TOWARDS ROBUST COMPUTERISED MARKING OF FREE-TEXT RESPONSES

Tom Mitchell, Terry Russell, Peter Broomhead and Nicola Aldridge

Towards Robust Computerised Marking of Free-Text Responses.

Tom Mitchell[1]
Terry Russell[2]
Peter Broomhead[3]
Nicola Aldridge[1]

1. Intelligent Assessment Technologies
2. Centre for Research in Primary Science and Technology, University of Liverpool
3. Department of Systems Engineering, Brunel University

tom@intelligentassessment.com

Abstract

This paper describes and exemplifies an application of AutoMark, a software system developed in pursuit of robust computerised marking of free-text answers to open-ended questions. AutoMark employs the techniques of Information Extraction to provide computerised marking of short free-text responses. The system incorporates a number of processing modules specifically aimed at providing robust marking in the face of errors in spelling, typing, syntax, and semantics. AutoMark looks for specific content within free-text answers, the content being specified in the form of a number of mark scheme templates. Each template represents one form of a valid (or a specifically invalid) answer. Student answers are first parsed, and then intelligently matched against each mark scheme template, and a mark for each answer is computed. The representation of the templates is such that they can be robustly mapped to multiple variations in the input text.

The current paper describes AutoMark for the first time, and presents the results of a brief quantitative and qualitative study of the performance of the system in marking a range of free-text responses in one of the most demanding domains: statutory national curriculum assessment of science for pupils at age 11. This particular domain has been chosen to help identify the strengths and weaknesses of the current system in marking responses where errors in spelling, syntax, and semantics are at their most frequent. Four items of varying degrees of open-endedness were selected from the 1999 tests.

These items are drawn from the real-world of so-called ‘high stakes’ testing experienced by cohorts of over half a million pupils in England each year since 1995 at ages 11 and 14. A quantitative and qualitative study of the performance of the system is provided, together with a discussion of the potential for further development in reducing these errors. The aim of this exploration is to reveal some of the issues which need to be addressed if computerised marking is to play any kind of reliable role in the future development of such test regimes.

Introduction

A broad classification of test items can be made in terms of 'Select' (coded, multiple-choice) or 'Generate' (open-ended) responses. A ‘generate’ form of response is one in which the mental processing of the respondent must include the importation of information which is not presented on the question page. This implies a different form of memory retrieval as compared with ‘select’ responses and the additional demand of constructing meaning, rather than choosing from a ready-constructed array.

The two item types are different modes that also assess qualitatively differently, so are more than simply arbitrarily selected equivalents. Due to limitations in assessment technology however, Select is the predominant question type used in CAA. A move towards automatic assessment of open-ended responses will enrich some crucial aspects of CAA.

This paper describes a software system that employs natural language techniques to mark open-ended (free-text) responses. The software, called AutoMark, has been under development for almost three years, and has been employed in a commercial eLearning product for the last twelve months (ExamOnline 2002). CAA procedures using the system are currently being developed at a number of higher education establishments, including Brunel University where an online Java test for first year engineering students is being developed.

The current paper describes AutoMark for the first time, and presents the results of a brief quantitative and qualitative study of the performance of the system in marking a range of free-text responses in one of the most demanding domains: statutory national curriculum assessment of science for pupils at age 11. This particular domain has been chosen to help identify the strengths and weaknesses of the current system in marking responses where errors in spelling, syntax, and semantics are commonplace and occasionally, at the limits of comprehensibility. This exploratory study provided feedback to enable the system developers to improve the robustness of the system for such demanding real-world applications. It also enabled the test developers to gain insight into the accommodations that might be needed (e.g. in mark scheme construction) to enhance the viability of free-text CAA applied to such items.

CAA of Free-Text Responses

Automated assessment of free-text responses has been the subject of some recent research. A good summary paper can be found in (Whittington, Hunt, 1999).

Perhaps the most well-known system is *e-rater* (Burstein, Leacock, Swartz, 2001), an automatic essay scoring system employing a holistic scoring approach. The system is able to correlate human reader scores with automatically extracted linguistic features, and provide an agreement rate of over 97% for domains where grading is concerned more with writing style than with content.

A novel approach was described by (Callear, Jerrams-Smith, Soh, 2001). The prototype Automated Text Marker is a content-based system for marking short free-text responses. The system identifies concepts in the text, and the dependencies between them. Matching is carried out between the concepts and dependencies found in students' answers and those found in the model answers.

A more generic technology that shows high promise is that of Latent Semantic Analysis (LSA) (Landauer, Dumais, 1997). LSA has been applied to essay grading, and high agreement levels obtained (Landauer, Foltz, Laham, 1998).

The system described in this paper employs the techniques of Information Extraction (Cowie, Lehnert, 1996) to provide computerised marking of short free-text responses. The system incorporates a number of processing modules specifically aimed at providing robust marking in the face of errors in spelling, typing, syntax, and semantics.

The System

It is axiomatic that providing computerised marking of free-text responses requires analysis of free-text. Traditionally there have been two approaches to this problem.

- Keyword analysis.
- Full natural language processing.

Keyword analysis, which analyses the text by looking for the presence or absence of predetermined key words, is simple to implement, but offers poor performance in practice. Full natural language processing (NLP) systems offer more potential, but are prohibitively complex and expensive to develop. In addition, robust in-depth analysis of free-text is still beyond the current state-of-the-art (Brill, Mooney, 1997).

More recently, a new type of NLP system has emerged employing a technique known as Information Extraction (IE) (Cowie, Lehnert, 1996). IE makes use of NLP tools (parsers, lexical databases, etc), but rather than attempting an in-depth language analysis, skims the input text searching for specific concepts. IE can provide real working systems for specific domains, with measurable performance metrics. (Glasgow, B., Mandell, A., Binney, D., Lila, G., Fisher, D., 1997) (Wenzel, C. 1997) (Mani, Maybury 1999).

Computerised marking of free-text responses can be framed as an information extraction task. Student responses represent the free-text that requires analysis. Each question represents a different domain for the system and the concepts of interest for each domain are correct (or specifically incorrect) responses for that question. A system based on this approach is described in the remainder of this section.

System Overview

AutoMark employs NLP techniques to perform an intelligent search of free-

text responses for predefined computerised mark scheme answers. This is analogous to the process carried out by human markers when marking free-text responses. And like human markers, the system attempts to identify the understanding expressed in a free-text response, without unduly penalising the student for errors in spelling, grammar, or semantics.

The system employs a mark scheme that specifies acceptable and unacceptable answers for each question. The system represents mark scheme answers as syntactic-semantic templates. Each template specifies one particular form of acceptable or unacceptable answer. For example, **Figure 1** illustrates a simple template for the mark scheme answer **The Earth rotates around the Sun**.

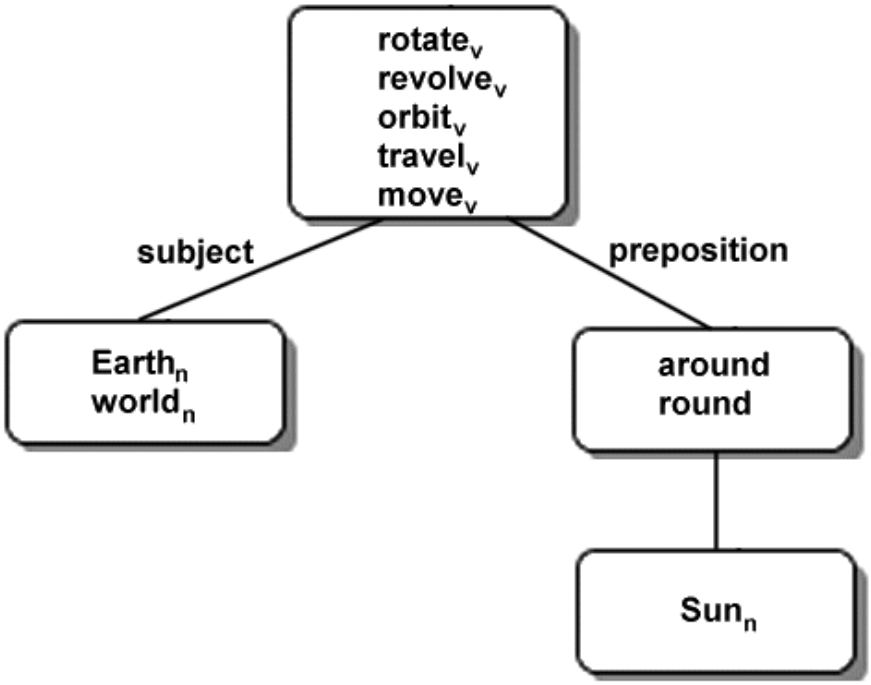

Figure 1. Illustration of a simple mark scheme template.

The template shown can be expected to match a student response if the response contains one of the stated verbs (**rotate, revolve, orbit, travel, move**) with one of the stated nouns (**Earth, world**) as it's subject, and **around / round** the **Sun** in it's preposition. Verbs in the student response are lemmatised (reduced to their base form, i.e. 'went' lemmatises to 'go') so that, for example, the following student responses will all be matched by the template shown above.

The world rotates round the sun.
The earth is orbiting around the sun.
The earth travels in space around the sun.

Development of the templates in the computerised mark scheme is an offline

process, achieved using a custom-written system configuration interface. **Figure 2** illustrates the operation of the system, and shows the main computational operations which the system performs.

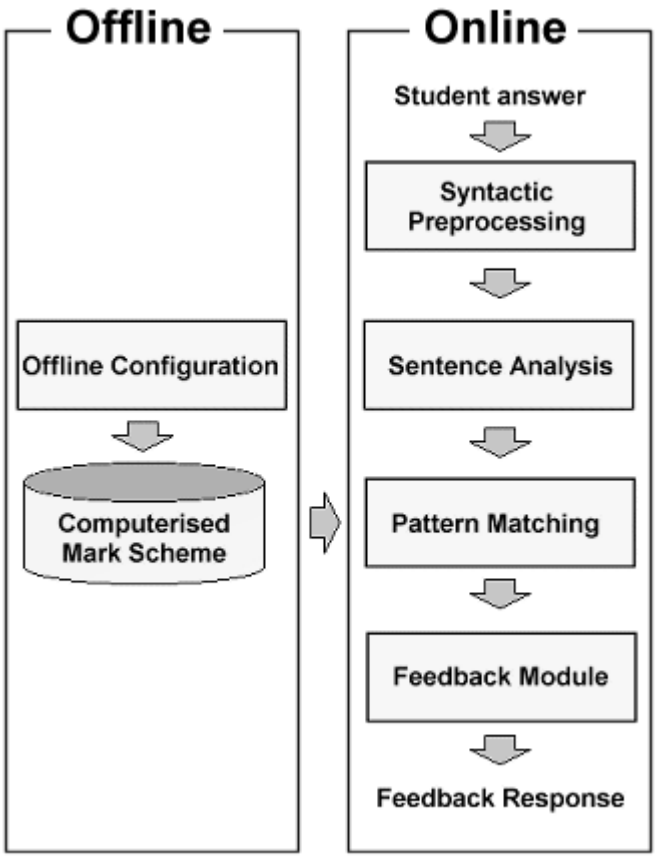

Figure 2. System Architecture.

The marking process progresses through a number of stages. First, the incoming text is pre-processed to standardise the input in terms of punctuation and spelling. Then, a sentence analyser identifies the main syntactic constituents of the text, and how they are related. The pattern-matching module searches for matches between the mark scheme templates and the syntactic constituents of the student text. Finally, the result of the pattern match is processed by the feedback module. Feedback will typically be in the form of a mark, but more specific feedback can be configured. An example question providing structured feedback for English comprehension can be found on the Intelligent Assessment website (Intelligent Assessment Technologies, 2002).

The Experiments

The performance of the system has been tested in collaboration with the Centre for Research in Primary Science and Technology (www.cripsat.org.uk). Statutory national curriculum assessment of science for pupils at age 11 was the domain explored, test items being drawn from archive material. For the purposes of this study, four items of varying degrees

of open-endedness were selected from the 1999 papers. The form of response of each of the four items, in increasing order of linguistic complexity, was as follows:

- Single word generation (1999, paper B, question 2a).
- Single value generation (1999, paper A, question 7d).
- Generation of a short explanatory sentence (1999, paper A, question 9b).
- Description of a pattern in data (1999, paper B, question 6d).

All four items were scored to one mark except the last, which was scored to two marks. 120 responses were randomly selected for each item. Hand-written pupil responses were transcribed using the exact spelling, syntax and punctuation as written on the test papers. The forty pupils at each level are the same for each of the four items.

Two experiments were devised to test the software using these data. The first we have termed the ***blind*** experiment, the second the ***moderation*** experiment. These are described below. The blind experiment, carried out by CRIPSAT, treated the system as a 'black box' and considered some of the educational implications of discrepancies between human and computer marking. The moderation experiment, carried out by Intelligent Assessment Technologies, examines the system errors inherent in the current system.

The Blind Experiment

An initial computerised mark scheme was developed by IAT for the blind experiment. Mark scheme templates were devised and tested using the model answers from the paper-based mark scheme, augmented by a small number (approx. 50) of answers devised to cover the range of expected pupil responses. This version of the computerised mark scheme is referred to in the remainder of the paper as ***unmoderated***.

Student responses were submitted to the AutoMark scoring process by CRIPSAT via the Intelligent Assessment website, using a 'blind' clerical procedure. The outcomes of human and computer marking were then compared.

The Moderation Experiment

Subsequent to the completion of the blind experiment, the responses used in the blind experiment were used to moderate the unmoderated computerised mark scheme. The improved version of the computerised mark scheme is referred to as ***moderated***. For the system being described, moderation is required to cope with :

- unexpected but allowable responses;
- unexpected but allowable synonyms;
- and unexpected but allowable phraseology

encountered in student responses. For example, for the following question /

mark scheme :

> **Question :** Why are some wild flowers highly scented with brightly coloured petals?
>
> **Mark Scheme Answer :** To attract insects.

the following responses were awarded marks by human markers, but were not covered by the unmoderated computerised mark scheme.

- *so they can show up more*
- *so the creature notice them and pollinates the flower*

Subsequent to moderation, a further test of the marking accuracy was then carried out using the moderated computerised mark scheme. For the purposes of these exploratory experiments, the same data (i.e. student responses) were used for the moderation experiment and the blind experiment. Consequently, the accuracy figures from the moderation experiment cannot be regarded as indicative of the expected performance on unseen samples. However they do serve the main purpose of the moderation experiment: to identify those errors which are inherent in the software, rather than those which can be addressed by moderation.

Results and Analysis

This section details the results of each experiment.

The Blind Experiment

The results of this experiment are summarised in **Table 1**.

	Paper B Q. 2a (n=120)	**Paper A Q.7d (n=120)**	**Paper A Q. 9b (n=120)**	**Paper B Q. 6d (n=120)**	**All (n=480)**
Item classification	Single word	Single value	Explanatory sentence	Pattern description	
Matches	118	119	111	100	448
% Match	(98.3)	(99.2)	(92.5)	(83.3)	(93.3)
False positives	1	0	1	0	2
False negatives	1	1	8	20	30

Table 1. Comparison of human and computerised marking outcomes for the blind experiment.

Key Stage 2 1999 Science National Test, Paper B, Question 2a: Single word generation

A photograph shows two children as they 'put an ice lolly in a dry glass jar'. A second photograph shows the outcome: 'After 105 minutes they saw that the ice lolly had turned to liquid. The question asks, '**What is the name of the process when a solid turns into a liquid**?'

This item is classified as posing a 'generate' demand (as there is no identification of the target process, 'melting', on the question page). A single word response will suffice, so the classification is 'generate/word', though pupils may gain credit for responses using more than a single word.

Two discrepancies were recorded between the human and computer marking. In the first case, a pupil had written '*meted*' and this was judged by the human marker to constitute a sufficiently close approximation to '*melted'* to be creditworthy. This can be interpreted as an experienced professionally-informed judgement concerning the likely errors and intentions of low-achieving pupils.

The second discrepancy was in respect of a response in which a pupil wrote, *'melting = condensaition'*. While the word '*melting'* has clearly triggered a positive computer response, the human marker has taken into account the additional incorrect element – the suggestion that melting can be equated with the process of condensation. This form of error is frequently encountered in national science test responses. The marking strategy for dealing with it refers to the 'list rule', so-called because pupils often respond with more than one response in a 'list' of offerings – perhaps intended as a kind of 'scatter-gun' shot at the target. In the case that some wrong science is included, it is treated as negating the correct science offered and results in a zero mark for that item. It might be anticipated that the identification of such incorrect science negating correct science could be very difficult to anticipate, simply because of the difficulty of delimiting the set of potential wrong ideas.

Key Stage 2 1999 Science National Test, Paper A, Question 7d : Single value generation

The second item-type required pupils to generate a single value (referring to a force in newtons). The context is balanced forces: a teacher asking children to turn a balance sideways. One child, Kerry, holds the scales at waist level while a second, Jason, pushes against them in a horizontal direction. It is explained that Jason's push measures 80N and the question is asked, '**What is the size of Kerry's push?**'. The required creditworthy response (80N) might be expressed in a number of acceptable ways, including a reference to the equivalence of the magnitude of the forces, as for example: '*The size of Kerry's push is the same*'.

In the 120 marking comparisons made, only one discrepant case emerged and this was a single false negative. The response offered was, '*It is about*

the same as Jason's.', marked correct by the human marker but incorrect by the software. A first reaction might be that the computer is acting according to its kind - insisting on greater accuracy, offering less leniency than the human. The marker's judgement is itself a professionally-informed decision and physics educators typically argue long and hard over such decisions. As such, the human-computer discrepancy is open to debate.

Key Stage 2 1999 Science National Test, Paper A, Question 9b : Generating a short explanatory sentence.

The third item to be discussed invited a more complex form of response than the two previously analysed, in that a one-word response would not suffice. The context presented was the life-cycle of flowering plants and the specific question posed was, '**Why are some wild flowers highly scented with brightly coloured petals?**'

A short response in the form of an explanatory sentence was the minimum required, though 'sentence' is used loosely, rather than in any strict grammatical sense.

Overall, the rate of human-computer matched marking outcomes was achieved with 92.5% of the 120 responses. In several cases the software appears not to have recognised vernacular grammar and vocabulary, including :

1. *To acttract insects*
2. *To aracted insects.*
3. *They are so the bee's see them*
4. *So that when the bees come for pollen they can spot them.*
5. *So bees come. They take pollin seeds they some times drop some on the floor and start a new life cycle.*

The marking discrepancy associated with another example is due to a human marker error, one which failed to convince the software:

6. *the bee's & wasp's*

This instance is a reminder that human errors in marking need to be taken into account is considering human-CAA reliability comparisons. In this last case, the computerised marking has proven more accurate than the human marking.

Key Stage 2 1999 Science National Test, Paper B, Question 6d : Generation of a description of a pattern in data.

The fourth item was the most complex in terms of the language required by students to frame a response. The context used a toy car and a 'push-meter' – a device which (like an inside-out newton meter) allows the starting push on a car to be calibrated in newtons. A table of data is presented in which the independent variable is defined as 'Starting force in N' and the DV as

‘Distance moved (by a toy car) in cm.’ The demand is for pupils to express the (direct) relationship between two continuous variables. To do so, they must resort to a form of language that differs from everyday usage. A creditworthy response must express both IV and DV in comparative terms, as in, ‘*The greater the starting force, the greater the distance the car moves*.’, or the converse – ‘*The smaller the force, the less the distance*.’ Clearly, there are several ways of expressing the same idea.

While the use of two comparatives gains two marks, the common error of referring only to one data pair, or to extremes in the data such as, ‘*The biggest push moves the car the furthest distance*.’ gains only one mark.

In total across 120 responses, there were 20 discrepant marking decisions, all of these being false negatives on the part of the software, (assuming the human markers’ judgements to be correct).

Undoubtedly, some responses occupy a middle ground between ‘definitely correct’ and ‘definitely incorrect’. In considering discrepancies, it is helpful to consider some of the responses that might be considered ‘problematic’ or ‘borderline’ for human markers:

‘If you pull back more the force that is pushing goes farer on the object.’
‘if it has a lot more starting force, the car would move more longer.’
‘The size of the starting affect by the N get bigger and the Distance get bigger.’
‘The longer she/he pulls the futier the car goes’.
‘The stronger the pull, the further the car goes’
‘the hever it is the more it moved’.
‘The size afects it because there is more power to make it go further.’
‘The harded the push the further and faster it goes.’
‘bigger the source the further it gos’.

The IV (‘starting force in N’) and DV (distance travelled in cm.) are clearly defined in the question presentation, and respondents would be best advised to refer to these explicitly in their answers. In fact, they do not, but re-define the variables in all kinds of ways. Since the validity of the assessment is judged on the extent to which the test succeeds in measuring children’s *science understanding* rather than their expressive parsimony, such variability must be accommodated.

The information brought to bear by the human marker is worth noting here. When the subject of a phrase is not explicit, the human marker will resort to inference to determine what the respondent is referring to; this usually involves looking back at the question stem to confirm an implicit reference. For example, ‘*bigger the source*’ and ‘*hever’* (read as ‘heavier’) were treated as referring to the magnitude of the starting force.

The CAA performance for this item type was significantly below that for the other item types.

The Moderation Experiment

Subsequent to the completion of the blind experiment, the responses used in the blind experiment were used to moderate the computerised mark scheme. A further test of the marking accuracy was then carried out using the moderated computerised mark scheme. **Table 2** shows the results and **Figure 3** shows a graphical comparison with the results from the unmoderated mark scheme.

	Paper B Qu 2a (n=120)	**Paper A Qu 7d (n=120)**	**Paper A QU 9b (n=120)**	**Paper B Qu 6d (n=120)**	**All (n=480)**
Item classification	Single word	Single value	Explanatory sentence	Pattern description	
Matches	120	120	118	105	463
% Match	(100)	(100)	(98.3)	(87.5)	(96.5)
False +	0	0	1	1	2
False -	0	0	1	14	15

Table 2 : Comparison of human and computerised marking outcomes for the moderation experiment.

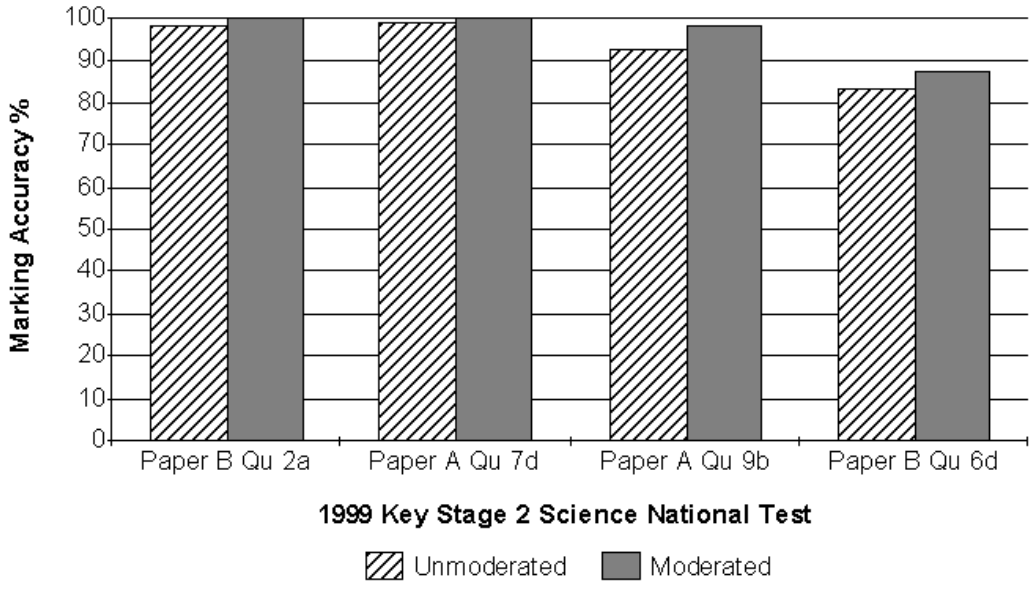

Figure 3. Graph showing the accuracy of computerised marking for both moderated and unmoderated mark schemes when compared to the marks awarded by a human marker.

System Robustness

Of the 420 positive student responses (there were 60 non-responses) to the test items, 388 were correctly marked by the unmoderated software and 403 by the moderated software. Included in these correctly marked responses

were many which included significant errors in spelling, syntax, and semantics. For example :

- *to etract the flys and other creatures*
- *because they atrackt insecs*
- *Because they want to atracted bugs*
- *The more the force the more ferther the car will travel*
- *it affects the distance bucuse the biger force and futher it goes back the futher it goes*
- *the more newtons her push was, the further the car went*

These responses illustrate the ability of the software to correctly mark a wide range of 'real-world' responses. However there still remain discrepancies between the computerised marking and the human marking. These errors are the result of limitations in the current system, and will be analysed in the next section.

System Errors

From a systems viewpoint, there are four recognisable sources of error in the computerised marking. These are :

- failure to correctly identify mis-spelled or incorrectly used words;
- failure to properly analyse the sentence structure;
- failure to identify an incorrect qualification;
- omission of a mark scheme template.

An explanation of each error category is provided below, followed by data quantifying the prevalence of each.

Failure to Correctly Identify Mis-spelled or Incorrectly Used Words

While over 90% of spelling mistakes were correctly identified by the software, 10% were not. The spelling errors found in student responses are often due to phonetic attempts at a correct spelling. This is illustrated by the following responses given by students :

- *To attract inisets*
- *to atracte insexe*
- *to atraked insects*

Note that the system correctly identified the intended words in all of the above examples.

It is worth stating that this category also includes mis-spellings which are inadvertently correct spellings of words other than those intended (e.g. '*meted'* for '*melted')*, of which a sub-set are homophones, e.g. '*grater'* for '*greater'*. A further sub-division of this error category is perhaps warranted.

Failure to Properly Analyse the Sentence Structure

The second biggest source of error (29.4%, 5 responses) is due to the sentence analyser failing to properly analyse the response. This class of error is largely responsible for the lower marking accuracy evident in marking the two-mark item. These kinds of errors commonly occur where there is a poorly constructed student response, and as such constitute a particular problem with children's writing, as illustrated in the responses below:

- *The size of the starting affect by the N get bigger and the Distance get bigger.*
- *The less push the less spead and it will not go far*

In cases such as these, the system is unable to identify a correct answer within a sentence because the sentence analyser is unable to process and identify the structure of that sentence. The solution to this class of error lies with the development of a more robust sentence analyser. Research into this area is ongoing within IAT.

Failure to Identify an Incorrect Qualification

Students' responses sometimes comprise a correct statement qualified by (or supplemented by) an incorrect statement. Invalid qualifications that negate a correct answer should result in a reduction of the marks awarded, as illustrated by:

- *melting = condensaition*

The correct response, '*melting'*, has been qualified by a reference to '*condensation'*, which negates the correct answer. This example poses a potentially serious problem for free text analysis. While the characteristics of the set of creditworthy responses may be increased iteratively, algorithms for recognising nullifying incorrect science may approach the infinite. Approaches to this problem are only partially implemented in the current system.

Omission Of A Mark Scheme Template

The third source of error is due to the limitations of the current CAA mark scheme format. This prevents inclusion of some correct answers in the computerised mark scheme. For example, the following response was not covered by the moderated mark scheme:

- *if you pull it back 5 N it will go ferther than 1 N*

This response is to the question, **What is the size of Kerry's push?** The answer given is very specific and includes a comparison of numeric values. It is infeasible to account for every possible direct numerical comparison which may arise, and for this reason it was not included in the computerised mark scheme. Problems such as this highlight the potential advantage to be gained by incorporating an expression evaluator into the sentence analyser. For example, for a mark scheme answer such as "Award a mark for any answer

over 40", it is obviously impractical to generate templates for each possible answer. Equally, simple numerical evaluation will not suffice if the response may be free-text. The real requirement is to include numerical expression evaluation as part of the free-text analysis process.

Quantifying System Errors

Figure 4 and **Figure 5** detail the percentage of each error category for the unmoderated and the moderated mark schemes respectively.

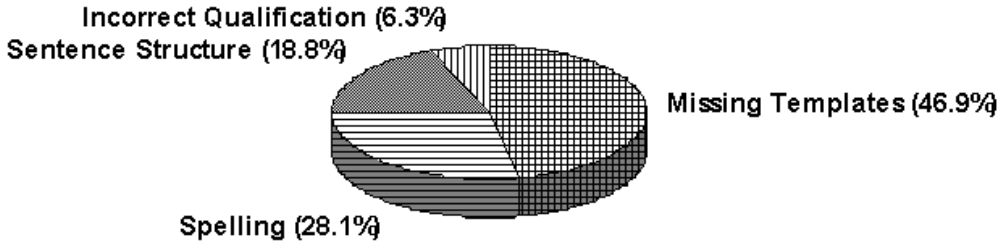

Figure 4. Analysis of errors with unmoderated mark scheme (32 responses incorrectly marked).

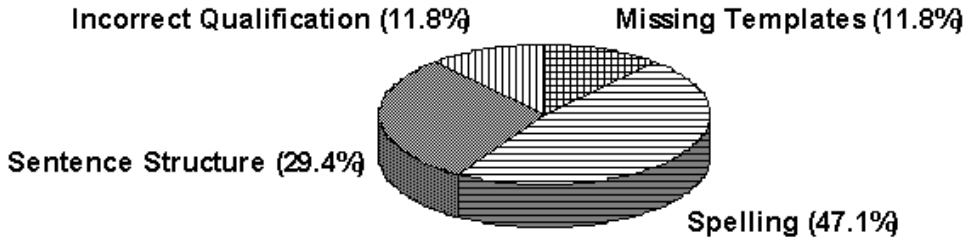

Figure 5. Analysis of errors with moderated mark scheme (17 responses incorrectly marked)

Conclusions on System Errors

The analysis of the errors encountered in the moderation experiment leads us to the following conclusions.

- With unmoderated mark schemes, the system generates a number of marking errors when faced with unexpected but allowable responses, synonyms, or phraseology.

- The system, after moderation, is able to provide high marking accuracy for items requiring word generation, sentence generation, and short explanatory sentence generation. This is true even with

the high incidence of poor spelling, syntax, and semantics evident in the student responses.

- The system performs less well on the item requiring generation of a description of a pattern in data. This is directly attributable to limitations in the current sentence analyser. More depth and detail is required in identifying the major syntactic relationships within the free-text responses. In particular, the current version proceeds on a clause-by-clause basis, but relationships between individual clauses are not maintained. For analyses of more complex free-text responses, this approach is fragile.

- Other parts of the system which will benefit from continued development are spell checking and semantic error checking. However the general approach in these areas is believed to be sound.

- The problem of incorrect qualification of correct answers is perhaps the most challenging. Other kinds of error will naturally be reduced by continued improvements in the sentence analyser, spell checker, and semantic processor. However the problem of incorrect qualification requires some innovative thinking.

The overall conclusion from a systems viewpoint is that the template-based approach, augmented by custom spelling and semantic processing modules of the type described here, can, with further development, lead to accurate and robust computerised marking of free-text responses across a range of item types and complexities.

Insights from an Educational Perspective

From the educational perspective, an efficient and reliable CAA implemented with large national cohorts is an attractive proposition. As such, discrepancies in human and computer marking judgements need to be considered from the perspective of maximising the benefits of each. CAA of free text offers the potential of speed and consistency of marking decisions; human marking adds professional judgement and inference which gives credit to badly expressed understanding. To characterise the two as ‘objective’ versus ‘subjective’ would be to misrepresent each. While the majority of responses are judged unequivocally correct or incorrect by both forms of marking, it is the ambiguous borderline responses that pose problems to both. The prospect of test papers and mark schemes accommodating the needs of CAA must be given serious consideration. What constitutes a ‘well-structured’ question, response format and mark-scheme as far as computerised marking is concerned? A consideration of this question could be a constructive and informative issue for test-item development more generally.

An alternative or complementary strategy to making software clever enough to seek and recognise divergent responses is to constrain (or gently corral) pupils in the intended direction. Such constraints, applied without sensitivity to the impact on the nature of the tests and the quality of student

performance, will rightly provoke protest from educators. On the other hand, 'test literacy' impacts on performance. Acceptable and unacceptable response procedures need to be made explicit, familiar and accessible to respondents, whatever the form of test.

Conclusions and Future Work

This paper has described a software system which employs natural language techniques to mark open-ended (free-text) responses. CAA procedures using the system are currently being developed at a number of higher education establishments, including Brunel University where an online Java test for first year engineering students is being developed. The paper has presented a brief quantitative and qualitative study of the performance of the system in marking a range of free-text responses in one of the most demanding domains: statutory national curriculum assessment of science for pupils at age 11.

From a systems viewpoint, we conclude that the template-based approach, augmented by appropriate spelling and semantic processing modules, can in due course lead to accurate and robust computerised marking of free-text responses across a range of item types and complexities.

From the educational perspective, perhaps the question is: do we want human markers to become more machine-like, or computers to become more human? Alternatively, can we use our analysis of the strengths and weakness of each to help us to define an improved middle ground of marking reliability and efficiency?

The experience of supporting students' use of OCR and ICR assessment technologies confirms that it is possible to guide students towards consistent use of response conventions. To what extent are such options open to the CAA of free text?

Another possibility for taking our work forward is to look closely at the human processing involved in implementing mark scheme criteria against student responses. Two positive outcomes are possible from such an enquiry: a modification of the manner in which mark schemes are constructed which will improve the reliability of human and computer marking; and secondly, some further clues as to how a CAA system may more closely emulate human markers.

References

Brill, E., Mooney, R.J. (1997) Overview of Empirical Natural Language Processing, AI Magazine, Vol 18, Part 4, pp 13-24, 1997.

Burstein, J., Leacock, C., Swartz, R., (2001) Automated Evaluation Of Essays And Short Answers. Fifth International Computer Assisted Assessment Conference Loughborough University 2nd and 3rd July 2001.

Callear, D., Jerrams-Smith, J., Soh, V. (2001) CAA of Short Non-MCQ Answers. Fifth International Computer Assisted Assessment Conference Loughborough University 2nd and 3rd July 2001.

Cowie, J., Lehnert, W.G. (1996). Information Extraction. In Communications of the ACM vol. 39 (1), pp. 80-91.

ExamOnline (2002)
http://www.examonline.co.uk

Glasgow, B., Mandell, A., Binney, D., Lila, G., Fisher, D., (1997). ,MITA : An Information Extraction Approach to Analysis of Free-Form Text in Life Insurance Applications. Innovative Applications of Artificial Intelligence, Providence, RI, USA, July 27-31, 1997.

Intelligent Assessment Technologies, (2002).
http://www.intelligentassessment.com

Landauer, T. K., Dumais, S.T. (1997) A Solution to Plato's Problem : The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. Psychological Review, vol. 25, pp 259-284.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.

Latent Semantic Analysis & CU Boulder
http://lsa.colorado.edu/

Mani, I, Maybury, T. (Ed.) (1999). Advances in Automatic Text Summarization, MIT Press, 1999.

Wenzel, C. (1997) Supporting Information Extraction from Printed Documents by Lexico-Semantic Pattern Matching. Proceedings of the International Conference on Document Analysis and Recognition, Los Alamitos, CA, Vol 2, pp 732 – 735, 1997.

Whittingdon, D., Hunt, H. (1999). Approaches to the Computerised Assessment of Free-Text Responses. Third International Computer Assisted Assessment Conference Loughborough University June 1999.