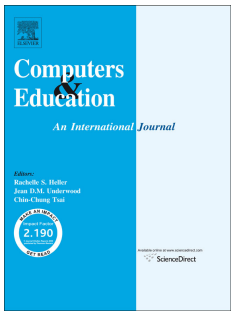
Accepted Manuscript



Title: A comparison of human and computer marking of short free-text student responses

Authors: Philip G. Butcher, Sally E. Jordan

PII: S0360-1315(10)00046-1

DOI: 10.1016/j.compedu.2010.02.012

Reference: CAE 1579

To appear in: *Computers & Education*

Received Date: 15 June 2009

Revised Date: 11 February 2010

Accepted Date: 12 February 2010

Please cite this article as: Butcher, P.G., Jordan, S.E. A comparison of human and computer marking of short free-text student responses, Computers & Education (2010), doi: 10.1016/j.compedu.2010.02.012

A comparison of human and computer marking of short free-text student responses

Sally E. Jordan[a,c]

[a]Centre for Open Learning of Mathematics, Science, Computing and Technology, The Open University, Walton Hall, Milton Keynes, MK7 6AA,United Kingdom.

b Corresponding author, p.g.butcher@open.ac.uk, tel. 0044 1908 653730, fax. 0044 1908 655787

c s.e.jordan@open.ac.uk

Abstract
The computer marking of short-answer free-text responses of around a sentence in length has been found to be at least as good as that of six human markers. The marking accuracy of three separate computerised systems has been compared, one system (Intelligent Assessment Technologies FreeText Author) is based on computational linguistics whilst two (Regular Expressions and OpenMark) are based on the algorithmic manipulation of keywords. In all three cases, the development of high-quality response matching has been achieved by the use of real student responses to developmental versions of the questions and FreeText Author and OpenMark have been found to produce marking of broadly similar accuracy. Reasons for lack of accuracy in human marking and in each of the computer systems are discussed.

Keywords
authoring tools and methods

Abbreviations
iCMA interactive computer marked assignment

Figure captions
Figure 1 Increasing feedback received on three attempts at a short-answer question embedded within OpenMark. (This is Question D in the analyses described in Sections 2 and 3.)

1. Introduction
Many purpose built computer-assisted assessment (CAA) systems and general purpose virtual learning environments include facilities for matching short-answer free-text responses. However, questions that use matching of this type are typically quite limited in their aims, frequently restricted to matching a word or two, with little account being taken of word order, negation, synonyms or spelling. Questions requiring longer or more complex responses are thus frequently framed as selected response items (e.g. multiple choice). Multiple choice items are generally regarded as robust and reliable, but concern has been expressed that they may not always be assessing what the teacher believes that they are, partly because they require 'the recognition of the answer rather than the construction of a response' (Nicol, 2007, p.54). The cognitive processes required of students are very different when they are asked to construct a response of their own, in their own language, and without any prompts from the question (Mitchell, Russell, Broomhead & Aldridge, 2002). Short-answer constructed response items are highly valued in traditional paper-based assessment and recent developments have seen the introduction of more sophisticated response matching into CAA systems, enabling the automatic marking of longer free-text answers. These response

matching systems can be classified into two distinct groups: those that use some form of computational linguistics and those based on the algorithmic manipulation of keywords.

Perhaps the most well known of the systems available for the e-assessment of free-text are e-rater (Attali & Burstein, 2006), an automatic essay scoring system that gives good agreement with human grading when the focus is more on writing style than on content, and Intelligent Essay Assessor (Pearson, 2009) which claims to mark reliably for style and for content. Both these systems are targeted at marking essays. However, the systems described in the current paper are designed for short factual answers, and although students are advised to write their answers in the form of a sentence of no more than 20 words, the focus of the grading is on the content of what they write rather than the style of their writing. Intelligent Assessment Technologies (IAT) FreeText Author system, the use of which at the UK Open University (OU) is described in detail in Jordan & Mitchell (2009), draws on the natural language processing (NLP) techniques of information extraction. The IAT software was selected for use at the OU because it provides an authoring tool that can be used by a question author with no knowledge of natural language processing. Two other systems based on computational linguistics are C-rater (Leacock & Chodorow, 2003) and the system developed by Sukkarieh, Pulman & Raikes (2003, 2004); these are reviewed in Siddiqi & Harrison (2008), who go on to appeal for a systematic evaluation of the different technologies.

At the OU, a bank of IAT short-answer questions has been developed, and student responses to developmental versions of the questions have been used to improve the answer matching. The questions are delivered to Open University students via the OpenMark assessment system (Butcher, 2006), which was developed at the OU but is now open source. OpenMark allows multiple attempts at each question, with an increasing amount of teaching feedback provided after each attempt (Figure 1). In this way, students (usually adults studying at a distance) are encouraged to act immediately on the feedback provided which, wherever possible, is targeted to the student's misunderstandings and so simulates a tutor at the student's elbow (Ross, Jordan & Butcher, 2006).

Figure 1 Increasing feedback received on three attempts at a short-answer question embedded within OpenMark. (This is Question D in the analyses described in Sections 2 and 3.)

OpenMark provides a range of question types, allowing for the free-text entry of words, numbers, scientific units and simple algebraic expressions as well as drag-and-drop, hotspot, multiple-choice and multiple-response questions. Until recently, the use of OpenMark's own free-text response matching for short-answer questions was limited to those questions where very short answers (usually no more than a word or two) were expected and where synonyms and alternative spellings were unlikely or disallowed. However, OpenMark response matching for longer answers has now been developed, again on the basis of real student responses, and the accuracy of the marking of 'unseen' batches of responses has been investigated.

This paper compares the accuracy of marking of six course tutors with that of the IAT software (Section 2) and then, for student responses to the same short-answer free-text questions, compares the accuracy of marking of the IAT software with that of OpenMark's own response matching and Regular Expressions as implemented in Java (Section 3). OpenMark's response matching and Regular Expressions are both based on the algorithmic manipulation of keywords. The work has been carried out as part of a broader practitioner-led initiative (Butcher, Swithenby & Jordan, 2009), investigating the use of interactive computer

marked assessment and funded by the Centre for Open Learning of Mathematics, Science, Computing and Technology at the OU.

2. A comparison of the marking of the IAT software and human markers

2.1 Methodology

Questions written using the IAT authoring tool were presented to students on the Open University's introductory science module S103 *Discovering Science* in a series of eight purely formative OpenMark interactive computer marked assignments (iCMAs). S103's final three presentations (each 9 months in length) started in October 2006, February 2007 and October 2007. The answer matching had been developed as far as possible before the questions were released to students but, as expected, students answered the questions in some unexpected ways, and the answer matching was developed in the light of the responses received during each presentation.

The author of the majority of the questions (the second-named author of this paper) has no expertise in computational linguistics or computer programming, but is an experienced academic author of OU course material and assessment tasks of all types, including conventional e-assessment questions. After an initial training stage of a few weeks, she was able to write short-answer e-assessment questions and appropriate IAT answer matching with relative ease. The time spent in the initial writing of a question and its answer matching varied between a few minutes and several hours, depending on the complexity of the question and the range of answers expected. Amending the answer matching in the light of student responses was even more dependent on the complexity of the question, taking more than a day for some questions. The inclusion of targeted feedback, and the response matching required to trigger this, added to the time required for development.

Seven short-answer questions (listed in Appendix A) were used as the basis of a human-computer marking comparison, with student responses from the February 2007 presentation being marked by six course tutors as well as by the IAT software. The IAT answer matching had been developed during the October 2006 presentation, but care was taken not to alter the answer matching further while the human-computer marking comparison was in progress. Three of the seven questions (questions D, F and G) were slightly rewritten for the February 2007 presentation, itself an illustration of the fact that very often questions can be improved by rewording the question as well as by refining the answer matching. The changes were considered to be insufficiently major to prevent the questions from being included in the human-computer marking comparison.

The tutors who took part in the human-computer marking comparison were volunteers, deliberately not selected to be particularly accurate or inaccurate markers of tutor-marked assignments. They had widely varying experience of OU work (having been employed by the University for between 2 and 30 years) and had academic backgrounds spanning the four main scientific disciplines taught in S103. The tutors were provided with marking guidelines, designed to be as similar as possible to those provided to ensure consistency of marking in tutor-marked assignments. The tutors were asked to mark all the responses to the short-answer questions as either correct or incorrect i.e. to give each response a score of 1 or 0. They were invited to comment on the marking of individual responses when they thought that this would be helpful and some did so extensively.

Note that the basic scores assigned by the computer were also simply 1 or 0. To ensure that the human-computer marking comparison did not assume that either the computer or the human markers were 'right', both the computer's and each tutor's marking of each response to each question were compared against:

- The median of all the tutors' marks for that response (used here as a measure of majority view)
- The mark awarded by IAT
- The mark awarded by the author of the questions (taken to be definite in cases of disagreement). The question author's marking was done 'blind', without knowledge of the way in which the course tutors had marked the question or the way in which the IAT system had responded to each particular response. However the author was very familiar with the mark schemes and model answers that the IAT system was applying.

The total of each marker's scores for each question was determined and chi-squared tests were conducted to test the null hypothesis that the marking of each of the markers (including the IAT system) was indistinguishable.

Responses in which there was any divergence between the human markers and/or the computer system were inspected in more detail, to investigate the reasons for the disagreement. When the investigation was complete, the tutors were contacted about a few responses to ascertain why they had marked them in the way that they did. They were also asked whether they felt that the marking guidelines provided had been adequate.

The Kappa, κ, inter-rater statistic was also used to assess the degree to which each marker agreed with the question author. It is calculated from the formula $\kappa = (P(a) - P(e))/(1-P(e))$ where $P(a)$ is the proportion of times the markers agree, and $P(e)$ is the proportion of times the markers are expected to agree by chance alone. For each question, the markers were ranked according to their inter-rater agreement with the question author. Analysis of variance was then used to ascertain whether there were any significant differences between the overall marking of each of the markers (again including the IAT system).

A second part of the analysis compared the overall computer grading out of three (depending on how a student modified their answer at second and third attempt, in the light of feedback received) with more conventional human marking out of three (with partial credit for partially correct responses). This analysis is not reported in detail here, other than for the information it provided about responses that the different human markers believed to be partially correct.

2.2 Results

Chi-squared tests showed that, for four of the seven questions (questions A, C, D and E as listed in Appendix A), the marking of all the markers (including IAT) was indistinguishable at the 1% level (Jordan & Mitchell, 2009). For the other three questions, the markers were marking in a way that was significantly different. However for all the questions the mean mark allocated by the computer system was within the range of means allocated by the human markers. In some cases the differences between human markers were large.

The percentage of responses where there was *any* variation in marking ranged between 4.2% (for Question A, which could be satisfied by a tightly constrained response) and 64.4% (for Question G, a more open-ended question). Table 1 details the sources of the variation in grading for Question A and illustrates a finding that was consistent across all questions, namely that although most of the course tutors were in good agreement most of the time with

their colleagues (i.e. with the median course tutor mark), and with the question author and the IAT system, the major source of variation was discrepancies in the grading of the course tutors. The Kappa inter-rater statistics measuring agreement with the question author are also given.

	Marker	**Number of responses that were marked differently (out of a total of 189):**		
		From median of course tutors	From IAT	From question author (κ inter-rater agreement)
Course tutors	Marker 1	5	6	5 (0.92)
	Marker 2	1	2	1 (0.98)
	Marker 3	3	4	3 (0.95)
	Marker 4	0	1	0 (1.00)
	Marker 5	2	3	2 (0.97)
	Marker 6	1	2	1 (0.98)
Computer marking	IAT	1	-	1 (0.98)

Table 1 Sources of the variation of grading in the 8 responses to Question A that were not marked identically.

Table 2 compares the marking of the course tutors and the IAT system with that of the question author, for all seven questions.

One course tutor (Marker 4) was in complete agreement with the question author's marking of one question (Question A) but there were frequently one or two course tutors who were at variance with the question author and their fellow tutors on a substantial number of responses (e.g. Marker 2 disagreed with the question author for 16.3% of the responses to Question B, Marker 1 disagreed with the question author for 19.3% of the responses to Question C, Marker 6 disagreed with the question author for 33.3% of the responses to Question G).

For six of the questions, the marking of the computer system was in agreement with that of the question author for more than 94.7% of the responses (rising as high as 99.5% for Question A). For Question G there was agreement with the question author for 89.4% of the responses. Further improvements have been made to the answer matching since the human-computer marking comparison took place in June 2007, and in July 2008, the marking of a new batch of responses was found to be in agreement with the question author for between 97.5% (for Question G) and 99.6% (for Question A) of the responses. This is in line with a previous study of the IAT engine's marking (Mitchell, Aldridge, Williamson & Broomhead, 2003) where an accuracy of >99% was found for simple test items.

Question	Number of responses in analysis	Percentage of responses (and κ inter-rater agreement) where the human markers and the IAT system were in agreement with question author		
		Means for the 6 human markers	Ranges for the 6 human markers	IAT
A	189	98.9 (0.97)	97.4 to100 (0.92 to 1)	99.5 (0.98)
B	248	91.9 (0.83)	83.9 to 97.2 (0.75 to 0.94)	97.6 (0.95)
C	150	86.9 (0.71)	80.7 to 94.0 (0.55 to 0.86)	94.7 (0.88)
D	129	96.7 (0.93)	91.5 to 98.4 (0.83 to 0.97)	97.6 (0.94)
E	92	95.1 (0.87)	92.4 to 97.8 (0.79 to 0.95)	98.9 (0.97)

F	129	90.8 (0.81)	86.0 to 97.7 (0.70 to 0.95)	97.7 (0.95)
G	132	83.2 (0.67)	66.7 to 90.2 (0.35 to 0.80)	89.4 (0.79)

Table 2 A comparison of the marking of six human markers and IAT, June 2007

For all questions the mean inter-rater statistics measuring agreement with the question author was higher for the computer marking system.

When the markers were ranked for each question by the inter-rater statistic and the mean ranks of each marker compared, the results (Table 3) reinforce the above table with the computer taking the top rank. According to analysis of variance, any two markers would be significantly different from each other at the 95% confidence level if their mean ranks differed by 1.8 or more. The results show that the markers fell into three distinct groups, significantly different from each other at the 95% confidence level:

- IAT and Marker 4
- Markers 2, 3, 5 and 6
- Marker 1

At the 95% confidence level only Marker 4 marked consistently as well as the computer.

Marker	Mean Rank
IAT	1.9
Marker 4	3.0
Marker 2	4.0
Marker 5	4.0
Marker 3	4.1
Marker 6	5.0
Marker 1	6.1

Table 3 Mean ranks of each marker according to the kappa inter-rater statistics

2.3 Reasons for inaccuracies in human marking

The following have been identified as possible reasons for the divergence in the marking of the course tutors:

The marking guidelines were not sufficiently clear and/or detailed. Every effort had been made to make the guidelines similar to those provided for tutor-marked assignments (which try to enable consistency of marking but also assume a certain amount of subject knowledge and professional judgment on behalf of the tutors) and two out of three tutors report being happy with them.

The response was partially correct, so there was uncertainty over whether a mark was justified or not. In some questions, responses such as these accounted for substantial divergence in grading. For example, the question author had decided that responses such as 'The forces are equal' were insufficient in answer to Question B and three of the human markers agreed. However the other three human markers marked similar responses as correct; in two cases because they considered the responses to be partially correct but worthy of credit. The third marker gave 3 marks for responses such as these in part two of the analysis, which indicates that he or she believed 'The forces are equal' to be a completely correct response.

The response was 'borderline' so there was uncertainty over whether a mark was justified or not. Responses in this category included those that implied the correct answer rather than

stating it explicitly and answers that included a correct answer with an addition that might imply that the student's understanding is flawed. Thus some markers gave no credit for 'They solidify from the molten state' in response to Question C (because this does not say what it is that solidifies) whilst other markers assumed that this information was implicit.

Lack of subject knowledge or understanding on behalf of the human marker. For example, in Question C, one of the course tutors marked responses such as 'From molten rock that has cooled and solidified' as incorrect, not appreciating that magma means 'molten rock' and so that this response is identical in meaning with the one given in the marking guidelines.

Slips/inconsistencies. In addition to the occasions on which one human marker marked consistently in a way that was different from the others, sometimes an individual marked identical or very similar responses as correct on some occasions and incorrect on others. For example, one marker marked 'Through cooling of molten rock' as correct but 'When molten rock is cooled' as incorrect in answer to Question C.

2.4 Reasons for inaccuracies in IAT marking

Mitchell et al. (2002) identified four reasons for inaccurate computer marking. There were examples of each of these in our human-computer marking comparison, though each was relatively rare:

Omission of a mark scheme template. These are essentially cases where the question author has failed to recognise a particular way in which a correct or incorrect answer might be expressed, so they were more common in questions where insufficient responses from students had been analysed. For example, prior to the human-computer marking comparison, the question author had not encountered answers such as 'a sedimentary rock would crumble easily' in response to Question G.

Failure to correctly identify miss-spelled or incorrectly used words. By the time of the human-computer marking comparison these failures were very rare; IAT's handling of poor spelling and grammar is generally excellent. However, misspellings are not recognised when the misspelt word has a different meaning. So, in an example from a different question, 'deceases' was not automatically recognised as a misspelling of 'decreases'.

Failure to properly identify the sentence structure. These failures are rare but can be very frustrating and difficult to overcome. The IAT authoring tool has particular difficulty with responses including the words 'and' or 'or' which means that the exemplary student response 'Gravity and wind resistance are the forces acting on the hailstone and they are equal' in answer to Question B has so far proved impossible to match.

Failure to identify an incorrect qualification (where a correct response is nullified by an incorrect one). Mitchell et al. (2002) identified the difficulty of accurately marking responses that include both a correct and an incorrect response as 'a potentially serious problem for free-text marking'. Jordan & Mitchell (2009, p.380) expand this point to say that 'Whilst any individual incorrect response of this nature can be dealt with by the addition of a 'do not accept' mark scheme in FreeText Author, it is not realistic to make provision for all flawed answers of this type.' An example of a response of this type is 'That there is balanced force acting on the hailstone with more downward force (gravity)' in answer to Question B. The first part of the answer is excellent, but the addition of 'with more downward force' indicates that the student's understanding is flawed. It is in marking responses of this type that human

markers are at an advantage to most computerised systems. However it should also be emphasised (Jordan & Mitchell, 2009, p.380) that 'contrary to e-assessment folklore, responses of this type do not originate from students trying to 'beat the system' …but rather by genuine misunderstanding'.

The human-computer marking comparison identified two other reasons for inaccurate marking by the IAT system:

Marking a correct response as incorrect because it matches a 'do not accept' mark scheme. This problem was very rare, but any issue that leads to correct responses being marked as incorrect must be taken seriously. For example, the correct response 'Extrusive rocks have smaller crystals, and intrusive have larger crystals' matched the 'do not accept' mark scheme 'Extrusive rocks have larger crystals' and so was marked as incorrect . 'Extrusive rocks have smaller crystals. Intrusive have larger crystals' would have been correctly marked.

Marking an incorrect response as correct due to misinterpretation of IAT confidence levels. The IAT system offers two levels of marking which indicate the system's confidence in the mark. The lower confidence mark uses 'flags' to indicate how 'close' a response may be to a correct response. It then lies with the author to decide how to interpret the flags and mark the response. Some of our decisions made using these flags resulted in students being told that their answer was correct when it was not. Problems of this type were relatively common for some questions and for subsequent uses the adjustment for flagged answers has only been applied when it has been shown to improve the overall marking accuracy rather than reducing it.

2.5 Discussion of human-computer marking comparison

Inaccuracy of human marking has been identified as a concern by Orrell (2008) and the UK Office of the Qualifications and Examinations regulator (reported by Frean, 2008) and the current study has demonstrated that computers can mark short-answer questions as accurately as human markers. Although the extent of errors in human marking caused by misunderstandings is alarming, it is not surprising that the computer's marking was more consistent than that of the human markers. In deciding which responses to mark as correct and incorrect, the question author was herself frequently in some doubt, and after marking large batches of responses she usually discovered some inconsistencies in her marking. It is perhaps worth highlighting one difference in the starting position for the human markers and the computer based system. The human tutors were provided with a mark scheme of intended actions while the training sets used by IAT encapsulated those intentions in concrete examples and perhaps these trainings sets provided a clearer interpretation of how the mark scheme should be applied.

The free-text responses in this trial were all marked as 1 or 0; no half-marks were permitted. The difficulty faced by human markers when confronted by a response that they considered to be partly correct has already been discussed. In addition when more complex responses are being marked and/or partial credit allowed, it is possible that the computer's grading will be in less good agreement with that of the question author (as reported by Pulman & Sukkarieh, 2005).

Where the time spent developing the question and the response matching can be justified (for modules with large number of students and where the questions will be reused) computer marking can provide more consistent results. It can also be used to free up course tutors from

the drudgery of marking simple responses, to enable them to concentrate on the marking of assessment tasks in which greater judgement is required and on supporting their students in other ways.

When the concern is with assessment for learning rather than assessment of learning, perhaps the accuracy of marking should not matter too much, but if marks are used to encourage students to engage with the assessment task, they will inevitably be concerned about the accuracy of the marking. Early evidence from summative use of IAT questions on S104 *Exploring Science*, the module that replaced S103, points towards the fact that, whatever the truth of the matter, students have less confidence in computers than in human markers. Rightly or wrongly, students are also likely to have less confidence in the computerised marking of short-answer questions than they have in the marking of more conventional e-assessment tasks, despite the fact that these tasks are sometimes flawed and so can lead to greater student disadvantage than free-text short-answer questions. Further investigation is needed into the impact of these threats to the wider summative use of free-text short-answer e-assessment questions.

Even in purely formative use, accuracy of marking is important because of the importance of giving correct feedback to students; evidence from student observation points towards the fact that if told that an answer is correct (even if it is not) students do not read the final answer provided by the OpenMark assessment system, and so may never realise that their understanding is flawed. The same has been observed of student reaction to inaccurate grading by human markers.

3. A comparison of the marking of different computer systems

3.1 Two algorithmic approaches to matching free-text responses

OpenMark's own response matching algorithm originated in the Computer-Based Learning Unit of Leeds University in the 1970s and has been further developed in recent years at the Open University. Regular Expressions are found in computer languages such as Java and PHP, and Rézeau (2008) has provided a Moodle question type which uses Regular Expressions. Both OpenMark and Regular Expressions rely on computational algorithms for their efficacy but these algorithms contain no knowledge of grammar or syntax.

The OpenMark response matching algorithm enables easy specification of words and synonyms while allowing for some misspellings. It has proved to be straightforward to use and easily comprehensible, and is described in brief in Appendix B.

Regular Expressions are well known to computer scientists as a form of short-hand for specifying search strings. The method is both short and powerful but not necessarily intuitive. For example the Regular Expression

```
'\b[A-Z0-9._%-]+@[A-Z0-9.-]+\.[A-Z]{2,4}\b'
```

will match email names. The Java™ tutorials (http://java.sun.com/docs/books/tutorial/) contains a chapter on Regular Expressions.

3.2. Methodology for computer-computer marking comparison

In summer 2008, an undergraduate student (not of Computer Science) was appointed to the task of trying to obtain adequate answer matching for responses to the same seven questions used in the human-computer marking comparison, using the two algorithmically-based systems described in Section 3.1. He was provided with seven sets of 'training' responses,

one for each of the seven questions, a software harness into which to enter his response matching and the documentation for the algorithms.

The training sets provided for use by the student in developing his answer matching were the responses from students on the October 2007 presentation whilst the IAT answer matching had been developed using responses from the October 2006 presentation; the two batches of responses were deemed to be comparable and by the end of the trial all systems had been tested against all student responses. The training sets ranged in size from 129 to 317 responses. The trainer aimed to match all sets as well as possible but the ease with which he was able to cater for most of the seen responses was unexpected. On the training sets the percentage agreement with the question author ranged from 94.3% to 100%.

The time required to understand how the two algorithms worked and the time needed to produce optimised response matching for each algorithm both indicated that the OpenMark algorithm was easier to use and faster to optimise. Typically two to three questions could be handled in a day with OpenMark whilst Regular Expressions more typically took a day per question, despite always being done second i.e. when the OpenMark matching had been completed. Two response sets proved trickier than the other five indicating the well known phenomenon that the major skill when creating free-text entry questions is experience in knowing what questions to ask.

The response matching was then tested (blind) against the same sets of responses used in the human-computing marking comparison; these were the student responses from the February 2007 presentation (Test 1). After further improvements to the each system's answer matching, all the student responses available at the time (some previously seen, some unseen) were marked by IAT, OpenMark and Regular Expressions, and the results were compared (Test 2). Each system's answer matching was then improved for a final time and the best results obtainable by each were compared (Test 3).

3.3 Results

Test 1 (using responses from the February 2007 presentation; all unseen)

Table 4 gives the results of the initial marking by IAT, OpenMark and Regular Expressions of the responses used in the human-computing marking comparison. None of these responses had been used in developing the answer matching of any of the computerised systems.

Question	Responses in set	Percentage of responses (and κ inter-rater agreement) where computer marking was in agreement with question author		
		Computational linguistics	Algorithmic manipulation of keywords	
		IAT	OpenMark	Regular Expressions
A	189	**99.5 (0.98)**	**99.5 (0.98)**	98.9 (0.97)
B	248	97.6 (0.95)	**98.8 (0.97)**	98.0 (0.96)
C	150	**94.7** (0.88)	**94.7 (0.89)**	90.7 (0.80)
D	129	97.6 (0.94)	96.1 (0.92)	**97.7 (0.95)**
E	92	**98.9 (0.97)**	96.7 (0.91)	96.7 (0.91)
F	129	**97.7 (0.95)**	88.4 (0.76)	89.2 (0.78)
G	132	**89.4 (0.79)**	87.9 (0.76)	88.6 (0.77)

Table 4 A comparison of the initial marking of IAT FreeText Author, OpenMark and Regular Expressions (as used by a summer student) for the same responses as used in the human-computer marking comparison

When Table 3 is extended to include the two new computer based marking systems these two systems are grouped with IAT and Marker 4:

- IAT, Marker 4, Regular Expressions and OpenMark
- Markers 2, 3, 5 and 6
- Marker 1

Test 2 (using responses from the October 2006, February 2007 and October 2007 presentations; one third unseen)
The February 2007 responses were now added to the training sets enabling further development of each system's response matching. Then all the student responses available at the time, the enlarged training set plus the remaining unseen responses, were marked by IAT, OpenMark and Regular Expressions. The results are given in Table 5.

Question	Number of responses in analysis	Percentage of responses where computer marking was in agreement with question author		
		Computational linguistics	Algorithmic manipulation of keywords	
		IAT	OpenMark	Regular Expressions
A	672	**99.6**	99.1	99.1
B	849	97.5	**98.8**	98.0
C	571	97.9	98.1	**98.4**
D	527	97.7	**97.9**	95.3
E	361	**98.9**	**98.9**	**98.9**
F	366	97.8	**98.6**	96.7
G	520	97.5	**98.7**	95.4

Table 5 A comparison of the marking of IAT FreeText Author, OpenMark and Regular Expressions (as used by a summer student) for the full response set (including training set and some further unseen responses)

Test 3 (using responses from the October 2006, February 2007 and October 2007 presentations; all seen)
Finally all responses were included in the training set and each system's response matching was optimised to match as many responses as possible. The best results for each system are compared in Table 6.

Question	Responses in set	Percentage of responses where computer marking was in agreement with question author		
		Computational linguistics	Algorithmic manipulation of keywords	
		IAT	OpenMark	Regular Expressions
A	672	**99.7**	**99.7**	99.6
B	849	98.7	**99.3**	98.6
C	571	**99.5**	**99.5**	99.0
D	527	98.5	**98.7**	95.5

E	361	**100.0**	**100.0**	99.7
F	366	**99.5**	**99.5**	97.3
G	520	**99.8**	99.4	95.8

Table 6 A comparison of the marking of IAT FreeText Author, OpenMark and Regular Expressions (as used by a summer student) : answer matching optimised in the light of the full response set.

The student was able to identify some features of responses that caused difficulties when constructing the response matching (see Section 3.4) and in response to these observations proximity controls have been added to the OpenMark algorithm (The description in Appendix B includes the latest additions). Since the original computer-computer marking comparison, improvements to the figures quoted in Table 6 have been obtained for IAT and, especially for OpenMark (using the improved algorithm and in the hands of a more experienced software developer). However the figures quoted in Table 6 are the original results from summer 2008.

3.4. Responses that were difficult to match with OpenMark

It is worth noting that OpenMark's response matching algorithm, although simple and intuitive to understand and use, is not a simple 'bag of words' system; it can cope with inaccuracies in spelling, and with word order and negation. The responses that were difficult to match at the time of the trial included the following, but improvements to the matching algorithm mean that the first can now also be successfully handled:

Responses where a qualifier could not be linked positively to its object. For example in the response 'if it was not fragmental and by looking for banding' it was not possible to associate the 'not' with just 'fragmental'.

Failure to correctly identify miss-spelled words. Rather than attempting to recognise a 'real' word (as IAT does), the OpenMark matching allows the omission and reversal of letters and thus copes well with many common typographical and spelling mistakes. Problems arise in very short words and when it is the first letter of the word that is incorrect or missing.

Failure to identify an incorrect qualification (where a correct response is nullified by an incorrect one). As for IAT, responses of this type, exemplified by 'That there is a balanced force acting on the hailstone with more downward force (gravity)' remain the most challenging. Any computer system is likely to match the correct part of the sentence but not the incorrect part, and it is in marking responses of this type that human markers are at an advantage, since they are able to spot the logical inconsistency. However, in situations requiring several human markers, the marking of responses of this type is likely to be inconsistent.

3.5 Discussion of computer-computer marking comparison

Tables 4, 5 and 6 all illustrate that OpenMark's response matching routine, a relatively simple algorithmically based system, and in the hands of a relatively inexperienced undergraduate and for a relatively short period of time, appears to be able to provide answer matching on a par with that developed by the question author with the assistance of IAT's authoring tool. This result was a huge surprise, and current work, comparing the marking of OpenMark's and IAT's answer matching more systematically and for a range of short answer questions in summative use, is pointing towards the conclusion that OpenMark is indeed able to provide high-quality answer matching for short answer free-text questions. The fact that the results for Regular Expressions were slightly less good is probably a result of the fact that this was more difficult for the undergraduate student to learn how to use.

Learning from student responses

It is important to note that, whether a system based on computational linguistics (IAT) or an algorithmically-based system (OpenMark) is used, the fact that responses from real students are used in developing the answer matching appears to have been a significant feature in developing answer matching that is, generally, more accurate and reliable than that of human markers. Previous users of similar software (e.g. Mitchell et al., 2003; Sukkarieh, Pulman & Raikes, 2003) have used student responses to paper-based questions in order to provide appropriate answer matching for the computer-based version, but this approach makes the assumption that there are no characteristic differences between student responses to the same question delivered by different media, or between responses that students assume will be marked by a computer as opposed to a human marker.

We were fortunate that we were able to develop the answer matching on the basis of responses from S103 students, although a drop down in use by S103 students (similar to that observed in other formative-only use of e-assessment) led to us having fewer responses than we might have hoped for on which to develop our answer matching.
Subsequent work has shown that student responses to the questions in summative use are the most useful in identifying misconceptions and phraseologies and we now have huge data-sets of such responses. This leads to the paradox that student responses to summative questions would be very useful as an aid to question development, but yet questions need to be fully developed (or as close to this point as possible) before being used summatively. An alternative solution would be to ask course tutors to mark responses (which could be gathered electronically from students) in the first instance. Once all the responses and marks were held in a computer we would have the raw material for training the computational response matching systems.

How many responses are required?

Mitchell et al. (2003) used paper-based marking guidelines and approximately 50 marked student scripts in developing their answer matching. Sukkarieh et al. (2003) used approximately 200 marked student answers per question for training, and approximately 60 answers per question for testing. Our experience, for improving the answer matching of the IAT system, OpenMark, and Regular Expressions, is that the number of responses required to develop sufficiently robust answer matching varies hugely from question to question.

Table 4 shows that Question G was initially badly marked by all three systems, indicating that this question provided more scope than the others for providing alternative correct and incorrect responses; Table 5 shows how the scoring on this question improved with a larger training set. For six of our questions training sets between 100 and 250 responses gave us a good base but for the seventh question we had to go well above this number.

Around 15 of the short-answer free-text questions are now in use, alongside conventional OpenMark questions, in regular summative but low stakes iCMAs on S104 *Exploring Science*. For typical summative use on S104, each data-set contains between 1500 and 2000 student responses; the problem moves from one of having insufficient data to one of struggling to find the time to allocate accurate marks for the purposes of training and evaluation. There is a need to do all that we can to automate the process, or at least to support the (human) question author with appropriate technology.

Advantages and disadvantages of computational linguistics and algorithmic approaches

Knowing that computational linguistics is behind the IAT response matching algorithms has provided an element of respectability to the marking process. However, from this preliminary study, it appears that the accuracy of marking of the algorithmically-based OpenMark matching is equally effective.

It took the question author several weeks to become proficient in the use of the IAT authoring tool. This can be compared with a couple of hours for the summer student to become capable in using OpenMark's response matching and a day or two for him to work through the description of Regular Expressions. From an untrained state the summer student completed all the work required for Tests 1-3 in 15 days, producing 14 sets of response matching (7 OpenMark, 7 Regular Expressions) in that time. Of the approaches using computational manipulation of keywords, the OpenMark approach was both simpler to operate and produced more satisfactory results (see Table 5). Further work on this project will concentrate on the IAT and OpenMark algorithms.

Is the achieved accuracy of the marking satisfactory? Mitchell et al. (2003) and Pulman & Sukkarieh (2005) report very similar percentage accuracy figures to ours for their computer-based marking of short free-text responses. Relative to these, and more significantly to the results of the human computer marking comparison reported in Table 2, we feel that all the figures for accuracy of marking quoted in Table 5 are very acceptable, with the possible exception of the two results for Regular Expressions that fall below 96%.

Do responses in formative and summative modes differ?
We now have data sets from questions used both formatively and summatively. Not surprisingly, responses to questions in summative use (even if the weighting is very low) are characteristically different from those in formative only use. They are more likely to be correct, more likely to be expressed in sentences and longer. In extreme cases, answers of more than a hundred words, written in several sentences have been received.

How complex is the response matching? This is quite variable. However for the response matching success rates reported in Table 5 the span is between 4 and 15 lines of response matching per question for OpenMark. That is for all questions at the level of success reported the response matching task is tractable. However for some questions (of those in the study, this is particularly the case for Question B), in trying to improve the answer matching beyond the point reported, a very large increase in complexity is introduced. For Question B, the OpenMark code increased to 40 lines, exemplifying the laws of diminishing returns both in time expended and number of responses matched by each new added line.

The IAT authoring tool specifies top level markschemes, then model answers within each markscheme and then synonyms for keywords in each model answer. It is the model answers that form the templates that are used to mark each student response. For Question B, the IAT project includes three mark schemes ('The forces are balanced'; 'The upward force is equal to the downward force' and 'There is no resultant force', with 9-13 model answers for each mark scheme and a range of synonyms (e.g. 'up', 'resistive', 'frictional' for 'upward')

Is there logic behind the OpenMark response matching? This is the key to using this technology more widely and has been studied most closely with questions A and B. There does indeed appear to be logic but extracting that logic and using it is the focus for the next set of work.

Recommendations for further work

Previous work (Mitchell et al, 2003; Jordan & Mitchell, 2009) and the current study show that with due care computational systems can provide response matching of short-answer questions that is on a par with human markers. However to widen the use of the approach it will be necessary to find more efficient routes to generate the appropriate response matching. A further study of the logic behind the derived response matching may provide insights as to how to improve the efficiency; Pulman & Sukkarieh (2005) have attempted to use machine learning to generate response matching patterns, with limited success.

The issue of ambiguous responses requires further work. This study has marked all responses as right or wrong and the author has agonised over whether to award a mark to responses that contain the right answer in combination with phrases that raise concern over the student's understanding. If such responses can be identified, it may be appropriate to award a partial mark, to refer these responses for human marking or to tell the student that the system cannot mark their response, so they should try again, phrasing their response more carefully.

All of the above response matching sits within an interactive assessment system that is designed to give instantaneous feedback. The university is committed to supporting its students in this way as a means of improving learning (e.g. Gibbs & Simpson, 2004) but with automated response matching that cannot guarantee 100% accuracy there is a danger of misleading students about their level of understanding. Clearly further investigation is needed into the ways in which students engage with e-assessment tasks and the feedback provided.

Acknowledgements

The authors gratefully acknowledge the financial support of the UK Higher Education Funding Council via the Centre for Open Learning of Mathematics, Computing, Science and Technology (COLMSCT), the insights and enthusiasm of the summer student, George Butcher, and assistance from Tom Mitchell of Intelligent Assessment Technologies Ltd.

References

Attali, Y. & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. *Journal of Technology, Learning and Assessment, 4* (3).

Butcher, P.G. (2006) OpenMark Examples http://www.open.ac.uk/openmarkexamples

Butcher, P.G., Swithenby, S.J. & Jordan, S.E. (2009). eAssessment and the independent learner. 23rd ICDE World Conference on Open Learning and Distance Education, 7-10 June 2009, Maastricht, The Netherlands. Available from http://www.open.ac.uk/colmsct/resources

Frean, A. (2008) GCSE and A level exam results inaccurate, Ofqual warns pupils. Retrieved 5 May 2009 from http://www.timesonline.co.uk/tol/life_and_style/education/article3941601.ece

Gibbs, G. & Simpson, C. (2004) Conditions under which assessment supports students' learning. *Learning and Teaching in Higher Education,* 1, 3-31.

Jordan, S. (2009) Analysis of the impact of iCMAs on student learning. http://www.open.ac.uk/colmsct/projects/sejordan

Jordan, S. & Mitchell, T. (2009) E-assessment for learning? The potential of short free-text questions with tailored feedback. *British Journal of Educational Technology, 40*(2), 371-385.

Leacock, C. & Chodorow, M. (2003). C-rater: Automated Scoring of Short-Answer Questions. *Computers and Humanities, 37*(4), 389–405.

Mitchell, T., Russell, T., Broomhead, P. & Aldridge, N. (2002). Towards robust computerised marking of free-text responses. 6th International CAA Conference, Loughborough, UK. Retrieved 5th May 2009 from http://www.caaconference.com/pastConferences/2002/proceedings/index.asp

Mitchell, T, Aldridge, N, Williamson, W & Broomhead, P (2003) Computer based testing of medical knowledge. 10th Computer Assisted Assessment Conference, Loughborough, July 2003. Retrieved 5th May 2009 from http://www.caaconference.com/pastConferences/2003/proceedings/index.asp

Nicol, D.J. (2007). E-assessment by design: using multiple choice tests to good effect. *Journal of Further and Higher Education*, *31*(1), 53–64.

Orrell, J. (2008) Assessment beyond belief : the cognitive process of grading, in A. Havnes and L. McDowell (eds) *Balancing Dilemmas in Assessment and Learning in Contemporary Education*. London: Routledge. pp 251-263.

Pearson (2009) Intelligent Essay Assessor. Retrieved 21st December 2009 from http://www.knowledge-technologies.com/prodIEA.shtml

Pulman, S. & Sukkarieh, J. (2005) Automatic Short Answer Marking. Proceedings of the 2nd Workshop on Building Educational Applications Using NLP, Ann Arbor, June 2005.

Rézeau, J (2008) Regular Expression Short Answer question type for Moodle, http://docs.moodle.org/en/Question_types

Ross, S.M., Jordan, S.E & Butcher, P.G. (2006). Online instantaneous and targeted feedback for remote learners. In C. Bryan & K.V. Clegg, K.V. (Eds), *Innovative assessment in higher education* (pp. 123–131). London: Routledge.

Siddiqi, R. & Harrison, C.J. (2008) On the automated assessment of short free-text responses. Paper presented at the 34th International Association for Educational Assessment (IAEA) Annual Conference, Cambridge, UK, September 2008.

Sukkarieh, J.Z, Pulman, S.G. and Raikes, N. (2003) Auto-marking: using computational linguistics to score short, free-text responses. Paper presented at the 29th th International Association for Educational Assessment (IAEA) Annual Conference, Manchester.

Sukkarieh, J.Z, Pulman, S.G. and Raikes, N. (2004) Auto-marking 2: using computational linguistics to score short, free-text responses. Paper presented at the 30th th International Association for Educational Assessment (IAEA) Annual Conference, Philadephia.

Appendix A The seven questions

A What does an object's velocity tell you that its speed does not?

B A snowflake falls vertically with a constant speed. What can you say about the forces acting on the snowflake?

C How are igneous rocks formed?

D You are handed two rock specimens and you are told that one is an intrusive igneous rock whilst the other is an extrusive igneous rock. How would you know which was the intrusive specimen?

E Why do intrusive igneous rocks have larger crystals than extrusive ones?

F You are handed a rock specimen that consists of interlocking crystals. How could you be sure, from its appearance, that this was a metamorphic rock?

G You are handed a rock specimen from a cliff that appears to show some kind of layering. The specimen does not contain any fossils. How could you be sure, from its appearance, that this rock specimen was a sedimentary rock?

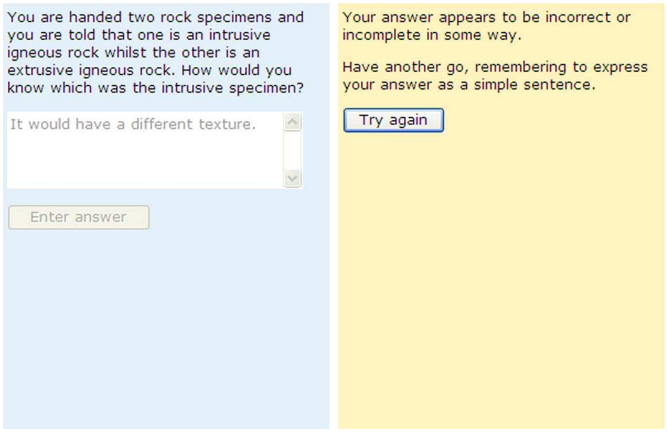
Appendix B OpenMark response matching features as at June 2009
(features indicated * have been introduced since the study described in this paper)

The matching options are:

Matching option	Description
allowExtraChars	Extra characters can be anywhere within the word.
allowAnyWordOrder	Where multiple words are to be matched they can be in any order.
allowExtraWords	Extra words beyond those being searched for are accepted.
misspelling: allowOneCharReplace	Will match a word where one character is different to that specified in the pattern. The pattern word must be 4 characters or greater for replacement to activate.
* misspelling: allowTransposeTwoChars	Will match a word where two characters are transposed. The pattern word must be 4 characters or greater for transposition to activate.
misspelling: allowOneCharExtra	Will match a word where one character is extra to that specified in the pattern. The pattern word must be 3 characters or greater for extra to activate.
misspelling: allowOneCharFewer	Will match a word where one character is missing from that specified in the pattern. The pattern word must be a 4 characters or greater for fewer to activate.
misspellings	This combines the four ways of misspelling a word described above.
* allowProximityOfN	Where 0 <= N <= 4. Sets the number of words allowed between words that are governed by the proximity rule.

Special characters provide more localised control of the patterns:

Special character	Description
Word AND	'space' delimits words and acts as the logical AND.
Word OR	| between words indicates that either word will be matched. | delimits words and acts as the logical OR.
* Proximity control	'_' between words indicates that words must be in the order given and with no more than N (where 0 <= N <= 4) intervening words. _ delimits words and also acts as logical 'AND'.
* Word groups	[] around multiple words enables word groups to be accepted as alternatives to single words in OR lists. Where a word group is preceded or followed by the proximity control the word group is governed by the proximity control rule that the words must be in the order given.
Single character wildcard	# matches any single character.
Multiple character wildcard	& matches any sequence of characters including none.



You are handed two rock specimens and you are told that one is an intrusive igneous rock whilst the other is an extrusive igneous rock. How would you know which was the intrusive specimen?

They would have different crystal size.

Enter answer

Your answer still does not appear to be correct.

You are on the right lines but your answer is not complete. You need to identify whether intrusive rocks have bigger or smaller crystals than extrusive rocks. See Book 2 Activity 5.1 and Section 5.2.1.

Try again

You are handed two rock specimens and you are told that one is an intrusive igneous rock whilst the other is an extrusive igneous rock. How would you know which was the intrusive specimen?

The intrusive rock would have bigger crystals.

Enter answer

Your answer is correct.

The crystals in intrusive igneous rocks are larger than those in extrusive igneous rocks. See Book 2 Activity 5.1. and Section 5.2.1.

☐ If you believe that the computer has marked your answer inaccurately please tick this box and your answer will be reviewed by the course team.

Next question