Automated Essay Grading System Applied to a First Year University Subject – How Can We do it Better?

John Palmer, Robert Williams & Heinz Dreher
Curtin University of Technology, Perth, WA, Australia

palmer@netunltd.com.au diagnosys@optusnet.com.au dreherh@cbs.curtin.edu.au

Abstract

Automated marking of assignments consisting of written text would doubtless be of advantage to teachers and education administrators alike. When large numbers of assignments are submitted at once, teachers find themselves bogged down in their attempt to provide consistent evaluations and high quality feedback to students within as short a timeframe as is reasonable, usually a matter of days rather than weeks. Educational administrators are also concerned with quality and timely feedback, but in addition must manage the cost of doing this work. Clearly an automated system would be a highly desirable addition to the educational tool-kit, particularly if it can provide less costly and more effective outcome.

In this paper we present a description and evaluation of four automated essay grading systems. We then report on our trial of one of these systems which was undertaken at Curtin University of Technology in the first half of 2001. The purpose of the trial was to assess whether automated essay grading was feasible, economically viable and as accurate as manually grading the essays. Within the Curtin Business School we have not previously used automated grading systems but the benefit could be enormous given the very large numbers of students in some first year subjects.

As we evaluate the results of our trial, a research and development direction is indicated which we believe will result in improvement over existing systems.

Keywords: assessment, assignment, automatic, essay, grading, marking, plagiarism

The Problem

Teaching staff around the world are faced with a perpetually recurring problem: how do they minimise the amount of time spent on the relatively monotonous tasks associated with grading their students' essays? With the advent of large student numbers, often counted in thousands in first year common core units, the grading load has become both time consuming and costly. A system that can automate the tasks is currently just a dream for most staff. One of the most thankless tasks in all academia is that of grading, particularly when there is no need to supply individual feedback as in the case of examination grading.

At the Curtin Business School we have about 2,000 first year students each year in several countries in the Australasian region. There are students in Malaysia, Singapore, Hong Kong and elsewhere in the world, including remote & remote Australia, all taking the same subjects but taught by staff local to their place of study or via distance education. To maintain consistency, all examination grading is centralized in Australia but the grading load on the Australian lecturers is horrendous.

As each of the 2000 students undertake an average of eight subjects each year, the number of final ex-

aminations could be as large as 16,000 scripts requiring grading. In addition to this already extensive problem we can add an additional 16,000 to 32,000 assignments of which possibly half would be graded too late to provide formative evaluation feedback to the students.

A system that can automate the grading process tasks is currently just a dream for most staff. However, there are a number of systems emerging from the research laboratories and finding their way into the production environment.

Four of these systems are described and evaluated in the following sections. The results of a trial of one of the systems is then reported.

Production Automated Essay Grading Systems

One of the earliest mentions of computer grading of essays in the literature was 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

Description

The idea behind PEG is to help reduce the enormous essay grading load in large educational testing programs, such as the SAT. 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 graded 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.

PEG has its origins in work begun in the 1960's by Page and his colleagues (Page, 1966).

> "…we coined two explanatory terms: *Trins* were the in*trinsic* variables of interest – fluency, diction, grammar, punctuation, and many others. We had no direct measures of these, so began with substitutes: *Proxes* were ap*prox*imations, or possible correlates, of these trins. All the computer variables (the actual counts in the essays) were proxes. For example, the trin of fluency was correlated with the prox of the number of words." (Page 1994, p. 130)

The multiple regression techniques are then used to compute, from the proxes, an equation to predict a score for each student essay. In the research reported in Page (1994), the goal was to identify those variables that would prove effective in predicting human rater's scores. Various software products, including a grammar checker, a program to identify words and sentences, software dictionary, a part-of-speech tagger, and a parser were used to gather data about many proxes.

Evaluation

Details of most of the predictive variables are not given in Page's work. However, amongst the variables found useful in the equation were the fourth root of the number of words, sentence length, and a measure of punctuation. One set of results, based upon a regression equation with twenty-six variables, showed correlations between PEG predicted scores and human rater scores varying between 0.389 and 0.743.

E_RATER

Description

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 criteria 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.

Evaluation

The system has been evaluated by Burstein, Kukich, Wolff, Lu & Chodorow (1998) and has found that it can achieve a level of agreement with human raters of between 87% and 94%, which is claimed to be comparable with that found amongst human raters. For one test essay question the following predictive feature variables were found to be significant.

1. Argument content score
2. Essay word frequency content score
3. Total argument development words/phrases
4. Total pronouns beginning arguments
5. Total complement clauses beginning arguments
6. Total summary words beginning arguments
7. Total detail words beginning arguments
8. Total rhetorical words developing arguments
9. Subjunctive modal verbs

Intelligent Essay Assessor - an LSA based system

Description

Latent Semantic Analysis (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.

A matrix represents the words and their contexts. 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) have 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.

Evaluation

Landauer et al. (1998), report that LSA has been tried with five scoring methods, each varying the manner in which student essays were compared with sample essays. Primarily this had to do with the way cosines between appropriate vectors were computed. For each method an LSA space was constructed based on domain specific material and the student essays. Foltz also reports that LSA grading performance is about as reliable as human graders (Foltz, 1996). Landauer reports another test on GMAT essays where the percentages for adjacent agreement with human graders were between 85%-91% (Landauer, 1999).

The Text Categorisation Technique (TCT)

Description

Larkey (1998) implemented an automated essay grading approach based on text categorisation techniques, 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, p. 90)

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 tech-

nique 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 (see also the discussion on PEG above), and the average sentence length.

Larkey conducted a number of regression trials, using different combinations of components. He also used a number of essay sets, including essays on social studies (soc), where content was the primary interest, and essays on general opinion (G1), where style was the main criteria for assessment. The results presented here are for these two essay sets only.

Evaluation

When all the criteria for assessment were used the proportion of essays graded exactly the same as human graders was 0.60 and scores adjacent (a score one grade on either side) was 1.00. For the general opinion essays the corresponding figures were 0.55 and 0.97. The system performed remarkably well.

The Trial at Curtin University of Technology

During the first semester of 2001 a trial of an automated essay grading system was conducted at Curtin University of Technology in Perth, Western Australia. One subject was chosen, a first year introduction to Information Systems, where we had about 1,000 students available to participate. Unfortunately the semester had already started by the time we were able to undertake this research. This meant that all assessment had already been determined. Once assessment has been published the policy at Curtin University is that it cannot be changed without the consent of the majority of students. In order to gain that consent and ensure a high rate of response to our trial, we announced that an additional voluntary essay-type question would be available for bonus marks. Needless to say we had a high rate of response.

The system we were trialling was an American system that required two hundred manually graded essays as input to their grading system. Between the three researchers we graded about 70 papers each and sent the electronic copies along with the marks to the US site. About another 330 ungraded essays were then forwarded to the site for grading.

A number of interesting outcomes were noticed when we analysed all the grades. Firstly, the grades from the three researchers as indicated in figures Fig.1, 2 and 3, had no significant difference in either absolute marks awarded or the standard deviation of marks. Grader “A” had always considered himself a “hard” grader and considered grader “B” rather soft.

However, the purpose was not to check our own grading but to see how consistent the computer system handled the assessment. We were delighted to have our suspicions confirmed; the computer system had the same mean and standard deviation of marks as the three of us. (see Fig. 4) We were satisfied that it worked.

There was an additional and quite unexpected result from the test. The system picked up several cases of plagiarism that we had failed to notice. In this case the plagiarism was really that of one student copying the work of another student rather than from extracting text from another source.

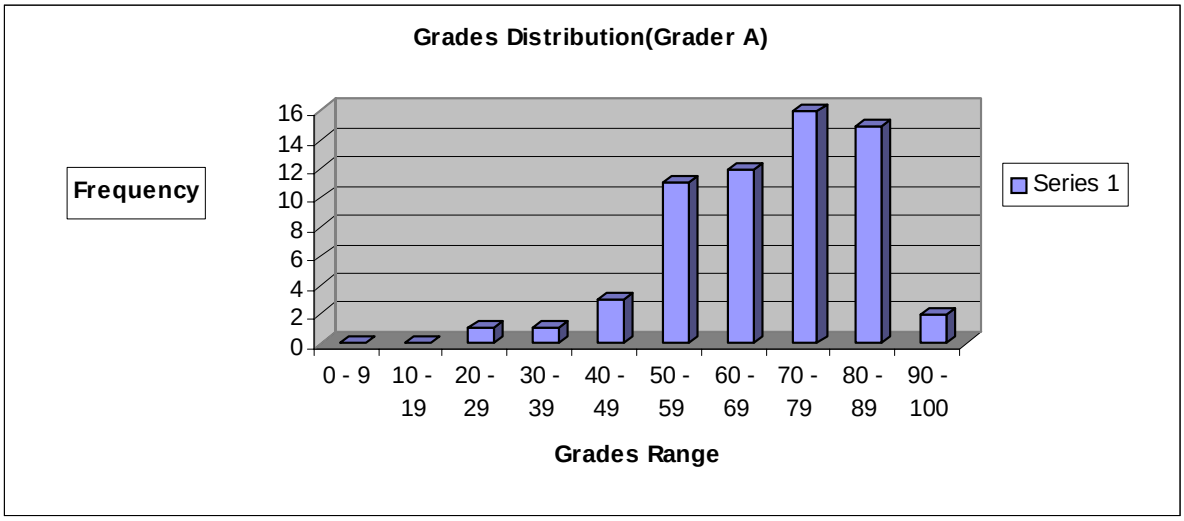

Fig. 1 Results of grades for Grader A

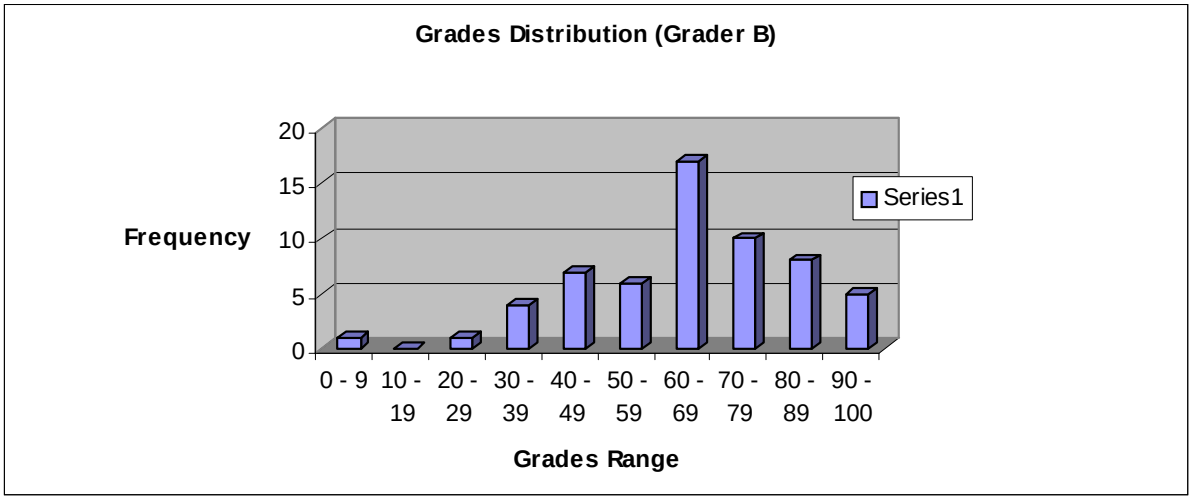

Fig. 2 Results of grades for Grader B

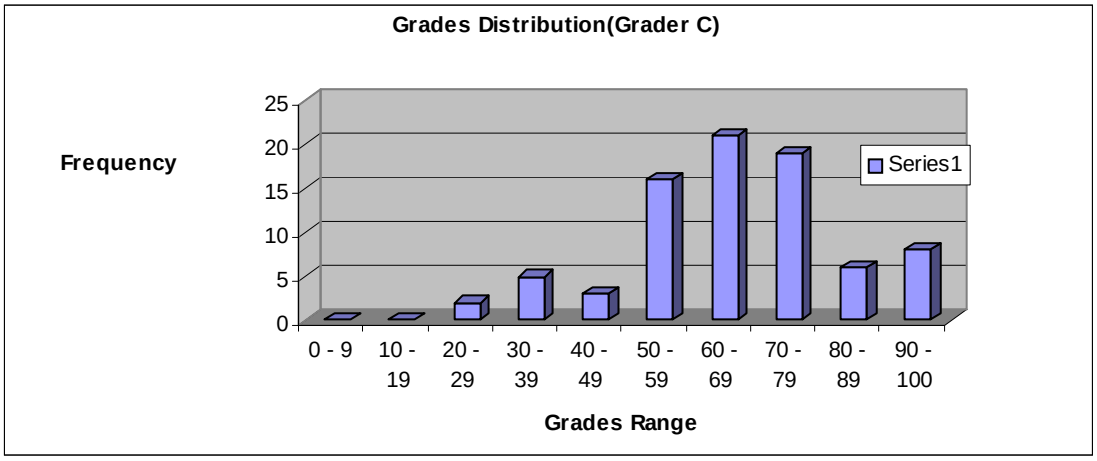

Fig. 3 Results of grades for Grader C

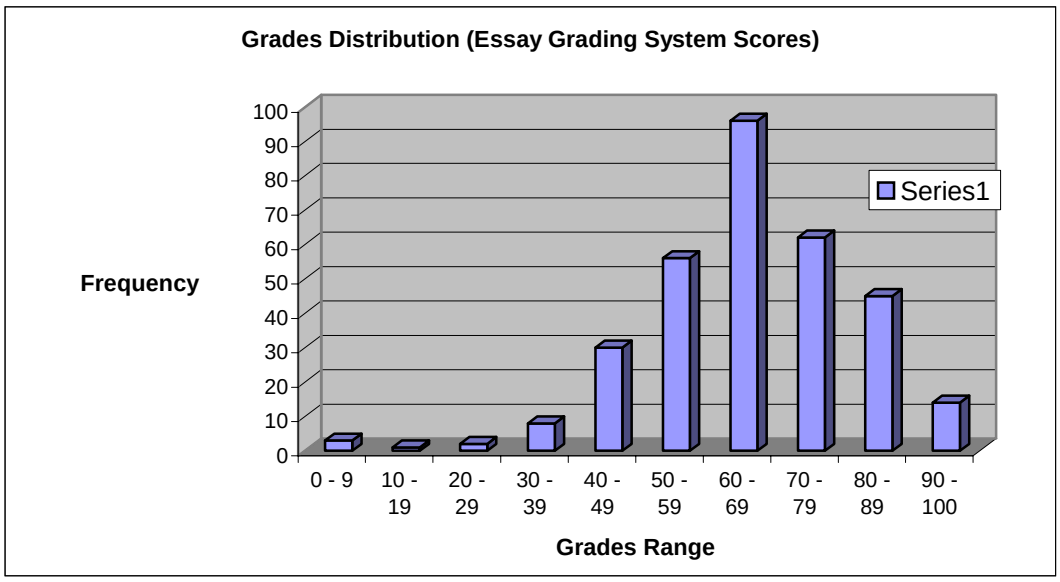

Fig. 4 – Results of grades from the Essay Grading System

The Weaknesses of the System

There are two important weaknesses and one minor weakness for our purposes in the system that we trialled. The first weakness is that for a successful implementation, one needs to manually grade 200 essays and feed them into the system. The computer will then accurately and dependably grade as many more essays on that topic as is required. In small classes of less than a few hundred students it becomes impractical.

The second weakness is the cost of using the system. As the system was American we had to pay in $US. With the exchange rate so poor it cost about A$11,400 to grade a few hundred essays. This is simply not cost effective. If we were to use the same essay for several semesters then the per-unit grading costs would reduce substantially. However it is highly unlikely that we would want to use the same essay questions in consecutive semesters or even twice ever.

There is a third factor. The system is run at a site in the USA rather than on our own computer network at Curtin University. There is some lack of control and potential security risk in having the process run remotely.

Costing Considerations

Ideally the system would be reasonably inexpensive, and certainly far cheaper than hiring grading staff. The grading system would be based on a single all-inclusive model answer supplied by the lecturer. Obviously the system would need to assess with the same degree of accuracy as a manual grader. And finally, the system should be available to be run in-house on a PC or central server.

We currently pay exam graders at a rate of about A$25 per hour. It should be possible to reduce the cost of grading through an automated essay grading system by 90%. Our single experience with the American system as described above was that it cost about A$33 per essay of up to two pages in length.

Based on supplying 200 graded essays at a cost of A$3 per essay, the initial cost before paying for the grading service would be A$600. The grading service costs that we experienced were another A$10,800, bringing the total to A$11,400. In the ideal case it would be beneficial for a University to own the grading system so the costs could be spread across many subjects and many departments. Even if the initial cost was in the thousands of dollars, the cost per essay or exam would become trivial.

There are economies of scale associated with the system, in that up to 2000 essays could have been graded for the A$11,400, but we did not have this number to grade, and so did not gain these benefits. If we had 2,000 essays to grade, the automated essay grading system would still have cost A$5.70 per essay, almost double the cost of grading manually.

Limitations to any Automated Grading System

To utilize any Automated Grading System the raw data, essays or examination answers, would need to be in a form that was computer readable. The most obvious form of this would be electronic documents in Word format. This is easily enough achieved where the student could write the essay on a computer. However, when students sit for examinations this is normally done at desks with paper and pen. The resulting examination script is not easily transferred to a computer readable medium. On the other hand we see that it is possible to have students sit an exam in a computer laboratory and submit their examination papers electronically. It would be difficult to have large numbers sit the exam simultaneously but it is not impractical to have two groups of students where as soon as the first group completes, the second group starts. In this way, with lab facilities of 200 PC's the same examination could be sat by up to 400 students without compromising the examination paper.

Another possibility would be to give the students a take-home examination due within 24 hours. Any number of students would then be able to sit the exam at the same time and submit the exam papers electronically.

Another serious limitation to an essay grading system is that it grades a students' knowledge of a given set of material. The model answer would contain only a set body of knowledge and would grade the student on the part of that knowledge the student was able to demonstrate. This may be acceptable in the early years of a course but probably not in more advanced studies.

Conclusion – Where we are Heading with our Research

We are developing an essay grading prototype system that overcomes some of the problems mentioned above. Our approach eliminates the need to grade 200 essays by humans - the prototype will work with one model answer. Secondly the system will operate on a standard Windows PC, and therefore can be shrink-wrapped for widespread distribution and local use.

The system relies on building a propriety representation of the knowledge contained in the model answer, and having the appropriate marks allocated (by the teacher) to the appropriate sections of the model answer. A student essay is processed using a combination of NLP techniques to build the propriety knowledge representation of it as well. 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.

The system will provide feedback to the student about topics that the student did not cover, or failed to cover adequately.

The prototype system is under construction, and its performance will be evaluated by processing the essays processed by the essay grading system used in the trial reported above, and the relative performances compared. We hope to publish further details about the system and its performance at a later stage.

References

Burstein, J., Kukich, K., Wolff, S., Lu, C. & Chodorow, M. (1998). Enriching Automated Essay Scoring Using Discourse Marking, Proceedings of the Workshop on Discourse Relations and Discourse Graders, Annual Meeting of the Association of Computational Linguistics, August, Montreal, Canada.

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

Foltz, P. W. (1996). Latent Semantic Analysis for Text-Based Research, Behavior Research Methods, Instruments and Computers, 28, 197-202.

Landauer, T. K. (1999). E-mail communication with author 8th June.

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.

van Rijsbergen, C. J. (1979.) Information Retrieval, 2nd ed., Butterworths, London.

Biographies

John Palmer. Having initially been educated in Canada as a Chartered Accountant and later as a Systems Engineer with IBM, John Palmer has been lecturing in Australia in the field of Information Systems at Curtin University of Technology, Western Australia for nearly thirty years. He has specialised in the areas of IT/IS Education and Computer Security. He has published widely in these areas and is now involved in research related to the automation of marking "essay type" student work.

Robert Williams has over 25 years experience in the Information Systems industry, as a practitioner, researcher and lecturer. 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 and educational systems. He has published a number of articles on system users' personalities and satisfaction, decision support systems, and automated essay grading systems.

Heinz Dreher has been working in the Information Technology Systems domain for 32 years. His first position was as computer programmer. This was followed with a move into the tertiary education sector in 1972 as senior tutor in Electronic Data Processing (EDP). Dr Dreher has expertise in Hypertext/Hypermedia systems and textual-knowledge-based systems, Computer Supported Co-operative Work (CSCW), Computer Mediated Communications (CMC), Project Management, Prototyping systems, Human Problem Solving Strategies, Decision Support Technologies, Knowledge Management, WWW and Electronic Commerce applications development and technologies, and Information Systems Research Methods. The Hypertext Research Laboratory, whose aim is to facilitate the application of hypertext-based technology in academe, business and in the wider community, was founded by him in late 1989.