of graph with label “worki…” just above the word “Concept” used as the legend for the horizontal axis.

There is a large mismatch between student (concept_frequency = 11) and model (concept_ frequency = 1) and the user wishes to see an explanation. Some activation method, for example a double-click, produces the window depicted in ***Error! Reference source not found.***. Now, the student, perhaps in conjunction with the teacher can review what may have been a conceptual error. Obviously not all the words reproduced from the thesaurus would be relevant and further refinement of this aspect of MarkIT, using the technique of ontological filtering for example, is being pursued. In the interim, users need to be discerning in the use of feedback. As with any technology, it is produced in the ‘service of mankind’ and in our case, is made to augment teachers so that this special resource may be applied to the truly human aspects of students and learning.

The MarkIT designers have provided for interactivity at the within-assignment-concept level, a

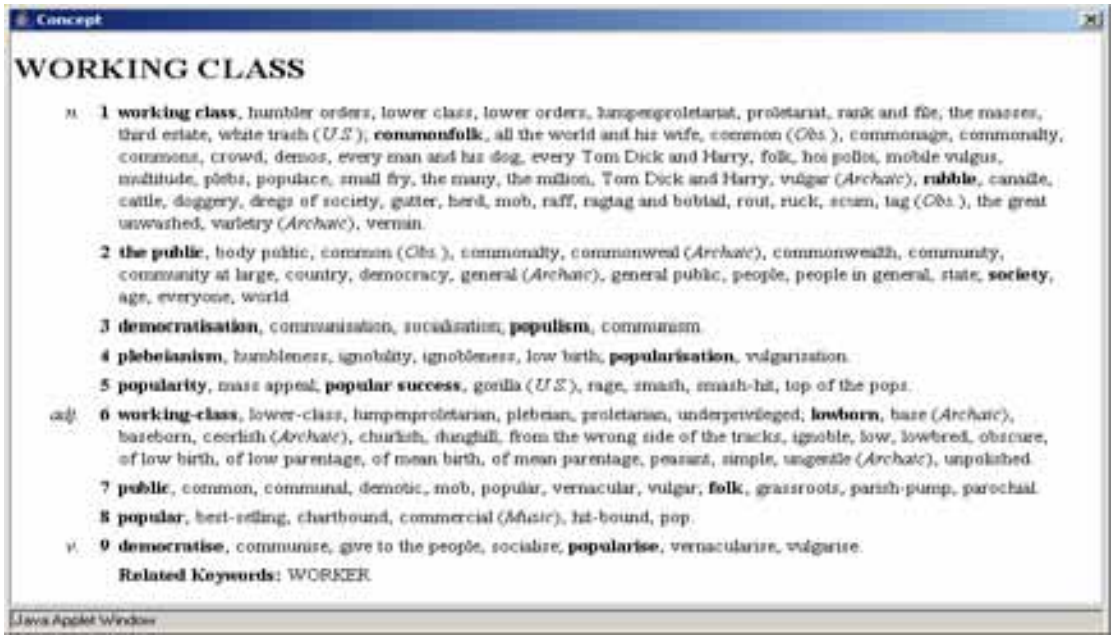

Figure 3: Within-assignment-concept level interactive explanation

feature already implemented as described above, but one can readily imagine an iterative cycle of assignment submission, improvement, resubmission, where MarkIT provides feedback based on the model answer and the series of repeated assignment submissions. Under these conditions a student’s rate of learning may also be appraised. Such ideas, and more, remain for further discussion and development. Ideas from interested readers are welcomed.

Never before has there been such a level of feedback available to students and teachers – in such detail and interactively, except perhaps in one-on-one student-tutor situations. We all know such arrangements are beyond the financial reach of the vast majority of students.

Conclusion

The proposition that human teachers can now be relieved of, or strongly supported in, the essay grading task is a reality in this new century. Skeptics still abound, but as AEG systems become embedded in the new wave of LMS, they too will experience the actual results of automated support for assessing essays and generating interactive visual and formative feedback for their stu-

dents. In time, the skeptics will surely make the transition in assessment-thinking and help their students perform to maximize individual potential within reasonable resource constraints. As explained in this paper, the *New Wave of Automated Formative Evaluation of Student Assignments* is here.

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

Blayney, P. (2005). Interactive assignments used for formative & summative assessment. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005* (pp. 1209-1214). Norfolk, VA: AACE. Retrieved from http://www.editlib.org/index.cfm

D'Mello, S., Craig, S., Witherspoon, A., Sullins, J., McDaniel, B., Gholson, B., & Graesser, A. (2005). The relationship between affective states and dialog patterns during interactions with AutoTutor. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005* (pp. 2004-2011). Norfolk, VA: AACE. Retrieved from http://www.editlib.org/index.cfm

Dreher, H., Scerbakov, N., & Helic, D. (2004). Thematic driven learning. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2004* (pp. 2594-2600). Norfolk, VA: AACE. Retrieved from http://www.editlib.org/index.cfm

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

Guetl, C., Dreher, H. & Williams, R. (2005). E-TESTER: A computer-based tool for auto-generated question and answer assessment. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2005* (pp. 2929-2936), October. E-Learn 2005. Retrieved from http://www.editlib.org/index.cfm

Heinrich, E., & Lawn, A. (2004). Onscreen marking support for formative assessment. In *Proceedings of World Conference on Educational Multimedia, Hypermedia and Telecommunications 2004* (pp. 1985-1992). Norfolk, VA: AACE. Retrieved from http://www.editlib.org/index.cfm

Lewis, B. A., MacEntee, V. M., DeLaCruz, S., Englander, C., Jeffrey, T., Takach, E., Wilson, S., & Woodall, J. (2005). Learning management systems comparison. Proceedings of the *2005 Informing Science and IT Education Joint Conference*, Flagstaff, Arizona, USA – June 16-19. Retrieved from http://proceedings.informingscience.org/InSITE2005/P03f55Lewis.pdf

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

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

Palmer, J., Williams, R., & Dreher, H. (2002). Automated essay grading system applied to a first year university subject – How can we do it better? *Proceedings of Informing Science 2002 Conference*, Cork, Ireland, June 19-21. Retrieved from http://proceedings.informingscience.org/IS2002Proceedings/papers/Palme026Autom.pdf

Valenti, S., Neri F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education, 2*, 19 – 330. Retrieved from http://jite.org/documents/Vol2/v2p319-330-30.pdf

Williams, R. & Dreher, H. (2004). Automatically grading essays with Markit©. *Issues in Informing Science and Information Technology, 1*, 693-700. Retrieved from http://articles.iisit.org/092willi.pdf

Williams, R. & Dreher, H. (2005). Formative assessment visual feedback in computer graded essays. *The Journal of Issues in Informing Science and Information Technology, 2*, 23-32. Retrieved from http://2005papers.iisit.org/I03f95Will.pdf

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*. Curtin University of Technology, Perth, Australia. Retrieved from http://lsn.curtin.edu.au/tlf/tlf2001/williams.html

Williams, R. (2006). The power of normalised word vectors for automatically grading essays. Paper to be presented at InSITE 2006, Manchester, England, June 25-28. http://2006.informingscience.org

Yasuko, N. (2004). Innovation of the formative assessments approach using WBT. In *Proceedings of World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education 2004* (pp. 453-460). Norfolk, VA: AACE. Retrieved from http://www.editlib.org/index.cfm

Biography

Heinz Dreher is senior lecturer and research fellow in the School of Information Systems, part of the Curtin Business School at Curtin University in Perth, Western Australia. 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 collaborating on Automated Essay Grading technology development, trial usage and evaluation, resulting in the creation of MarkIT (www.essaygrading.com); 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 Graz University of Technology, and continues to collaborate in teaching & learning and research projects with European partners.