A LOW TECHNOLOGY COMPUTER ASSISTED MARKING STRATEGY FOR LAW ESSAYS.

Kevin Bampton

A Low Technology Computer Assisted Marking Strategy for Law Essays.

Kevin Bampton,
The Derbyshire Business School,
University of Derby
Kedleston Road
Derby DE22 1GB

'Everything should be made as simple as possible, but not simpler'
(attributed to either Albert Einstein or *Readers' Digest,* depending on who you believe.)

Abstract

This paper provides an explanation of the rationale and underlying principles leading to the development and implementation of an automated student feedback tool for Public Law essays at the University of Derby. It also reports on the successes and limitations of the approach and suggests that the use of automated feedback at an early stage in formative assessment is a theoretically supportable and practically useful strategy.

The criteria of validity of formative assessment tools for essays

The potential of the essay as a vehicle for learning development has rather suffered from the demands of mass education. Kipka explains the problem:

"Students at university often receive mixed messages about essay writing. They are advised to go through a process of drafting and redrafting, yet (with the notable exception of postgraduates) they typically receive no formal credit or feedback on intermediate drafts, as if only the final product is to be valued. Given commonly experienced pressures and procrastinatory tendencies, it is hardly surprising that all too many university essays are submitted in a rushed, undigested state. Feedback along the lines of "take more time", "where's your argument?", "?", "so...", etc. is unlikely to get to the heart of the problem of mismatches between student and staff expectations."[1]

The aim of the project was to automate repetitive aspects of marking and giving feedback, so students could check their work before they handed it in

[1] Kipka, P (2001) Premises, principles, procedures, prudence: a useful taxonomy of learning objectives. In K. Chanock (ed.) *Sources of Confusion*, Proceedings of the National Language and Academic Skills Conference, La Trobe University).

and correct obvious errors. It was hoped that the program could provide interactive guidance on where they might improve their work, encouraging self-assessment and reflection before handing the work in. It was embarked upon because there seemed to be qualities particular to law that could be capitalised upon to achieve a relatively low-technology solution to the essay-marking problem.

Legal terminology has evolved as a efficient means of encoding and communicating legal knowledge. Law uses complex cross-referencing and complex shorthand to facilitate the storage and manipulation of information[2]. The use by law students of correct linguistic identifiers, including case law and statutes in the correct contexts is a key indicator of understanding and is essential for engaging in legal discourse.[3] Thus law lends itself to fairly simple parsing techniques, the domain-specific language resulting in a narrower range of acceptable words[4] to indicate comprehension and use of concepts.[5]

When dealing with relatively small groups it may be more efficient and more reliable for an expert to encode linguistic constructs into a program, rather than "teaching" a system to mark, for example by using latent semantics analysis. The volume of essays needed to "teach" as system and the ever-changing context of law would undermine the value of some approaches to automated essay marking.[6] Limiting the semantic scope for acceptable answer to legal questions reinforces the fact that students seldom benefit from deviating from the conventions of normal legal discourse[7]. Indeed there are practical advantages to using a domain specific language approach to[8] as has been demonstrated in the context of police training[9].

---

[2] see for example, Bruce Yandle and Andrew P. Morris, "The Technologies of Property Rights: Choice Among Alternative Solutions to the Tragedies of the Commons," 28 Ecology L. Q. 123, 127-130 (2001); Summers, Robert S. 1982. *Instrumentalism and American Legal Theory.* Ithaca: Cornell University Press.

[3] Birnbaum, L. (1991). Rigor mortis: A response to Nilsson's "Logic and artificial intelligence." *Artificial Intelligence, 47*, 57-77, p. 65

[4] Burstein, Jill C. and Randy M. Kaplan. (1995). On the Application of Context to Natural Language Processing Applied to the Analysis of Test Responses. In Proceedings from the Workshop on Context in Natural Language Processing, IJCAI, Montreal, Canada.

[5] Conrad, J. Guo, X. Jackson, P. Meziou, M. Database Selection Using Physical and Acquired Logical Collection in a Massive Database Selection Using Domain-specific Operational Environment, , Research & Development: Thomson Legal & Regulatory – West Group.

[6] Gerstl, P. (1991). A Model for the Interaction of Lexical and Non-Lexical Knowledge in the Determination of Word Meaning. In J. Pustejovsky and S. Bergler (Eds), Lexical Semantics and Knowledge Representation, Springer-Verlag, New York, NY.

[7] Cruse, D.A. (1986). Lexical Semantics. Cambridge University Press, Cambridge, UK cited in Kaplan, Randy M. and Randy E. Bennett. (1994). Using the Free-Response Scoring Tool To Automatically Score the Formulating-Hypothesis Item. (RR-94-08). Princeton, NJ: Educational Testing Service.

[8] Gerstl, P. (1991). A Model for the Interaction of Lexical and Non-Lexical Knowledge in the Determination of Word Meaning. In J. Pustejovsky and S. Bergler (Eds), Lexical Semantics and Knowledge Representation, Springer-Verlag, New York, NY.

[9] Burstein, J., Wolff, S., & Lu, C. (1999). Using Lexical semantic techniques to classify free-responses. In N. Ide & J. Veronis (Eds.), *The depth and breadth of semantic lexicons.* New York: Kluwer Academic Press.

Implementation

a) Objectives of the project

Building on general criteria for validity for automated feedback, the following objectives were decided upon:

- Underlying assumptions should be explicitly stated as part of the feedback so that students could challenge them, if necessary[10];
- Students should be guided by feedback to make improvements and to reflect on their essay[11];
- The emphasis and language of feedback should represent the sort of feedback that a human marker would give, but should be as detailed and comprehensive as required. It should be structured to be positive or constructive in line with best practice[12];
- Students should be able to act on the feedback and get a reaction (including feedback) to any improvement[13];
- The facility should be voluntary and in the pilot study, students should continue to benefit from the existing support of human marking until the facility had shown it was equal to the task.

b) Integration with teaching and learning strategy

Using the program to assist with formative assessment was only one aspect of the teaching and learning strategy. The tutorial problems to be assessed were set in advance and were designed to deal comprehensively with the issues to be covered classes. Students were directed to undertake written preparation of work in essay form and told about the 'essay check' facility. It was presented as a means of undertaking initial checking on their work before the tutorial and before they submitted the work for marking by the tutor. It was assumed that because the program was only capable of assessing superficial elements it might only serve to encourage 'shallow' or 'surface' learning[14].

---

[10] Carless, D. (2003) Putting the learning into assessment. The Teacher Trainer, 17(3), 14-18.

[11] MacLellan, E. (2001). Assessment for learning: The differing perceptions of tutors and students. Assessment and Evaluation in Higher Education, 26(4), 307-318. Note also "…the parent or inventor of an art is not always the best judge of the utility or inutility of his own inventions to the users of it." Plato *Phaedrus* 14, 275b.

[12] P. Ramsden, *Learning to Teach in Higher Education*, 1992, Routledge, London, 193

[13] Black, P., & Wiliam, D. (1998). Assessment and classroom learning, Assessment in education, 5(1), 7-74 .

[14] Gibbs, G. (1992) Improving the Quality of Student Learning through Course Design. R. Barnett (Ed). *Learning to Effect.* London: Open University Press.

c) Outcomes

It was hoped that students would check their work and correct common, annoying traits before submitting it. It was further anticipated that rapid, but imperfect feedback might be better than good but later responses from the tutor[15]. The elements that preliminary checking aimed to identify included (in no particular order):

- Correct spelling and usage of case names and terminology;
- Conventions of legal citation;
- Core concepts relevant to the question;
- Relevant authority including core and peripheral cases and statutes;
- Referring to leading and dissenting judgments within cases;
- Issues raised by the question;
- Beneficiaries of advice in problem scenarios.

d) Structure

The model developed to analyse student responses was adapted from Mosenthal and Kirsch's model of cognitive analysis of document literacy tasks[16]. This model was designed into the procedure for parsing of the student essay. The program searches hierarchically for indicators of (a) organizing categories, (b) specific categories derived from organizing concepts or from other specific features, and (c) the semantic features. Each is represented by text elements (strings or words) comprising metonyms, synecdoches and autonomasia, grouped within broad concept groups. The structure of these groupings reflects Bergler's layered lexicon approach[17]. Words are listed in their base forms together with a limited range of encoded suffixes to provide for the desirable flexibility within the domain and to compensate for the anticipated word ambiguity problems[18]. Some elements are also procedurally related to specific roles within argumentation, based on Toulmin's model[19] of reasoning. Although Anderson and Twining's model held considerable attraction, the Toulmin model was the only one that seemed replicable within the limits of the system.

---

[15] This seems to be the case with respect to peer feedback: Gibbs, G. & Simpson, C. (2003). Does your assessment support your students' learning? http://cehep.open.ac.uk/cehep/ssrg/reports/index.htm

[16] Mosenthal, P.B., & Kirsch, I.S. (1991). Toward an explanatory model of document literacy. *Discourse Processes*, *14*(2), 147-180.

[17] (n70)

[18] See, for example the approach taken by Burnstein *et al* (1995), p.8 (n65).

[19] Toulmin, S. (1958) *The Uses of Argument* (Cambridge University Press).

Specific item feedback response is triggered by a single instance of a matching text element, but only if matches had been registered all the way up the hierarchy of categorization. For example, to trigger feedback relating to a given semantic feature, the student would need to also have registered a match within the concept group of the specific category within which it is nested, but also have identified the relevant concept within the organizing category. Additional feedback is triggered by the frequency with which student answers match concepts within a group, following a simple Item Response Theory model[20].

Feedback was constructed on a modular basis. The model was conditioned by ideas derived from Marcu's Rhetorical Structure Theory approach[21]. This allows for fairly dynamic feedback with a realistic feel, while avoiding repetitious or overlapping strings. Depth of rhetorical effect is achieved by encoding strings that relate to the heirarchical position of the corresponding text element registered. Rhetorical effect is further achieved by modelling statements on the basis of nucleus statements (triggered by organizing categories, for example) and satellite statements (triggered by specific categorizations, for example)[22]. Generalised feedback based upon Item Response Theory is grouped towards the end of the feedback statement. The outcome matches the complexity and depth of the essay.

While scoring was not a significant objective of the exercise, it became obvious that for credibility, the indicative grading system, especially increments would need to accord with human marking. No attempt were made to develop an expert rule-based system, having reviewed the findings of Clauser *et al*[23] which indicated that weighted linear regression represented the best strategy to predict human scoring within expert domains. Instead the presence or absence of a text element was simply counted and used in a regression-based scoring model calibrated against the judgement of the human markers. This was in line with the approach reported by Clauser *et al.*

A significant feature of the programme is the approach to using feedback as a means of assisting students to restructure their answers. Specific feedback constitutes not only confirmatory statements, but further guiding questions encouraging students to map the problem more closely to the encoded model. Responses also assume the possibility of misattribution of meaning to the incidence of text elements, and therefore explicitly state what the meaning

---

[20] Baker, Frank (2001). *The Basics of Item Response Theory.* ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, College Park, MD, available online: http://edres.org/irt/baker/.

[21] Echihabi, A and Marcu, D (2003). A Noisy-Channel Approach to Question Answering. Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL), July 7-12, Sapporo, Japan.

[22] Mann, William C., Christian M. I. M. Matthiessen and Sandra A. Thompson (1992). Rhetorical Structure Theory and Text Analysis. Discourse Description: Diverse linguistic analyses of a fund-raising text_. ed. by W. C. Mann and S. A. Thompson. Amsterdam, John Benjamins : 39-78.

[23] Clauser, B., Margolis, M. Clyman, S. and Ross, L. (1997) Development of automated scoring algorithms for complex performance assessments. Journal of Educational Measurement, 34, 141-161

that has been assumed might be. By these two methods the feedback is designed to assist students in constructing models of understanding by situating the data within the model. The program encourages the student to reconstruct his or her essay in a progressively more detailed way, while forcing students to examine the essay in detail and several times over. It was intended that by critically revisiting the same material they would become more habituated into basic approaches to legal discourse and problem solving will become second nature to them. The approach follows the cognitive approach outlined by Lesh & Lamon[24]. At the same time the process aims at providing the schemata that form the building blocks of legal discourse step by step in an interactive way, as if teaching basic comprehension[25].

Implementation

The program was always designed to be lightweight using minimal computer resources, no specialist software or plug-ins, and operable entirely by client-side scripting. Each essay question was therefore programmed as a self-standing module, built entirely of elements of javascript for interactive functions, Perl elements were used as the basis for parsing and html formed the shell. The sophistication of the program was put into its conceptualization and design, rather than into the software itself. As a formative self-assessment tool it required none of the usual security protection, based on the simple principle that students would have nothing to gain by cheating themselves. The test of the success of the project was to be measured on the basis of comparison of the checker's performance with blind marking; the extent of student use; improvement in student performance.

The core trial involved encoding four Public Law tutorials, based on questions that had been successfully used in previous years and for which criteria had been developed. The essays ranged from a discursive essay within a broad political context to a stipulative problem question. In addition, a tutorial using the same technology, but consisting of a series of short-answers was coded as well as a free-text test for second years evaluating a stipulative data set derived from client interviews. These controls were created in order verify several of hypotheses and underlying assumptions about free-text assessment.

Aside from initial research, programming the first checker involved around 30 hours and each subsequent question took around 10 hours to code. Typically the checker is programmed to identify around 100 different concept groups, distributed across organizing categories, specialized categories, more

[24] Lesh, R. & Lamon, S. (1992) *Assessment of authentic performance in school mathematics* Washington, DC: American Association for the Advancement of Science.
[25] Rumelhart, D.A. (1980). Schemata: The building blocks of cognition. In R.Spiro, B. Bruce, & W. Brewer (Eds.), *Theoretical issues in reading comprehension* (pp. 33-58). Hillsdale, NJ: Erlbaum.

specialized categories. A free floating group of ancillary concepts, representing additional features of good answers that students might include as a result of reading outside the recommended range of sources was also programmed.

Results

The program was evaluated using a framework more commonly associated with the evaluation of continuing professional education[26] because of the skills and problem-based dimension of some of the assessments. Outcomes were measured based on their effectiveness in the following categories:

- Realising the design objectives
- Student participation
- Learner satisfaction
- Acquisition of knowledge, skills and attitudes as evidenced in examination and coursework performance
- Continued application of learning and second order effects

Realising the design objectives

Apart from occasional bugs, related to the scoring computation and the problem that cookies, essential to convenient repeated use of the program, were not enabled on University Learning Centre computers, the program operated without any technical problems. Students had no difficulty operating it, given that is only one button for them to press to gain feedback, indicative grade and refresh. The file size averages 50KB for each self-standing essay checker, giving it a swift loading time, the parsing process for a 2,500 word answer is faster than the refresh time of the page.

Student participation

Student participation was critical objective[27] since the primary purpose of developing the program was to provide them with speedy feedback. This and the general perceptions of the users were measured by questionnaire, by means of verbal reports and through entries in their course diaries. The questionnaire was issued at the end of the module examination. The return rate was higher than the general module questionnaire, with 84 out of 108 students completing it. One student reported frustration at not being able to get the program to work on her home computer, and a further student

[26] Cervero, R (1988) *Effective continuing education for professionals* (San Francisco: Jossey-Bass)

[27] Note also "…the parent or inventor of an art is not always the best judge of the utility or inutility of his own inventions to the users of it." PPlato *Phaedrus 14, 275b*

reported that he had not used the facility because he did not word-process his work. The tutorials were run over two consecutive years for cohorts of around 80 students per year. In all the method was used to evaluate around 1650 tutorial pieces, each student using each checker on average around 4 times. All students thought that the checker was useful, or very useful and that they encouraged them to improve their work.

Students indicated that they read through most of the feedback carefully and made significant changes to their work before resubmitting their work to the checker, rather than making superficial modifications to try and enhance their score. This was confirmed from superficial analysis of the feedback from the "mailto" feature built into two of the checkers, which allowed student use to be remotely monitored. Qualitative feedback suggested that the calibration of grading in the first checker should have been set a little lower so as not to demoralise students early on in the course. Students soon requested checkers to be available for every tutorial, rather than simply every other tutorial, which was taken to be an indicator of some success. Several students reported that it significantly assisted them in starting off their essays or in structuring them. One second-year non-standard entrant who had struggled through the first year with essays expressed the view that it had helped to develop her question answering abilities. Insofar as it is possible to draw inferences from changes in performance, there was a significant effect on student work[28].

Learner satisfaction

Of those who used the checker, all but one expressed the view that the program was worth the inconvenience of having to word-process work and deal with the problems of the University's network system. Every user rated the feedback as being useful, with approximately 65% expressing the view that the feedback was very helpful indeed. 76 students expressed the view that they had learned a significant amount of substantive information from the checkers. All of the students indicated that the initial feedback had prompted them to do further research or reading. Almost all students indicated that they would want to use the facility again. Four students commented that the checker occasionally did not recognise material when they felt they had already included it. Suggestions for improvement of the system centred upon extending its availability to other modules and increasing the sensitivity of the grading system. 56% of the 2003-2004 cohort, although still free to avail themselves of a human marker as well as the checker, relied entirely upon the checker for formative feedback.

[28] For consideration of the validity of drawing such inferences, see Mislevy, R (1996) Evidence and Inference in Educational Assessment, cSE Technical Report 414.

Acquisition of knowledge, skills and attitudes as evidenced in examination and coursework performance

It is notoriously difficult to draw conclusions, year in year in relation to performances of different cohorts of students. However, the examination performances in Public Law 1 over the two years that the checkers have been in operation have seen a statistically significant increase in the grouping of grades in the middle range, rather than the lower pass range. More significantly, the phenomenon of the "empty answer", with little of substance or substantive content has almost disappeared. Coursework quality has improved along similar lines, as one might expect.

Blind marking of computer assessed work yielded little substantial difference in grading and no differences of more than one grade step (e.g. C+ to C), for lower and mid-range grades. A significant divergence was apparent in the C+ and above range, where grades are very dependent on qualitative features. The human markers tended to recognize attributes in the work that the program would be incapable of detecting, such as persuasiveness, cogency and aspects of subtlety and sophistication. There tended, however, to be a convergence of grading when it came to first class work, which had technical features of precision and included material that could be encoded into the program. Modifications to the algorithm for calculating the grade were made for the second year to compensate for these disparities based purely on statistical adjustments to model predictive grading. The disparities disappeared as a result of this remodelling.

Generally material was presented with better referencing, more authority and fewer basic mistakes. Summative coursework feedback still included comments by the markers on mundane, but important aspects of the work, but much of the need for this had been removed, so the focus was on strategy and improvement of style, structure and emphasis. The improved quality of the work also improved the speed of marking turnaround. In 2003-2004, the full portfolio of work for 103 students (8 pieces each) was processed within fourteen days, partly on account of the work being generally well-presented and the students having acted on feedback where it was available. Of the two cohorts, the 2002-2003 cohort were generally perceived to be a weak year (which corresponds with their average entry qualifications) in relation to their performance in other subjects, where students would be expected to perform as well, of not better, than in Public Law. The 2003-2004 group are acknowledged to be a stronger year (as evidenced again by the entry qualifications), and their performance was still marginally stronger in Public Law than in other first year subjects.

Continued application of learning and second order effects

As yet, it has not been possible to devise a systematic form of long term follow up on the more general impact of the intervention. Some improvement in the work that is still being perceived in coursework submitted for the Human Rights second year module may be attributable to further encouragement to use the first year writing guide. The impression gained in the successor module Public Law 2 is the basic objectives have been met.

Conclusions

The checker program is still at a very early stage of development and as this paper demonstrates has not been subjected to testing on a particularly rigorous basis. The aims of the project were relatively modest. However the principle of adopting considerable shortcuts in order to attain these seems sustainable. The focus on coherent and useful feedback seems one that has been a worthwhile one. Much feedback presented to students is after they can make practical use of it within the context of a learning activity. The possibility of getting instantaneous, if not always perfectly targeted, feedback that can be acted upon and then instantly reassessed is one that has been an impossibility using traditional educational means. The instantaneous and detailed nature of the feedback generated by the model discussed in this paper seems to actively encourage students to check their own work and to read further, while maintaining a healthy suspicion of the computer's capacity to make qualitative judgements about essays. The private nature of the interaction seems to allow students the opportunity to try and then try again and anecdotal evidence suggests that students experiencing fundamental problems have been more, rather than less willing to approach tutors for help.

The initiative has helped make the process of preparing tutorials in written form a more interactive one. One of the advantages of the checker system is that the substance of the dialogue between marker and marked is moved to the time of the preparation of work, when students are most receptive to feedback and guidance. Students so often make the same mistakes regardless of how often they have been warned about them in general terms. The facility to use relatively low-level programming to help them spot and correct the "obvious mistake" using prompts and questions to get the student to re-examine their work is one that has potential even within the sophisticated world of free text assessment.

As a supplement to the human marker, the computer marker can remove from the process the demoralizing grind of repetitively identifying common, sometimes trivial, but often irritating and habit-forming mistakes. By signalling, during the formation of answers where students might further develop their work it provides an opportunity to independently develop and to experiment with strategies, simply not available with once and for all human marking (as it invariably must be).

The inevitable compromises involved with such a low technology approach may open this program up to criticism if essay-marking systems are judged purely on the accuracy of scoring as feedback. However, if the judgement is based on the extent to which students are stimulated into checking their work and improving and reflecting upon it, then even a low technology system such as this one has a contribution to make to educational technology.