Analysing tutor feedback to students: first steps towards constructing an electronic monitoring system

Denise Whitelock,* Stuart Watt,** Yvonne Raw* and Emanuela Moreale*
*The Open University, **Robert Gordon University
email: d.m.whitelock@open.ac.uk

Virtual Learning Environments provide the possibility of offering additional support to tutors, monitors and students in writing and grading essays and reports. They enable monitors to focus on the assignments that need most attention. This paper reports the findings from phase one of a feasibility study to assist the monitoring of student essays. It analyses tutor comments from electronically marked assignments and investigates how they match the mark awarded to each essay by the tutor. This involved carrying out a category analysis of the tutors' feedback to the students using Bales's 'interactional categories' as a theoretical basis. The advantage of this category system is that it distinguishes between task-orientated contributions, and the 'socio-emotive' element used by tutors to maintain student motivation. This reveals both how the tutor makes recommendations to improve the assignment content, and how they provide emotional support to students. Bales's analysis was presented to a group of tutors who felt an electronic feedback system based on this model would help them to get the right balance of responses to their students. These findings provide a modest start to designing a model of feedback for tutors of distance education students. Future work will entail refining these categories and testing this model with a larger sample from a different subject domain.

Introduction

The digital university is becoming more fact than fiction with the adoption of Virtual Learning Environments (VLEs), with perceived pedagogical and administrative advantages (Hazemi and Hailes, 2002). Students are encouraged to submit their course work electronically with tutors commenting electronically on the scripts. This enables the feedback process to be speeded up. This was found to be the case with the Open University's bespoke electronic Tutor Marked Assignment (TMA) system. VLEs offer

additional support to tutors, monitors and students for writing and grading essays and reports. This enables monitors to focus on the assignments that most need their attention by providing tutors with high-quality feedback.

There are a number of options that can be taken to monitor the marking of students' electronic assignments. One is to develop an automatic essay grading system to analyse the text of a TMA and award an appropriate mark. This mark can then be checked against the one given by the tutor. Such an automated system enables monitors to focus on assignments that most needed their attention. Tutors can then be given high-quality feedback to assist them in developing their marking skills.

The basis for this approach comes from the development of automatic essay-grading systems in the United States since the mid-1960s. Although these offered fairly good correlations with human graders, they relied on superficial features which are easy to extract such as length, average word length, use of punctuation and use of certain key words. This is still largely the case, although the features used are often far more computationally intensive than they used to be (see Hearst, 2000).

Two essay grading systems are worth describing in particular, e-rater (ETS) and the Intelligent Essay Assessor (IEA). E-rater (ETS) developed about 100 different linguistic features, and then used regression analysis to develop a scoring model that compared well with human graders. ETS developed this into a system that scored essays with a 90 per cent correlation with expert assessors. The Intelligent Essay Assessor uses latent semantic analysis (LSA), which is a statistical technique, designed to estimate how similar the content of one body of text is to another, at a semantic level rather than at the word level. In estimating this similarity, it corresponds remarkably well to human scorers. Using LSA for essay grading involves indexing pre-graded essays, with associated feedback, so that new essays can be graded by finding the best matches among the pre-graded ones. The strength of IEA is that it can be used to give constructive feedback. Finally, it is considered psychologically sound as it is a theory of language and matching several significant psycholinguistic effects, for example, Landauer, Foltz and Laham (1997).

The FRAMES project set out to explore how techniques such as these could be used to assist monitoring as well as student and tutor support, in the essay assessment process. Initially, we began with a set of readability and text scoring measures that were claimed would account for the grade awarded to an undergraduate essay (Burstein, Marcu, Andreyev and Chodorov, 2001). These non-content metrics were used to build a predictive model from a training set of scripts using LSA. This model allows a previously unseen script to be analysed and awarded a grade. We have found that the measures suggested by the literature were not sufficient to construct an adequate model for master's level essays and that other indicators of student understanding needed to be included in order to produce a workable model (see Moreale, Whitelock, Raw and Watt, 2002).

A second option to monitor the marking of students' assignments is to focus on tutors' comments that have been inserted into the students' essays. The comments found on the TMAs form a rich data set from which to extract some generic findings with respect to comments and the mark awarded. This raised the question of how these trends could be identified and translated into rules for an electronic monitoring system. One approach is to construct an analytical framework of the types of interactions that occur between tutor

and student in the tutor comments; then to count these interactions and see if a trend emerges according to the mark that was awarded to the assignments in question. The simplicity of such an analysis is deceptive since the categories under investigation must be well operationalized and the classification schema must be consistently adhered to as opposed to being dependent on the views of the individual observer. It would, therefore, be better to use a tried and tested system of interactional categories and see if it would fit our context rather than to construct one from scratch.

One system we considered was Flanders's (1970) set of interactional categories. These were designed to record what goes on in classrooms. Flanders's scheme uses ten separate categories to record teacher and pupil interactions. Three of the categories that refer to pupil talk are irrelevant to the TMA context. In our case the pupil 'talk' is the essay itself. The other Flanders's categories are concerned with teacher talk in response to the student-initiated responses, such as accepts ideas or feelings and praises, while the remaining categories are concerned with teacher-initiated events which are directed at controlling classroom behaviour. These include 'criticizes behaviour' which is expected to change through the teacher interaction. 'Gives direction or orders' is another controlling category. These latter categories did not fit into the distance tutoring model for this master's level course and so this system was rejected.

All distance students require feedback from their tutor not only about the subject matter but also an acknowledgement of their effort and progress. Hence an explicit level of socio-emotive support as well as direct instruction is necessary and is emphasized in the training of OU tutors. In fact the types of interactions that go on in face-to-face situations are encouraged in the remarks of both the online and paper-based tutor. A set of categories devised by Bales (1950) appeared to be more suitable since it distinguishes between task-orientated contributions and 'socio-emotive' interjections. However, it was devised to analyse face-to-face interactions and not text dialogues. We also considered a system developed by Angeli, Valanides and Bonk (2003), who distilled a set of eleven categories to help tutors provide online assistance but again the socio-emotive role is not explicit. The analysis reported here adopted Bales's framework and set out to classify the tutor comments typed on to the essay. This is the particular feedback addressing issues as they appear in the students' written text. The study aimed to:

- investigate whether a Bales interactional analysis of the tutor comments could provide an adequate model of the tutors' written feedback on the student assignments;
- identify trends in these interactions that accompany the grade awarded to the assignment;
- translate these trends into a set of heuristics which will form the basis of an automated assessment tool which will be used by the examinations office to select TMAs for monitoring purposes.

Procedure

The electronically marked TMAs chosen for analysis were taken from the MA module in Open and Distance Learning entitled 'Foundations of Open and Distance Education'. The total number was 194 selected from 42 students. The students were required to submit five

TMAs before submitting their final dissertation; however, the fifth TMA involved developing a proposal for a subsequent dissertation, and was therefore omitted from this study. Some students did not submit all their TMAs. The marks from the four TMAs contributed to 50 per cent of the student's final grade. The students were seeking substantial feedback and guidance from their tutors' comments in their assignments in order to improve their marks during the course and this type of feedback was considered to be the most personally helpful to them throughout the duration of the course.

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 a well argued and informed account of current theories and research into open and distance learning.

Students

This cohort of students consisted of international educational professionals, from Greece, Switzerland, Japan and the United States. There were no face-to-face tutorials; tutoring took place online. A small number had already obtained Ph.D.s and were currently working in universities but wanted to understand more about distance and online learning as they were about to embark on devising such courses themselves. Others had a software-design background. All were committed, conscientious participants, although they had not studied for a number of years and were unused to being cast in a student role.

Tutors

The three tutors for this presentation of the MA module also wrote the course materials and were experienced researchers in the field. They had tutored at least three other OU courses and were adept users of the electronic TMA system. Tutors' comments should therefore not only illustrate a sound knowledge of the domain, but also provide exemplars of positive constructive advice to students on how to improve their TMA score – a facet of tutoring that can be detected by Bales's categories. The first tutor marked seventy scripts, the second sixty-three scripts, and third sixty-one scripts. The small difference in number is due to initial difference in tutorial group numbers and some students not submitting all their TMAs. It is also worth noting that the number of comments made by the tutors increased in the later assignments, as shown in Figure 1.

Using Bales's interaction analysis

Bales' twelve interactional categories are shown in Table 1. They were designed to record what was being achieved during group interaction sessions. The strength of this system according to Sapsford (1999) is that it is a subtle, rich and sophisticated measuring

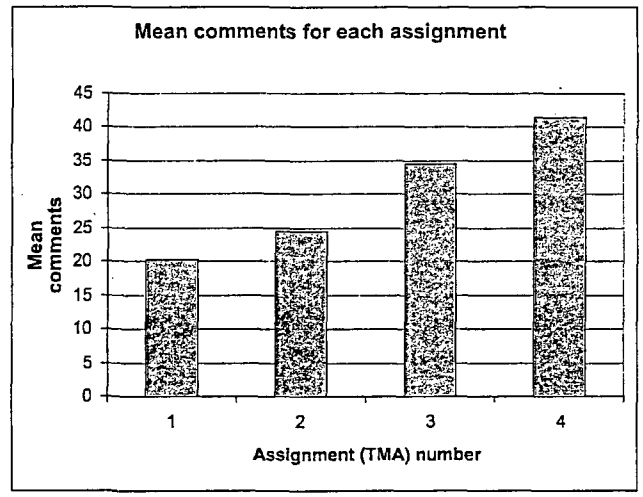

Figure 1: Changes in mean comments through the assignments

instrument that was designed to distinguish between task-orientated and socio-emotive contributions in a group session. Open University training for tutors stresses the importance of praise and constructive guidance and these categories had the potential to capture this type of tutor feedback.

Categories		Specific examples
	Positive reactions	
A1	1. Shows solidarity	Jokes, gives help, rewards others
A2	2. Shows tension release	Laughs, shows satisfaction
A3	3. Shows agreement	Understands, concurs, complies, passively accepts
	Attempted answers	
B1	4. Gives suggestion	Directs, proposes, controls
B2	5. Gives opinion	Evaluates, analyses, expresses feelings or wishes
B3	6. Gives information	Orients, repeats, clarifies, confirms
	Questions	
C1	7. Asks for information	Requests orientation, repetition, confirmation, clarification
C2	8. Asks for opinion	Requests evaluation, analysis, expression of.feeling or wishes
C3	9. Asks for suggestion	Requests directions, proposals
	Negative reactions	
D1	10. Shows disagreement	Passively rejects, resorts to formality, withholds help
D2	11. Shows tension	Asks for help, withdraws
D3	**12. Shows antagonism**	**Deflates others, defends or asserts self**

Table 1: Bales's interaction categories

Categories		Specific examples
	Positive reactions	
A1	1. Shows solidarity	Jokes: Rewards e.g. 'excellent', 'good point', 'well done' etc.
A2	2. Shows tension release	'Yes now this is good
A3	3. Shows agreement	'Agrees: 'yes', 'I agree', 'I can accept ...' 'That's right' etc.
	Attempted answers	
B1	4. Gives suggestion	Directs: 'notice that ...', 'please explain further', 'could expand this ...', 'discuss further', 'enlarge upon' etc.
B2	5. Gives opinion	Evaluates: 'I'm wondering ...', 'my reaction is ...', 'I assume that you are ...', 'I think', 'in my opinion', etc.
B3	6. Gives information	Orients: e.g. 'As you know quite a few authors are ...', 'Your examples show the potential for ...', 'in fact much new technology ...', etc.
	Questions	
C1	7. Asks for information	Questions: '?' 'Is this specific to ...?' 'Why is this different..?' 'What would this be?' etc.
C2	8. Asks for opinion	e.g. 'so you're saying ...?', 'is the content and pedagogy here likely to be ...?'
C3	9. Asks for suggestion	Would the research design look like this or ...?
	Negative reactions	
D1	10. Shows disagreement	Disagreement: 'No', 'I don't think I agree', 'I'm afraid that isn't the point' ... etc.'
D2	11. Shows tension	and again this does not follow'
D3	12. Shows antagonism	No examples found

Table 2: Examples of incidences of Bales's interaction process

Each tutor comment was coded with respect to Bales's categories (see Table 1). The code was marked up on the assignment next to the comment and later entered onto an electronic spreadsheet. Two researchers undertook this task and the inter-rater reliability factor was 0.89. All the comments on the TMA were coded with the Bales system, although one category, D3 ('shows hostility') was redundant, which is encouraging as it indicated that there was no evidence of tutors displaying antagonism towards the students. It was also helpful to discover that no extra categories needed to be added to account for all the tutor comments.

The categories which contained minimal comments with a mean value of less than two were A2 ('Shows tension release'), C3 ('Asks for suggestion') and D2 ('Shows tension'). The low values for A2 and D2 categories complement each other revealing a very low incidence of tension. C3 is also low because this is not a record of a face-to-face interaction and so the tutor rarely asks for a further suggestion. Bales's interactional categories appear to provide an appropriate analytical framework for these tutor comments despite being designed to monitor face-to-face interaction. This issue will be discussed further following

the more detailed analysis of the comments illustrated in Section 4 below. Instances of Bales's interactions can be seen in Table 2.

Results

The main objective of this phase of the analysis was to identify a set of trends in the tutor interactions that matched the grade awarded. In order to account for the variation in student background and the even bigger difference in tutoring style (one tutor wrote a third more comments than the other two), the mean number of comments per category was calculated for each level of pass awarded. These pass levels were given to students in their assignment guide and were as follows:

Pass 1 = 85–100
Pass 2 = 70–84
Pass 3 = 55–69
Pass 4 = 40–54
Bare Fail = 30–39
Fail (with the option of resitting) = 15–29
Fail outright = 0–14

The practicalities of the analysis then meant that the number of incidences of each of Bales's categories for each standard of pass was counted as can be seen in Figure 3. The categories were then conflated so that A category comments (positive reactions), B category comments (direct teaching comments), C category comments (questions), and D category comments (negative reactions) were grouped and counted for each standard of pass (see Figure 2). This was done to see whether there was a notable differential in the number of incidences of any specific categories within each standard of pass.

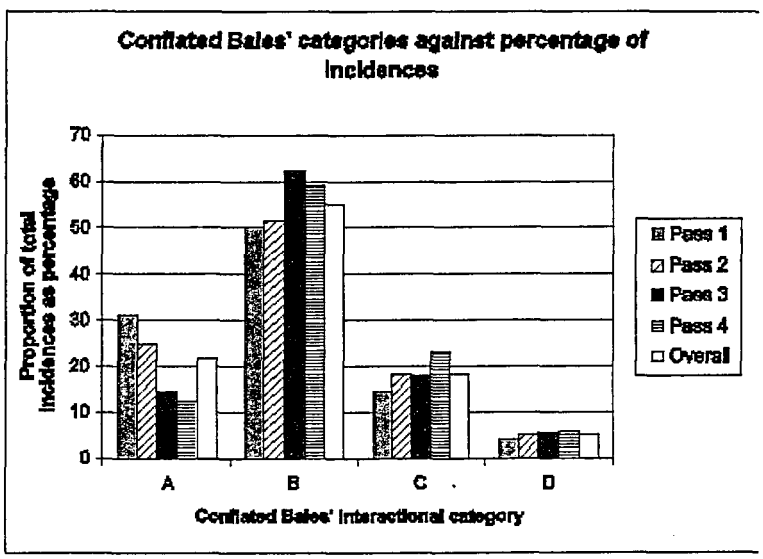

Figure 2: Graph to show conflated Bales' categories against proportion of incidences

The general picture that emerges when the categories were conflated is that over half of the tutor comments give suggestions, directions and opinions about the work (category 'B'). The rest of the comments are spread between the following two groups – questions (categories 'C') and providing socio-emotive support (category A). The former category was used by tutors to illustrate things that were incorrect but in a non-confrontational manner. The comments for categories 'B' and 'C' point out difficulties and problems with the assignment and offer constructive help to sort these out. The remaining comments (one sixteenth) demonstrate direct tutor disagreement with the student.

A more generic picture emerges which offers a basic teaching model of the tutors' comments but more importantly the data illustrate how this model is dynamic and shifts to match the competency of the student. This issue is explored by examining the variations that are found within the pass levels. For example, with the higher passes, group B, that is, tutor direction, forms the bulk of the tutor comments. These students receive more praise and are questioned less about their presentation. They are not asked to reflect upon so many problems with the text as they are clearly not there. The converse is true for the lower passes where category 'B' (such as direct teaching comments) still form the bulk of the tutor responses but there is more questioning and less praise.

These findings suggest that trends exist between the types of tutor comments per pass level, but can these be translated into a set of heuristics for our monitoring system? In order to address this question the full set of comments was scrutinized with respect to pass level as illustrated by Figure 2.

A pass at Level 1 reveals that the tutor shows broad agreement with the student, using such phrases as 'I agree' and 'I can accept', together with 'that's right' type comments. The tutor offers more opinions about details in the assignment opening up more of a dialogue with the student. There is less direction given with a smaller amount of questioning the student for information. The same general pattern emerges with a Level 2 pass but there are now equal amounts of 'B1' and 'B2' comments. This means the tutor is not emphasizing opinion generation as much as giving direct suggestions for improvement with these students. There is also more use of questioning to draw attention to problems in the assignment. Both for a Level 3 and 4 pass the pattern changes. The tutor is giving far more direction ('B1') and asking more questions that highlight inconsistencies and draw attention to problems in the text. There are more disagreement comments too, although these are small in number. These trends indicate that students who exhibit well integrated arguments obtain a higher grade. Critical argument is acknowledged by tutors explicitly with comments such as 'I agree', etc. The teaching model lent itself to the formation of a number of metrics. These were:

- more of the comments will be in the 'B' category (that is, directive interactions such as 'gives suggestion', 'gives opinion', or 'gives information') compared with the other categories (Wilcoxon signed ranks test, B compared with A, $z = -8.34$; $p<0.001$; B compared with C, $z = -9.36$, $p<0.001$; B compared with D, $z = -10.37$, $p<0.001$);
- the number of times comments from the 'D' category occur (such as 'shows disagreement', 'shows tension', or 'shows antagonism') is always less than the number of incidences in the other categories (Wilcoxon signed ranks test, D compared with A, $z = -8.66$, $p<0001$; D compared with B, $z = -10.37$, $p<0.001$; D compared with C, $z = -9.03$, $p<0.001$);

Categories		
A	**Positive reactions**	
	Conflated	*An overall significant positive correlation with score (r_s=0.196, p<0.05)*
A1	1. Shows solidarity	Significant positive correlation with both score (r_s=0.179, p<0.05) and a similar trend at pass level, overall and for two tutors
A2	2. Shows tension release	No correlations or trends identified.
A3	3. Shows agreement	Trend to a positive correlation with score; two tutors showed significant correlations with both score and pass level
B	**Attempted answers**	
	Conflated	*A strong overall significant negative correlation with both score (r_s=–0.236, p<0.005) and pass level (r_s=–0.266, p~0.001)*
B1	4. Gives suggestion	Significant negative correlation with both score (r_s=–0.427, p<0.001) and pass level (r_s=–0.415, p<0.001) identified both overall and for each tutor.
B2	5. Gives opinion	One tutor showed a positive correlation with both score and pass level, and a second showed a similar trend.
B3	6. Gives information	No correlations or trends identified.
C	**Questions**	
	Conflated	*A strong overall significant negative correlation with both score (r_s=–0.314, p<0.001) and pass level (r_s=–0.317, p<0.001).*
C1	7. Asks for information	Significant negative correlation with both score (r_s=–0.333, p<0.001) and pass level (r_s=–0.327, p<0.001), overall and for two tutors.
C2	8. Asks for opinion	No correlations or trends identified.
C3	9. Asks for suggestion	No correlations or trends identified.
D	**Negative reactions**	
	Conflated	*A trend towards negative correlation with both score and pass level*
D1	10. Shows disagreement	Significant negative correlation with both score (r_s=–0.178, p<0.05) and pass level (r_s=–0.170, p<0.05) overall, not generally significant for individual tutors.
D2	11. Shows tension	No correlations or trends identified.
D3	12. Shows antagonism	No examples found

Table 3: Main correlations for Bales's categories, individual and conflated

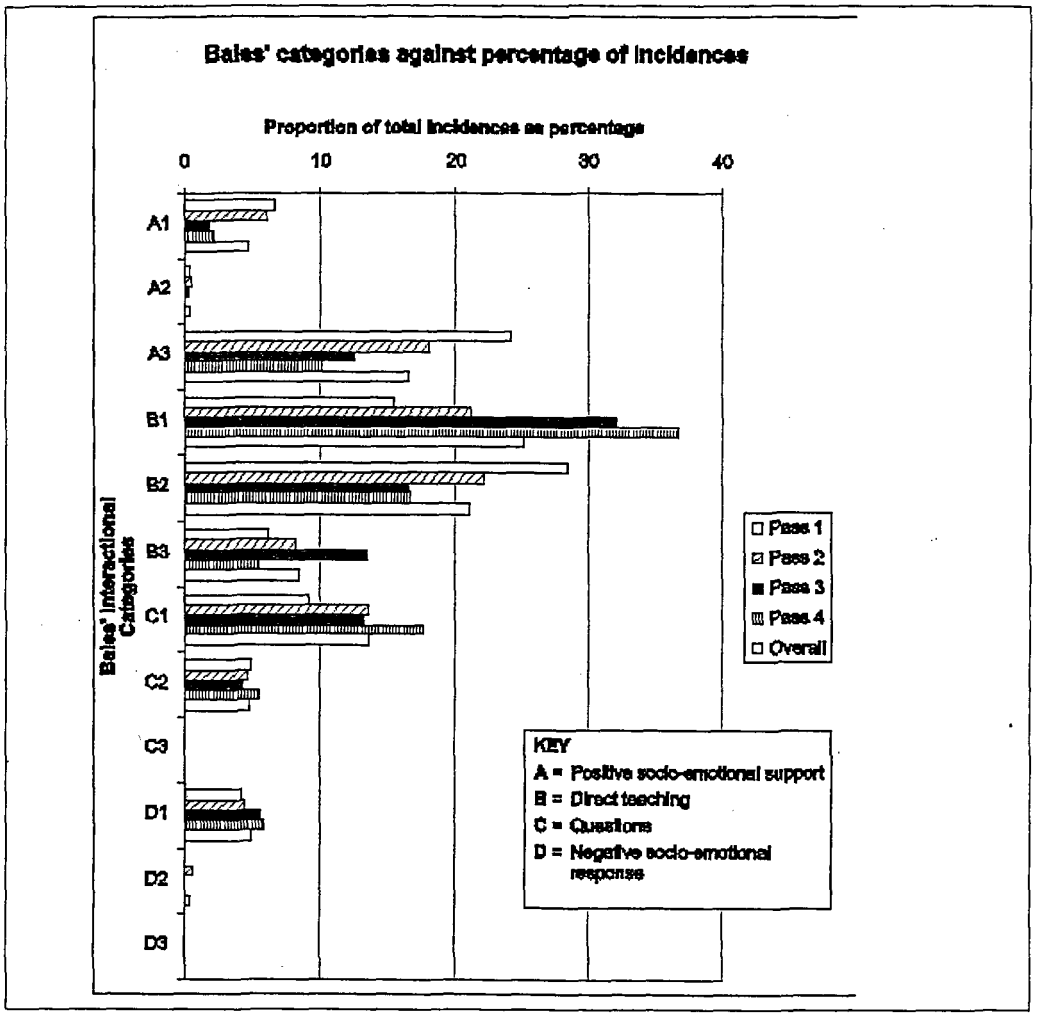

Figure 3: Graph to show Bales's categories against proportion of incidences

- there is an inverse trend in the number of incidences of categories 'A' and 'C' corresponding with the standard of pass. For example, a 'Pass 1' will receive more comments within category 'A' than a 'Pass 4' script, and a 'Pass 4' script will receive more comments within category 'C' than a 'Pass 1' script. Category 'B' is also negatively linked to the strength of the assessment. In effect, positive reactions are given to stronger scripts, where attempted answers and questioning are used to address weakness.

A full set of correlations were used to analyse the relationships between grades and comments, summarized in Table 3 (Spearman's rank correlation coefficient, N=145). Overall, conflated category 'A' showed a statistically significant positive correlation with score, and conflated categories 'B' and 'C' showed significant negative correlations with score.

Perhaps most interestingly, while all kinds of comments in category 'A' are positively linked with pass level, in category 'B' the picture is more complex, within its overall negative link with pass level. B1 'Gives suggestion' is a strongly significant negative correlation, with more suggestions being given to weaker assignments. However, B2 'Gives opinion' shows a weaker positive correlation: for one tutor it is strongly significant (r_s=0.525, p<0.001), a second tutor shows a similar trend and the third tutor no effect at all. So, while there seem to be important individual differences in teaching style, opinions do seem to be used more in response to stronger assignments. This behaviour both needs and deserves further exploration.

Given the significance of these effects regarding the balance of comments in these categories, it seems probable that there could be a sound basis for generating a rule-based system to analyse these comments. However, before this can be achieved we need to see if the Bales categories can account for comments in other subject domains such as science, technology and business studies. These analyses are currently in progress.

Conclusions

The Bales category system accounted for all the tutor comments found on the assignments. This suggests that it is a useful model for analysing tutor comments but needs testing in other subject domains. It provided a general tutoring model which showed dynamic variation with level of pass rate. To summarize, a pattern emerges that could form the basis for some expectations about how assignments should be marked. For example, for the best students obtaining the top grades there would be more praise given. Less direct teaching comments would be needed but there still should be some questioning. This would then stimulate the student to reflect upon their answers and to improve in subsequent assignments. There would be few negative comments found. The balance of comments should change as the mark awarded decreases. The students with the lowest marks need more direct teaching and so the number of 'B', that is, teaching comments, should increase. However, some praise should be given where it is due so as to encourage and motivate the student to complete their studies.

The advantage of Bales's system is that it distinguishes between task-orientated contributions, and the 'socio-emotive' element used by tutors to maintain student motivation. Our analysis has detected that the tutor not only used questions to stimulate further reflection but also employs this category to point out constructively where there are problems with an essay.

The findings to date have been presented at tutor workshops and tutors have been enthusiastic about developing an electronic system based on this model because it makes explicit the training they have already received from the University. They felt feedback from such a system would help them to achieve the right balance in their responses to students. One tutor remarked that he now realized he did not give enough explicit praise to his best students as he believed a good mark said it all!

The findings presented to date provide a modest start to the design of a model of feedback for tutors of distance-education students. However, to achieve the primary goal of providing automated support for monitoring, and supporting both students and tutors will require further research in two areas. First, the Bales approach needs to be validated in

other disciplines, work that is currently in progress. Second, we need to build a semantic rule-based system to allot tutor comments to the correct categories, and validate it against a substantial corpus of analysed assignments. However, the work has other potential: for example, it could help to analyse tutor comments in computer-based conferences and could provide a form of quality assurance for these teaching approaches. Properly validated, the resulting system will have the potential to make a significant difference to quality assurance and learning support in virtual learning environments.

References

Angeli, C., Valanides, N. and Bonk, C. J. (2003), 'Communication in a web-based conferencing system: the quality of computer-mediated interactions', *British Journal of Educational Technology*, 34, 1, 31–43.

Bales, R. F. (1950), 'A set of categories for the analysis of small group interaction', *American Sociological Review*, 15, 257–63.

Burstein, J., Marcu, D., Andreyev, S. and Chodorow, M. (2001), *Towards Automatic Classification of Discourse Elements in Essays*, Annual Meeting of the Association for Computational Linguistics,Toulouse, France.

Flanders, N. (1970), *Analyzing Teacher Behaviour*, Reading, MA: Addison-Wesley.

Hazemi, R. and Hailes, S. (eds) (2002), *The Digital University: Building a Learning Community*, London: Springer.

Hearst, M. (2000), 'The debate on automated essay grading, IEEE intelligent systems', September–October, *http://www.knowledge-technologies.com/presskit/KAT_IEEEdebate.pdf*

Landauer, T. K., Foltz, P. W. and Laham, D. (1997), *Introduction to Latent Semantic Analysis, Discourse Processes*, 25, 259–84. *http://lsa.colorado.edu/papers/dpl.LSAintro.pdf*

Moreale, E., Whitelock, D., Raw, Y. and Watt, S. (2002), 'What measures do we need to build an electronic monitoring tool for post graduate tutor marked assignments?' Sixth International Computer Assisted Assessment Conference, 9–10 July, Loughborough, pp. 253–67.

Sapsford, R. J. (1999), *Survey Research*, London: Sage.