WHAT MEASURES DO WE NEED TO BUILD AN ELECTRONIC MONITORING TOOL FOR POSTGRADUATE TUTOR MARKED ASSIGNMENTS?

Emanuela Moreale, Denise Whitelock,
Yvonne Raw and Stuart Watt

What Measures do we Need to Build an Electronic Monitoring Tool for Postgraduate Tutor Marked Assignments?

Emanuela Moreale[1]
Denise Whitelock
Yvonne Raw[2]
Stuart Watt[3]

KMI[1]
The Open University
Walton Hall
Milton Keynes
MK7 6AA

IET[2]
The Open University

Robert-Gordon University[3]

E.Moreale@open.ac.uk

Abstract

This paper reports the findings from a preliminary study, which set out to test a number of metrics and heuristics that will form the basis of an electronic monitoring system for postgraduate level assignments. Such a (web-based) system would augment the University's electronic 'Tutor Marked Assignment' (e-TMA) infrastructure with new services. These services would support tutors, monitors, and students as they learn to grade and write essays and reports.

This study has illustrated that some readability measures can act as excellent diagnostics, such as picking up over-length essays, but that sentence length is not a good one because of problems concerning the complexity and style of the text. Other heuristics such as 'indicators of analytical thinking' formed promising measures that could be included in our system.

Overall this study has provided some benchmark metrics to begin to construct a monitoring system that will include both readability metrics and content heuristics from both tutors and students.

Introduction

Universities and other further education establishments are concerned not only with providing their students with first class tuition, but also that the most salient quality procedures are implemented and correctly documented. There are, of course, real problems with the scalability of such quality systems when student numbers are high. For example at the Open University, the student population for some courses can be as high as 15,000 students. Large numbers of tutors are needed to support these students and in any given year there is 10 percent intake of new tutors. These new recruits receive training and a random sample of their marking is monitored. Electronic tools are therefore needed to focus the monitoring process to where it can most support staff and improve quality. Using these tools monitoring can, in principle, be extended to every assignment on a course, improving the quality of monitoring by focussing it where it is most pertinent. Such a system would be able to give tutors some immediate feedback on their assessment, be more supportive and help them develop their skills in this area.

The Open University is currently undertaking a feasibility study to ascertain whether an essay-marking tool can be developed for such quality assurance purposes. This project does not plan to introduce automatic assessment of students work but to produce a system that actively supports the monitoring process. The software tool combines a number of metrics derived from an investigation of student assignments that will facilitate the construction of an electronic monitoring system. Such a tool will form part of a software suite (Goodfellow et al 2002) that are being developed at the Open University which utilise a latent semantic analysis approach to the diagnostic. This has been adopted because Latent Semantic analysis produces a useful representation for text research (Foltz et al 1998).

This paper reports on some of the findings from one of the preparatory stages of the project. This stage involved carrying out an analysis of the main contributory factors that are influencing the grade awarded to the tutor-marked assignments from the Open University course H801 – ‘Masters in Open and Distance Learning’.

The aim however of the current phase of the research is to investigate:

- The investment required to prepare the raw data for analysis.
- whether the factors suggested by current prototype essay grading systems can predict the grade for these Masters level assignments
- a set of surface (non-content related) metrics allowing a reasonably-reliable prediction of the grade level of each assignment

Background to the Course and Students Used for Test System

The electronically marked TMAs (Tutor Marked Assignments) of the 1998 cohort, for the MA module in Open & Distance Learning entitled 'Foundations of Open and Distance Education' were selected from the University’s archive. The total number was 194. They came from 61 students and consisted of four TMAs per student. The prescribed length was 2000 words for TMAs 1 and 2, and 4000 for TMAs 3 and 4. The marks from all these TMAs contributed to 50% of the student’s final grade for this module and hence were a substantial contribution to their final grade. The syllabus for this module covered the following topics:

- The theory and practice of open and distance learning

- Terms and rationales in open and distance education
- Becoming a critically reflective practitioner
- Theories of open and distance learning
- Characteristics and needs of learners
- Interaction in open and distance learning

The TMAs were designed to examine student understanding of all the topics in the course and required them to submit well argued and informed account of current theories and research into Open and Distance learning.

Students

The students who were from this cohort comprised of a number of educational professionals who were resident in the UK, Greece, Switzerland, Japan and the United States. All the tutoring took place online. There was no face-to-face interaction. A small number had already obtained PhDs and were currently working in Universities. These students wanted to understand more about distance and on-line learning as they were about to embark on devising such courses themselves, for the first time in their careers. Other participants were from a software design background.

Tutors

The three tutors for this presentation of the MA module had also written the course materials. They had all worked in the Open University for over 15 years and were familiar with the University's tutoring system.

Preparation

The essays were submitted as Microsoft Word documents. The essays were anonymised by removing all the students' and tutors' personal details that occurred within the first page on the 'PT3 form'. This was carried out in accordance with data protection guidelines for research of this nature. A new identifier was then allocated to each essay.

The removal of the PT3 form proved at times not to be sufficient, as some students had included their personal details within the format of their essays. There were also cases where tutor comments appeared within the body text of the essays.

It was also necessary to separate the essay proper (the argumentative part of each essay) from the references and appendices sections in order to ensure that appropriate sections of the text were analysed. A Perl script was written to carry out these tasks, although a manual check was required afterwards. The essays were then ready for statistical analysis. The level of complexity within these proceedings was heightened by the non-standard format of the essays, and therefore the need for a more restrictive formatting of student essays was identified.

Procedure

A number of researchers (Christie 1999, Williams 2001) find Page's (1994) dichotomy between content and style-based approaches to automated essay grading a useful approach. We propose a taxonomy that is better suited to marking postgraduate essays:

Taxonomy for Postgraduate Essay Marking/Monitoring		
Deep measures	Topic-related content	What is said: "does it match the question?" - "usage" / "coverage" (Christie 1999)
Surface measures	Non-specific content	What is said: All other content e.g. presence/ absence of analysis and synthesis
	Style	The way something is said

Table 1 Taxonomy for Postgraduate Essay Marking and Monitoring

This feasibility study concentrated on general "surface" (style and non-specific content) measures, while specific topic-related content-based approaches (such as LSA) were left for a later study. One aim of the project was to show that surface measures can be usefully employed in the automated monitoring of essay marking.

The measures employed in this study include essay length and over-length (style and content), readability statistics (style), number of references (non-specific content) and indicators of critical thinking and use of experience (non-specific content).

Results

We therefore started off by calculating some very general surface measures, such as essay length.

Essay Length

Previous studies had revealed a strong surface correlation between essay length and score and in particular, between the fourth root of the number of words and the score (Page, 1994).

Measurements used:

- normalised word count,
- log of word count and
- fourth root of word count (as in Page).

Results revealed weak but significant correlations for all these measures when calculated both over the whole set and the set of longer essays (assignments 3 and 4). However, no significant correlation was found for the shorter essays (assignments 1 and 2).

This finding appeared rather strange at first as the results did not appear to fit the expected pattern for the shorter essays. Did this mean that there was something wrong with this sample? This is a finding that we would want any monitoring system to detect. Therefore in order to explore further we decided to investigate essay 'over-length', which is something that the external examiners would not want to see occurring. If this was happening too frequently then the shorter essays would not match the pattern.

A summary of correlation results between essay length and Score is provided in Table 2.

Measurement	Correlation Coefficient	Number of Cases	Significance level
All Essays:			
- Normalised Word Count	r = 0.175	N = 128	p < 0.05
Assignments 1 & 2:			
- Word Count		Not significant	
- Log of Word Count			
- Fourth Root of Word Count			
Assignments 3 & 4:			
- Word Count	r = 0.272	N = 65	p < 0.05
- Log of Word Count	r = 0.274	N = 65	p < 0.05
- Fourth Root of Word Count	r = 0.277	N = 65	p < 0.05

Table 2 - Pearson Correlation – Word Counts and Score

Essay Over-length

The hypothesis was that essays that were remarkably over-length would be unfocused and would therefore score lower marks according to the course rubric. However, it was discovered that essays that were well over the word limit tended to have higher minimum scores.

This could have been down to the following concurrent reasons:

- Longer essays tend to be more complete than shorter (generally, but not necessarily, lower-scoring) essays.
- Tutors tended not to enforce the word limit (i.e. they did not mark down essays even when complaining about essay length in their comments).

Degree of Over-Length	Minimum Grade
Less than 20%	No Pattern
20 % or greater	50
30 % or greater	60

Table 3 - Degree of Over-Length and Minimum Grade Exploratory Analysis - Working Set of 128 Essays

Readability Metrics

Readability metrics are sometimes used in automated assessment on the assumption that more readable essays get higher marks. Commonly-used metrics are the Gunning Fog Index, the Flesch Reading Ease Score and the Flesch-Kincaid Grade Level Score. These indexes are based on simpler metrics, such as the average number of syllables per word.

The test revealed that in this case none of these readability indexes showed any correlation with scores in our sample.

The analysis revealed that in dealing with master-level essays, while readability scores had a role in ensuring that the texts were not too difficult for primary and secondary school pupils to read; it was reasonable to expect a certain lexical and syntactic complexity (as well as longer average sentence length) from writings by adult masters-level students.

Sentence Length

'Sentence length' is a simple metric that has been reported both to be a good measure of readability (Si & Callan 2001) and to be correlated with grade (Page).

The results revealed no such correlation in our sample.

Analysis revealed one possible indication, which was that readability was not significantly correlated with the grade at postgraduate level. On the other hand, this result could have been due to the nature of the essays that were being analysed. These essays required students to submit longer more complex answers, as opposed to traditional "one-piece" essays. They also required students to use bulleted lists and tables, and to discuss several references (such as URLs) in the text.

Difficult words

Actual word difficulty is usually not 'measured' directly in readability metrics that instead gauge word difficulty through word length or number of syllables (e.g. Gunning Fog Index). The assumption behind measuring difficult words is that they are best avoided in one's writing.

The following assumptions were made:

- that word rarity could also be used as an indicator of word difficulty.
- that some amount of "difficult words" could be expected in Masters-level essays.

Results revealed that the qualitative exploratory analysis of the essay sample allowed us to identify a set of words that tended to be present in high-scoring essays. These included: 'dystopian', 'oxymoron', 'dualistic', and 'typology'.

This finding was significant to the study because it revealed that while the presence of such words could not be used to predict a score, a low-scoring essay that used such word(s) may warrant further attention by the monitors.

References

The hypothesis was that the more references there were in an essay, the higher the grade it would achieve.

The assumption was that masters-level essays were expected to provide a critical analysis of seminal works and research papers, which should have been be referenced appropriately (and listed in a separate references/bibliography section).

The rationale here was that an essay containing a substantial number of references gave some indication that the student had read around their subject, and was therefore more likely to get a higher grade.

The results revealed that there was a good correlation between the number of references and the grade awarded.

The following measurements were used:

- a count of references in reference section of each essay
- a count of actual references to publications etc in essay text

The number of references listed by students in the references or bibliography section of their essays, were measured. Two different counts were used; the amount of 'proper' references (books, journals), and the total amount of references (inclusive of 'proper' references, and other references such as URLs).

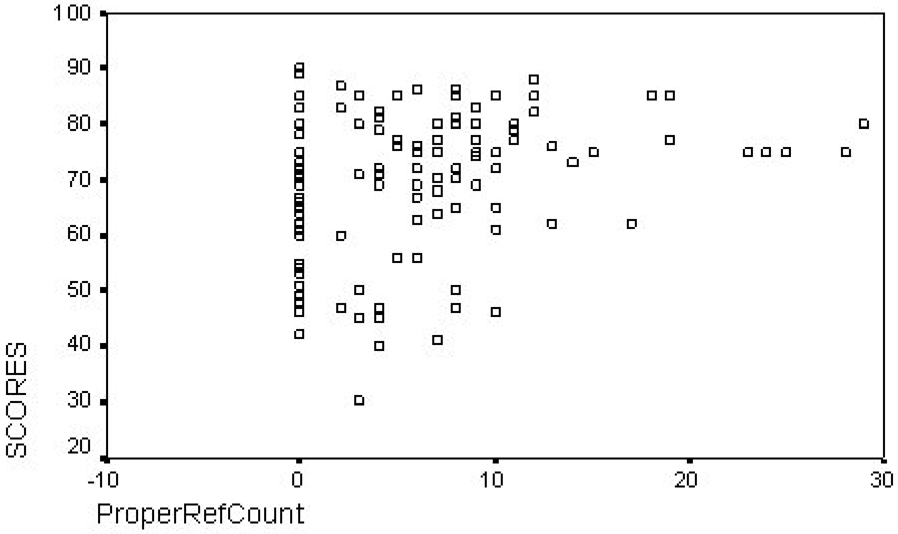


Figure 1- Scatter Diagram of Proper References Count and Score

Results revealed a correlation between the number of references and the score in both cases. The correlation was stronger in the case of 'all references' (r=0.270;N=125;p<0.01) than where only "proper" references are allowed (r=0.250;N=126;p<0.05).

Analysis revealed one possible explanation that said that, since the essays were from taken from a 'modern' course in distance teaching, the references were more likely to originate from online resources. As well as this explanation, it was also observed that students were encouraged to draw on their own experience, hence their references to the messages that were exchanged in online discussions.

References in Essay Text

A third count was also carried out, whereby the number of references to authors/works within an essay was calculated.

The rationale here was that a count of references in the references section would not always be reliable, as students sometimes simply compile a list of 'useful references' instead of listing only the works that were actually referenced in the text. Therefore, a count of references to authors/works within the essay text would be a more reliable indicator of the amount of analytical thinking that took place in the essay.

One problem emerged then, and that was that this kind of counting proved to be quite difficult, somewhat subjective and rather time-consuming. This count was therefore limited to a sample of 59 essays.

As expected, a highly significant correlation (r=0.467;N=59;p<0.01) was found between the number of references to authors/works in the essay text and the actual score. While this measure was applied to approximately half the essay set, and the counts were somewhat subjective, it still represented a useful measure that could be applied for monitoring purposes.

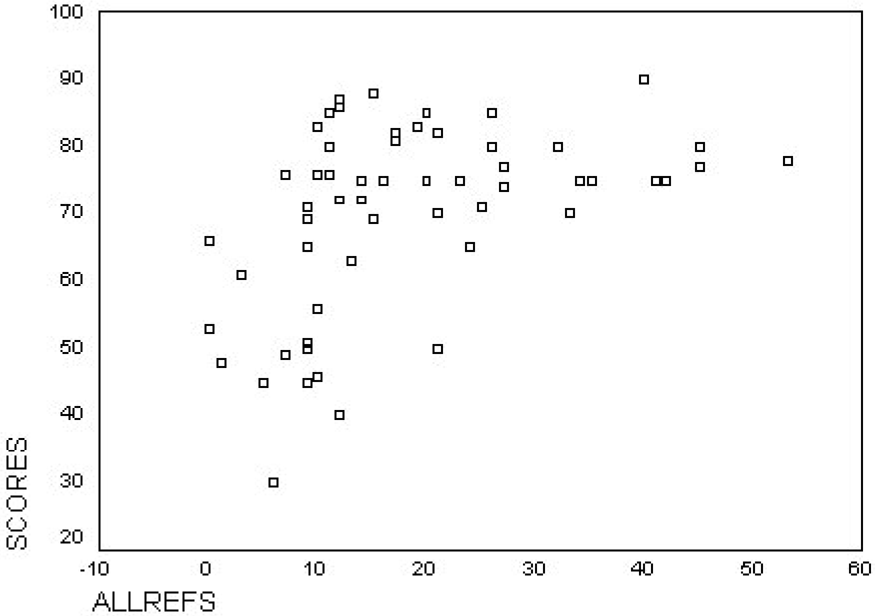


Figure 2 - Scatter Diagram of References Count Within Essay Text and Scores

Overall, reference counts seem to represent a useful measure for automated assessment monitoring purposes.

Analytical Thinking Indicators

The hypothesis was that essays containing greater numbers of words that indicated analytical thinking, (analytical thinking indicators), would score higher than essays with a lower incidence of such words.

The rationale was that analytical thinking is an important requirement in Masters-level essays, and therefore essay that contained more analysis should score better marks.

The list of words in Table 4 were analysed for correlations with score.

Results showed good correlations with score, for example correlations were found for ‘argue’, ‘explor(e)’, ‘model’, and ‘experience’. Some of the more interesting findings are also illustrated.

Analytical Indicator	Correlation	
argue	r = 0.337; N=128; p< 0.01	
explor(e)	r = 0.224; N=128; p< 0.05	
model	r = 0.202; N=128; p< 0.05	
experience	r = 0.180; N=127; p< 0.05	
defin (define, definition)	r = 0.173; N=128; p< 0.1	
impl(imply, implement)	r = 0.167; N=128; p< 0.1	
suggest	r = 0.147; N=128; p< 0.1	
theor(y)	r = 0.122; N=126; p< 0.2	
research	r = 0.118; N=127; p< 0.2	+
summaris(e)	r = 0.161; N=127; p< 0.1	+
assum (assume)	no significant correlation	–
assert	no significant correlation	
characteris(e, istic)	no significant correlation	
discuss	no significant correlation	
contrast	no significant correlation	
analysis / analyse	no significant correlation	
maintain	no significant correlation	
limitation	no significant correlation	
relate	no significant correlation	

NOTES + removing outlier increased correlation significance
– correlation was significant before removing outlier

Table 4 – "Analytical Thinking Indicators" (Stems) and Correlations with Grade

ARGUE:

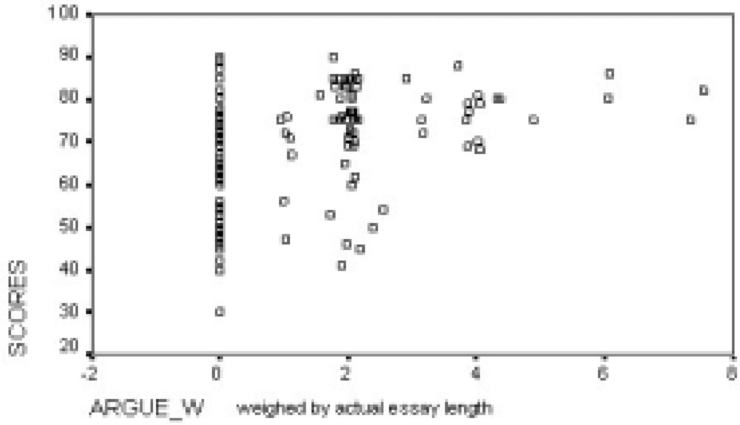
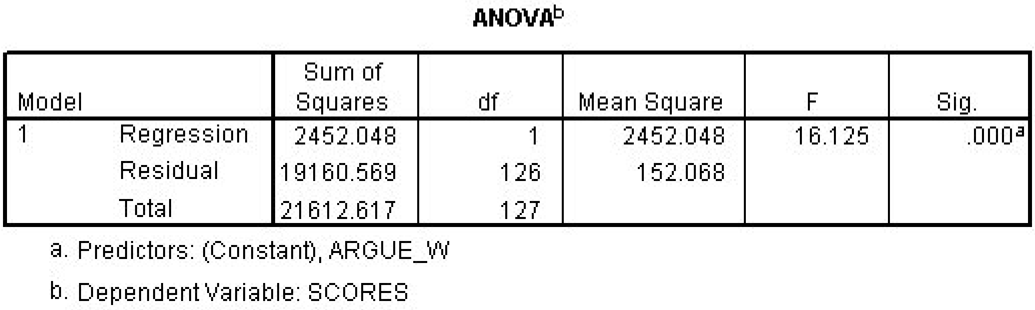


Figure 3 - Scatter Diagram - Occurrences of "Argue" and Score



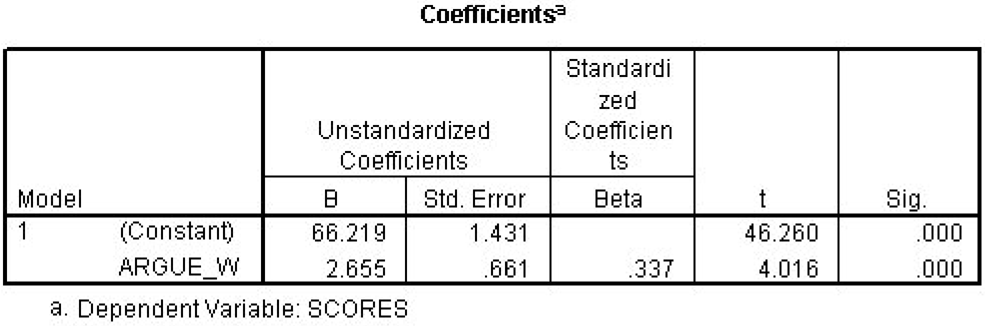
ANOVA^b

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2452.048	1	2452.048	16.125	.000^a
	Residual	19160.569	126	152.068		
	Total	21612.617	127			

a. Predictors: (Constant), ARGUE_W

b. Dependent Variable: SCORES

Table 5 ANOVA for "Argue"



Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	66.219	1.431		46.260	.000
	ARGUE_W	2.655	.661	.337	4.016	.000

a. Dependent Variable: SCORES

Table 6 - "Argue": Regression Coefficients

EXPLOR(E):

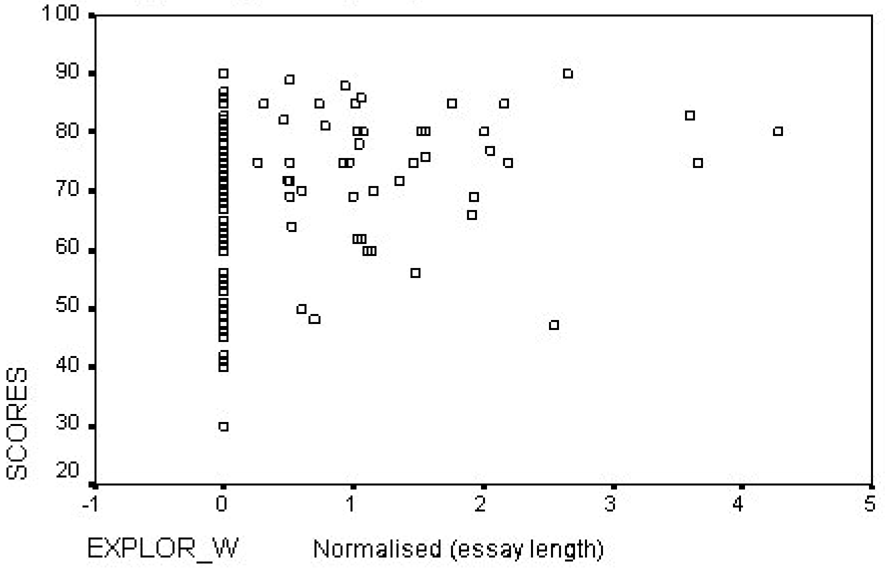


Figure 4 - Scatter Diagram - Occurrences of "Explor(e)" and Score

Analysis showed that whilst the use of some ‘critical thinking indicators’ instead of others is partly a matter of personal style, it would seem that the use of words such as ‘argue’, ‘explore’, ‘model’, ‘define’ and ‘imply’, are reasonable indicators of analytical thinking. Hence they are good score predictors when combined with other measures.

Experience

Another important element that tutors are looking for in essays at postgraduate level is ‘experience’ and the students’ use of their experiences in developing logical arguments in their essays.

Results indicated that with the word ‘experience’, there was a positive correlation with score (r=0.180; N=127; p<0.05). However, the expression ‘my (own) experience’ does not.

Analysis suggested that:

- the positive correlation of the word ‘experience’ with score (as shown in Table 7) was due to factors other than references to experience in the essay (e.g. where reference were made to ‘learning by doing’).
- A better measure must be found to gauge experience leveraging in essays.

Tutor Comments

A second type of analysis of the data explored whether there was a relationship between the number of tutor comments and grade awarded for each assignment. The exercise was undertaken because it was hypothesized that one of the heuristics that might be adopted for our monitoring system, could well be that the more comments that are awarded the lower the standard of pass.

A scattergraph of these findings as shown in Figure 5, illustrates a tendency towards this rule but no significant correlation. However, the aim was to identify a trend, and therefore the average number of comments per script for each standard of pass was analysed.

The dots that lie along the ‘y’ axis along the bottom of this graph depict the average number of comments for each of the Bales Categories for TMA number 5. The nature of TMA number 5 was such that it required the student to submit a dissertation proposal. This TMA was formative and therefore no mark was given.

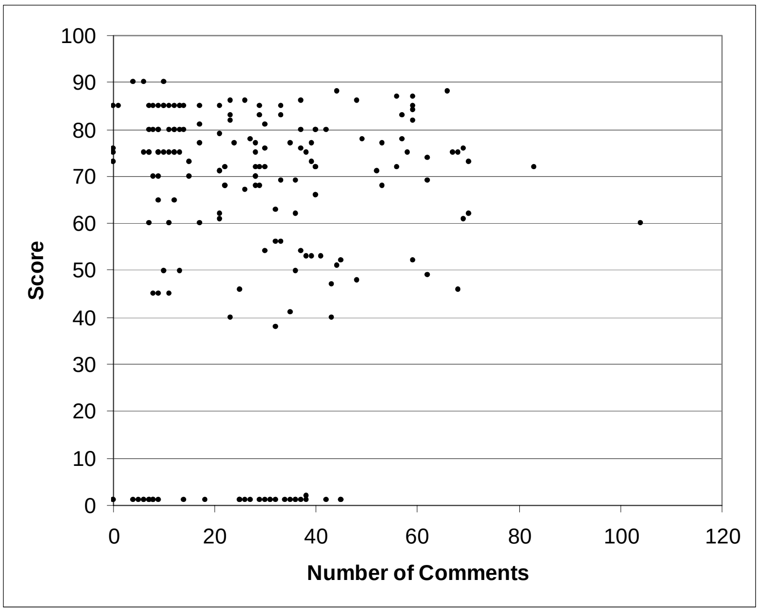


Figure 5 - Graph to illustrate the number of comments to each score

The ‘pass’ thresholds for the H801 module of the MA in Open and Distance Learning are as follows:
Pass 1 = 85 – 100
Pass 2 = 70 – 84
Pass 3 = 55 – 69
Pass 4 = 40 – 54

The following graph was created and shows that it is easier to identify trends by looking at the average number of comments per script for each standard of pass.

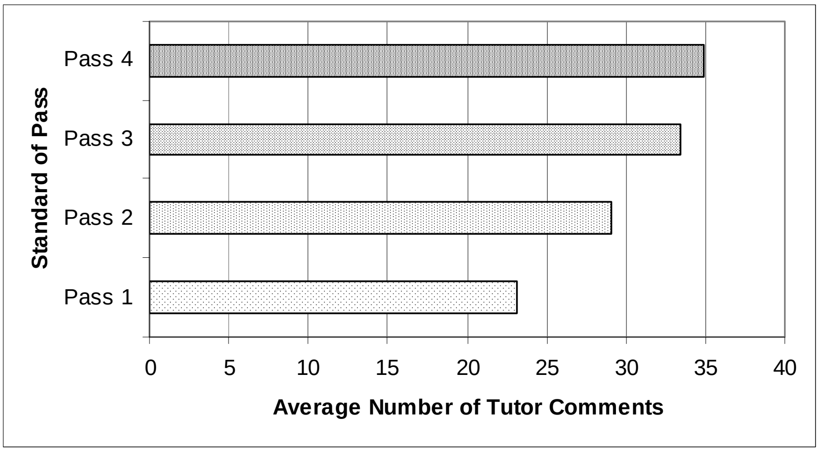


Figure 6 - Graph to show conflated scores against the average number of tutor comments

Figure 6 demonstrates that the hypothesis with respect to the correlation between the number of comments and the overall score, was correct. It can be seen that students who achieved a ‘Pass 1’ received less comments on average per script, than those students who achieved a ‘Pass 4’.

Conclusions

This study set out to investigate the investment required to prepare the raw data from the current electronic Tutor Marked Assignment (eTMA) system. The time required for this was considerable and taken up with programming script to carry out anonymisation, and to separate the references and appendices from the essay proper. The latter took weeks to write and performed correctly approximately 90% of the time. This was due to the variations in formatting and structuring in the essays, which in turn required post-processing and manual checking. In the absence of a standard format, the processing of the essay text highlighted potential issues for concern such as workload, margins for error, and scalability. Any future systems would require the standard formatting of student scripts in order to minimise pre-processing.

The second parameter that was investigated was whether the factors suggested by current essay grading systems could predict the grade for these masters level assignments. Our findings reflected the complexity of the postgraduate material and the different forms of presentation of that same material. This is because for example, the sentence length varied in these essays in a way not accounted for by the index measure. Bullet points were employed to summarise a position, and some of the questions required the student to produce bulleted lists of researchers positions in a given domain. It was the fact that these tried and tested measures were not behaving as we would have expected, that led us into a problem-solving route.

It was found that some of the standard metrics were not suitable for our sample, (such as sentence length), but that ‘essay over-length’ is a metric we should include as it proved to be a good diagnostic that revealed a problem with this sample of students. The occurrence of over-length assignments is indeed a finding that we would want any monitoring system to highlight.

The third parameter investigated in this study was whether a set of surface metrics such as the use of references would allow a reasonably reliable prediction of grade. In fact it was the number of authors referenced in the text rather than at the end of the essay that proved to be a highly significant predictor of grade. Further investigation into the content of the essays, revealed the occurrence of verbs such as ‘argue’, ‘model’, define’ and ‘imply’ which are reasonable indicators of analytical thinking and correlate with the human marker’s score.

The tutor comments on these essays also proved to be a valuable source of information. We have found that the number of comments here does vary with grade of pass, i.e. the highest grade receives fewer comments from the tutor. This will provide another metric to our monitoring system. This study has provided some benchmark metrics to start to construct a monitoring system, which will include both readability metrics and content heuristics from both tutors and students.

References

Christie, J.R. (1999) Automated Essay Marking – for both Style and Content, *CAA 1999*, Loughborough.

Flesch, R. (1979). *How to Write Plain English*, New York.

Foltz, P. W., Kintsch, W.,& Landauer, T. K. (1998). The measurement of textual Coherence with Latent Semantic Analysis. *Discourse Processes,* 25, 285-307.

Gunning, R. (1981). *The Technique of Clear Writing*. 2nd Edition, McGraw-Hill.

Page, EB, Lavoie, MJ, & Keith, TZ (1995). *Computer Grading of Essay Traits in Student Writing. Project Essay Grade* Extended Working Paper

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education*, 62(2), 127-42

Rudner, Lawrence & Phill Gagne (2001). An Overview of Three Approaches to Scoring Written Essays by Computer. *Practical Assessment, Research & Evaluation*, 7(26). Available Online:
http://ericae.net/pare/getvn.asp?v=7&n=26

Si, L. and Callan, J. (2001) A Statistical Model for Scientific Readability, *CIKM’01*, Nov. 2001, Atlanta, GA.

Whitelock, D., Raw, Y., Watt, S., and Moreale, E. (2002) *Improving Feedback to Tutors: Managing an Electronic Monitoring System.* Programme on the Learner Use of Media. Technical Report Series No.145, The Open University, UK

Williams, R. (2001) Automated Essay Grading: An Evaluation of Four Conceptual Models, *Proceedings of the Teaching and Learning Forum (TL Forum) 2001*.