Green, S., Johnson, M., O'Donovan, N. and Sutton, P. (2003). *Changes in key stage two writing from 1995 to 2002*. Paper presented at the United Kingdom Reading Association Conference, University of Cambridge, UK, July 2003.

Heppell, S. (2003). Assessment and new technology: New straitjackets or new opportunities. In C. Richardson, (Ed.), *Whither Assessment?* London: Qualifications and Curriculum Authority. Retrieved August 31 2005 from http://www.qca.org.uk/downloads/combined_whither_assessment.pdf

Kennedy, A. (1992). The spatial coding hypothesis. In K. Rayner (Ed.), *Eye movements and visual cognition*. New York: Springer-Verlag.

Laming, D. (2004). *Human judgment: The eye of the beholder*. London: Thomson.

Milanovic, M., Saville, N. and Shuhong, S. (1996). A study of the decision-making behaviours of composition markers. In M. Milanovic and N. Saville (Eds.), *Performance testing, cognition and assessment: Selected papers from the 15th Language Testing Research Colloquium, Cambridge*. Studies in Language Testing 3. Cambridge, UK: Cambridge University Press/UCLES.

O'Hara, K. (1996). *Towards a typology of reading goals*. (Rank Xerox Research Centre Affordances of Paper Project Technical Report EPC-1996-107). Cambridge, UK: Rank Xerox Research Centre.

O'Hara, K. and Sellen, A. (1997). A comparison of reading paper and online documents. *Proceedings of the conference on human factors in computing systems (CHI '97)*, 335–342. New York: Association for Computing Machinery.

Piolat, A., Roussey, J-Y. and Thunin, O. (1997). Effects of screen presentation on text reading and revising. *International Journal of Human-Computer Studies*, **47**, 565–89.

Pommerich, M. (2004). Developing computerized versions of paper-and-pencil tests: Mode effects for passage-based texts. *The Journal of Technology, Learning and Assessment*, **2**, 6. Retrieved August 31, 2005 from http://www.bc.edu/research/intasc/jtla/journal/pdf/v2n6_jtla.pdf

Price, B. and Petre, M. (1997). *Teaching programming through paperless assignments: An empirical evaluation of instructor feedback*. Centre for Informatics Education Research, Computing Department, Open University, UK. Retrieved April 20, 2005 from http://Mcs.open.ac.uk/computing papers/mzx/teaching.doc

Ragat, P. and Hevey, D. (1995). *Sufficiency of evidence: The development of guidance for Assessors and Verifiers*. (Research and Development Report No. 32), Sheffield, UK: DFEE.

Reinking, D., Labbo, L. D. & McKenna, M. C. (2000). From assimilation to accommodation: A developmental framework for integrating digital technologies into literacy research and instruction. *Journal of Research in Reading*, **23**, 2, 110–22.

Royal-Dawson, L. (2003). *Electronic marking with ETS software*. AQA Research Committee. Paper RC/219. In D. Fowles and C. Adams (2005), *How does assessment differ when e-marking replaces paper-based marking?* Paper presented at the IAEA Conference Abuja, Nigeria. Retrieved February 5, 2006 from www.iaea.info/abstract_files/paper_051218101528.doc

Salmon, G. (2004). *E-moderating*. The key to teaching and learning on-line. London, UK: Routledge Falmer.

Sanderson, P. J. (2001). Language and differentiation in examining at A level. PhD thesis, University of Leeds.

Sellen, A. and Harper, R. (2005). Personal correspondence.

Sellen, A. and Harper, R. (2002). *The myth of the paperless office*. Cambridge, MA: MIT Press.

Sturman, L. and Kispal, A. (2003). *To e or not to e? A comparison of electronic marking and paper-based marking*. Paper presented at the 29th International Association for Educational Assessment Annual Conference, Manchester, UK, October 2003.

Whetton, C. and Newton, P. (2002,). *An evaluation of on-line marking*. Paper presented at the 28th International Association for Educational Assessment Conference, Hong Kong SAR, China, September 2002.

Wolf, A. (1995). *Competence-based assessment*. Buckingham, UK: Open University Press.

The Cambridge Assessment/Oxford University automatic marking system: Does it work?

Nicholas Raikes Research Division

In the first issue of *Research Matters*, Sukkarieh *et al*. (2005) introduced our work investigating the automatic marking of short, free text answers to examination questions. In this article I give details and results of an evaluation of the final prototype automatic marking system that was developed.

Introduction

Background

Cambridge Assessment funded a three year research project that investigated the application of computational linguistics techniques to the automatic marking of short, free text answers to examination questions. The research, conducted at Oxford University by Professor Stephen G. Pulman and Dr. Jana Z. Sukkarieh, focussed on GCSE Biology as a suitable context since the Biology question papers contained large numbers of questions requiring short, factual, written answers.

The researchers took two broad approaches to automatic marking. The first approach involved writing by hand what were, loosely speaking, machine marking schemes for the items to be automatically marked. This approach is referred to as the 'Information Extraction' approach. The second approach – dubbed the 'Machine Learning' approach – involved trying various machine learning techniques to, again loosely speaking, learn the marking scheme from a sample of human marked answers. A hybrid approach using semi-automatic methods to produce the machine marking scheme was also investigated.

Much useful research with some promising results was done in relation to the machine learning and hybrid approaches, offering the prospect of reducing the amount of specialised work required to set up new items. For details, see Pulman and Sukkarieh (2005). A complete prototype marking system was developed using Information Extraction techniques, and it is this system that is the focus of the evaluation reported in this article.

How the system works

We gave information about the system in our last *Research Matters* article (Sukkarieh *et al*., 2005). In essence the system matches answers to be marked against pre-written patterns to extract pertinent information previously judged by human examiners to warrant the award or forfeiture of a mark. The patterns can include syntactic information to specify parts of speech, verb groups and noun phrases, and essentially a pattern covers the synonyms for each pertinent piece of information.

Patterns are written by hand and based on the marking scheme used by human examiners, together with a sample of 200 human-marked answers. The sample answers are annotated by the human examiners to indicate precisely the part(s) of each answer which gained or forfeited marks – this annotation is done to minimise the need for the person writing the patterns to make these judgements.

Method

Two multi-part Biology questions were chosen from a 2003 GCSE Double Science examination. They were chosen because:

- They were common to both Foundation and Higher tiers and therefore could be used with the widest range of candidates.
- Every sub-part required a short, factual, textual, answer. This means that the whole questions could be used, providing a cohesive mini computer based test that can subsequently be given to volunteers to demonstrate or further research the system.
- There were eight 1-mark items (sub-parts) and five 2-mark items and so the questions covered the range for which this automatic marking technique might be suitable.

In the real GCSE examination, candidates answered on paper. Since automatic marking requires machine readable text, a random sample of 748 paper scripts was obtained and the relevant answers – excluding the totally blanks! – keyed into a computer file. Two hundred of these answers to each item were used to help with writing the patterns, while the remaining answers were held back for use in the marking trial.

All answers for each item had been live marked (i.e. marked for real in the GCSE examination) by human examiners; we had the resulting item-level marks keyed into a database. For the evaluation we recruited two senior examiners – both of whom had led teams of examiners in the live marking – to independently mark the transcribed answers a further two times. These examiners marked hard copies of the transcriptions; their marks were also keyed into the database. We therefore had three human examiner marks for each non-blank answer: one live[1] and one from each of the two Team Leaders recruited for this evaluation. In addition to marking the answers, we asked the two Team Leaders to annotate 200 of the answers to each item to show, by highlighting and labelling, precisely which parts of the answer matched each numbered point in the examiners' written marking scheme. The two hundred answers to each item were chosen according to their live marks as follows. For the 1-mark items, random samples of 50 0-mark answers and 150 1-mark answers were drawn. For the 2-mark items, the proportions were 50 0-mark, 75 1-mark and 75 2-mark. Where there was a shortage of higher mark answers, half of those available were used in the training data (the balance made up of lower scoring answers), and half were retained for use in trial marking.

The researchers in Oxford were provided with the following material to help them write the patterns:

- copies of the question paper and of the examiners' written marking scheme;
- both sets of the 200 annotated sample answers for each item;
- all three sets of marks for these answers (one live mark and two evaluation marks).

The Oxford researchers were **not** provided with any details of the remaining answers used for trial marking. These answers are referred to as the 'unseen answers'.

Oxford's patterns were sent to Cambridge Assessment and compiled into the automatic marking system running on a Cambridge Assessment server for trialling.

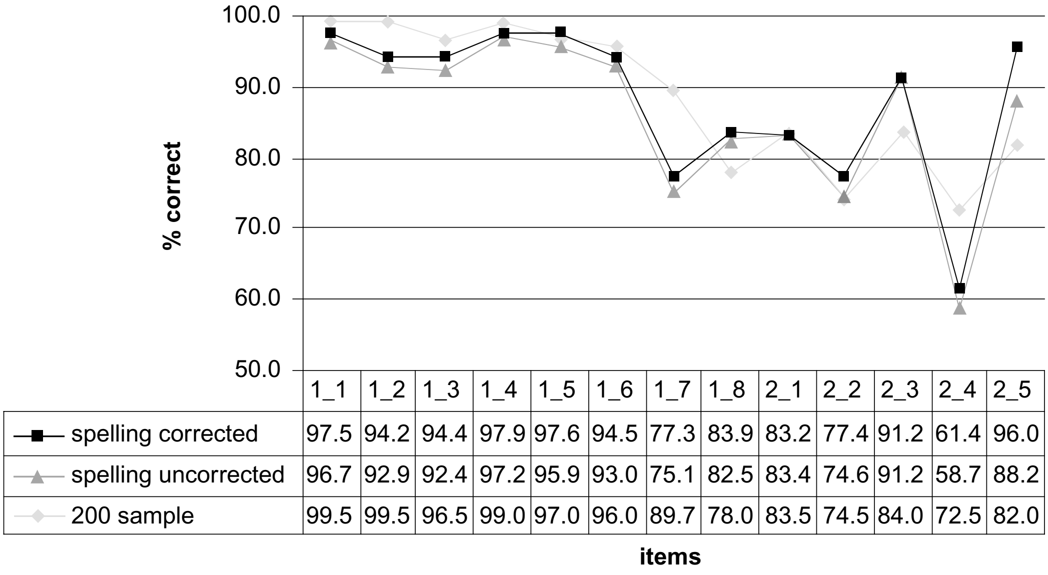
The unseen answers were marked automatically using the patterns developed by Oxford. The output included details of words that were unrecognised by the system, generally due to spelling errors made by candidates, together with suggested alternatives. If the test had been taken by candidates on computer, the system would have used this output to provide a spelling checking facility to candidates. It was therefore decided to correct the errors and run this corrected data through the marking engine. In this way the best and worst case scenarios for spelling mistakes could be compared.

In addition to the Oxford patterns, Cambridge Assessment also commissioned a temporary worker previously totally unacquainted with the project to write patterns for three of the items. This person had a background in psychology and computer science, but no experience of computational linguistics. He relied primarily on Oxford's documentation to train himself in how to write the patterns, and when writing his patterns had access to exactly the same material as Oxford had had. These patterns – the Cambridge Assessment patterns – were compiled into the system and the unseen answers were marked using them. In this way 'a proof of concept' investigation was conducted into the feasibility of transferring pattern-writing skills to persons not involved in developing the system – a key requirement for commercial use of the system.

Results

We will present and comment on the results in the following order. First we report on the correctness of the automatic marks using the Oxford patterns. Next we report inter-marker agreement levels, comparing each of the four markers (one automatic and three human) with each other; again, the Oxford patterns were used by the automatic marker. Finally, we report similar results for when the Cambridge Assessment patterns were used, and compare them with the previous Oxford results.

[1] The live marks were not all due to a single examiner, since sample scripts were chosen at random.



	1_1	1_2	1_3	1_4	1_5	1_6	1_7	1_8	2_1	2_2	2_3	2_4	2_5
■ spelling corrected	97.5	94.2	94.4	97.9	97.6	94.5	77.3	83.9	83.2	77.4	91.2	61.4	96.0
▲ spelling uncorrected	96.7	92.9	92.4	97.2	95.9	93.0	75.1	82.5	83.4	74.6	91.2	58.7	88.2
◆ 200 sample	99.5	99.5	96.5	99.0	97.0	96.0	89.7	78.0	83.5	74.5	84.0	72.5	82.0

Figure 1: Percentages of answers marked correctly by the automatic system

The automatic marks' correctness: Oxford patterns

Figure 1 shows, for the eight 1-mark items (1_1 to 1_8) and five 2-mark items (2_1 to 2_5):

- **Spelling corrected:** The percentage[2] of the **unseen** answers marked correctly by the system, after the answers were corrected for spelling.
- **Spelling uncorrected:** The percentage of the **unseen** answers marked correctly by the system, with no spelling corrections.
- **200 Sample:** The percentage of the **200 sample answers** to each item, used by Oxford for pattern writing, that the system marked correctly[3].

Table 1 presents the percentage of answers-per-item correctly marked by the system, averaged separately across the 1-mark and 2-mark items. The 1-mark average was also calculated with the two outlying items – 1_7 and 1_8 – excluded.

Key points concerning Figure 1 and Table 1 are:

- The system's correctness levels were high – above 90% – for six of the eight 1-mark items and two of the five 2-mark items.
- Spelling correction improves the performance of the system, though not by much.
- The figures for the 200 sample answers are very similar to those for the full marking trial sample, indicating that generally the 200 answers chosen to help with pattern writing were sufficient to cover the range of answers in the larger sample.

It might be supposed that the system would get the marks right for 100% of the 200 sample answers, since the Oxford researchers wrote the patterns with these answers in mind. However, the patterns are designed to extract key pertinent information from the answers and are not simple copies of the answers themselves; the challenge is to write patterns that are general enough to match paraphrases of the intended content but not so general as to also match unintended content.

Table 1: Average percentage of answers-per-item correctly marked automatically

Item tariff	*Average % correct per item*			
	n	*Corrected*	*Uncorrected*	*200 sample*
1 mark items	3519	92.1	90.7	94.4
1 mark items without 1_7 & 1_8	2744	96.0	94.7	97.9
2 mark items	1793	81.8	79.2	79.3

There is not an obvious reason why the system performed less well on some items than others. It does not seem to depend simply on the facility (difficulty) of the item or the number of alternatives given in the examiners' written marking scheme – item 1_6 (94.5% correct) contained more alternatives than item 1_7 (77.3% correct), whereas item 1_1 (97.5% correct) contained fewer alternatives than 1_7. However, it would seem reasonable to suppose from the results given in Figure 1 that if marking accuracy approaches 100% for the 200 sample answers, the accuracy will also be high for the larger sample of answers. In this way, the sample answers used for writing the patterns might be used to screen for items likely to be unsuitable for automatic marking.

Tables 2 and 3 indicate the correctness of the automatic marks for the spelling-corrected answers, broken down by the correct mark. The percentages relate to the column totals, i.e. in Table 2, 94.1% of the 784 answers with a correct mark of 0 were correctly given a mark of 0 by the automatic system. Key points to note are that:

- The automatic marker is a little more likely to under-mark (give too low a mark) than over-mark. This implies that for these data the system is more likely to fail to credit an unusual right answer than it is to be 'fooled' by a wrong answer.
- For the 2-mark items, very few responses received an automatic mark that was wrong by the full two marks.

[2] Strictly, this is the percentage of the unseen answers for which we have undisputed human marks that were marked correctly by the system. See footnote 3.

[3] Two approaches were used for determining the 'correct' marks. For the 200 sample answers used by Oxford, the Oxford researchers made their own judgement of what the correct mark should be. For the unseen answers, we considered that when all three human marks agreed, the mark was definitive; we therefore only used these definitively marked answers for the analyses involving 'correct' marks.

Table 2: % correct by mark – (1 mark items)

		Correct mark	
		0	*1*
Auto	0	**94.1**	7.9
	1	5.9	**92.1**
	All	100.0	100.0
	n	784	2735

Table 3: % correct by mark – (2 mark items)

		Correct mark		
		0	*1*	*2*
Auto	0	**89.5**	17.7	2.3
	1	10.4	**77.0**	18.0
	2	0.1	5.3	**79.7**
	All	100.0	100.0	100.0
	n	733	666	394

Inter-marker agreement: Oxford patterns

No matter how careful or well trained they are, human markers inevitably make occasional mistakes[4]. Moreover, some candidates write ambiguous answers that leave room for markers to differ in their expert opinion of whether an answer fulfils the requirements for a particular mark to be awarded – indeed unbiased marking requires that for an answer on the cusp of being worth a mark, half of a hypothetical population of all markers would award the mark and half would withhold it. Of the 6,357 unseen answers, at least two human markers differed in the marks they awarded to 1,047 of them. Table 4 gives pairwise agreement percentages between every pair of markers. For example, the top left cell shows that on average the live and automatic marks exactly agreed for 87.6% of the answers to one-mark items.

Key points to note are:

- 1-mark items:
  - When the two items previously identified as outliers (1_7 and 1_8) are ignored, the average exact agreement rates for the auto-human pairs are broadly similar, though a little lower, than for the human-human pairs;
  - All the figures are higher when the outlier items are removed, implying that the human markers were also more likely to disagree on these items – though the differences are smaller for the all-human comparisons than for the auto-human comparisons.
- For the 2-mark items, the average exact agreement rates for the auto-human pairs were lower than for the human-human pairs, though none fall below 70%.The averages reported in Table 4 mask considerable variation by item, however. For example, the auto-marker agrees more often with the live mark for item 2_5 (91.8%) than do either of the human markers (85.1% and 89.7% respectively). There is no obvious reason for this.

Table 4: Pairwise agreement between markers

Item tariff	*Average % exact agreement per item*					
	Live v auto	*Live v exr1*	*Live v exr2*	*Auto v exr1*	*Auto v exr2*	*Ext1 v exr2*
1 mark items	87.6	89.5	93.5	87.6	87.9	92.7
1 mark items without 1_7 and 1_8	91.6	91.8	95.1	92.3	92.5	94.8
2-mark items	72.7	80.1	81.7	73.7	76.6	87.1

Cambridge Assessment's patterns

Patterns for three items were written by the temporary worker recruited by Cambridge Assessment. The accuracy of the automatic marker when using these patterns, compared with the Oxford patterns, may be found in Table 5. Table 6 gives the inter-marker agreement figures. The results for the two sets of patterns are very similar.

Table 5: Percentage of answers correctly marked automatically using Oxford & Cambridge Assessment patterns

Item	*% correct (uncorrected spellings)*		
	Oxford	*Cambridge Assessment*	*n*
1_3	92.4	94.2	449
1_4	97.2	95.3	465
2_3	91.2	93.4	385

Conclusion

We evaluated the automatic marker using eight 1-mark items and five 2-mark items. The items were all taken from a GCSE Biology question paper, and answers from a sample of paper scripts were keyed into a computer file for automatic marking.

The automatic marker marked all but two of the 1-mark items with a high degree of correctness – more than 90% of the answers for which we had a definitive (undisputed) human mark were marked correctly.

[4] Marking reliability is controlled through quality control checks – both during marking and also prior to results issue, when checks may be targeted on candidates close to grade boundaries – and by virtue of the fact that many mark-decisions contribute to a candidate's final result.

Table 6: Inter-marker agreement, Oxford and Cambridge Assessment patterns

Item	*% exact agreement (uncorrected spellings)*					
	Oxford			*Cambridge Assessment*		
	Live v auto	*Auto v exr1*	*Auto v exr2*	*Live v auto*	*Auto v exr1*	*Ext1 v exr2*
q2biii	91.1%	91.1%	89.4%	93.2%	92.8%	91.1%
q4a_fur	91.3%	93.6%	90.6%	89.2%	92.3%	88.4%
q2cii	71.2%	82.9%	85.9%	72.4%	88.2%	87.9%

Agreement levels between the automatic marker and human markers were also broadly similar – for these items – to those found between human markers. We could find no simple explanation for why the remaining two 1-mark items were marked less well by the system – suitability for automatic marking does not appear to depend simply on item difficulty or the number of alternatives given in the examiners' written marking scheme. However, the 200 sample answers used for pattern-writing appear likely to be sufficient for screening 1-mark items for automatic marking. The system was generally less often correct, and there were bigger differences between auto-human and human-human agreement levels, for 2-mark items.

Patterns were written for three of the items by a temporary worker recruited by Cambridge Assessment. This worker was highly qualified in psychology and computing, but had had no previous exposure to the project or computational linguistics. The correctness and inter-marker agreement levels were similar for both sets of patterns, implying that it is possible to transfer pattern-writing skills from the developers to new staff. This is an important step for the commercialisation of the system.

We conclude that automatic marking is promising for 1-mark items requiring a short, textual response. More work is needed to see how the findings generalise to subjects and qualifications other than GCSE Biology, and to investigate why some items are less suitable for automatic marking using this system than others.

References

Pulman, S. G. and Sukkarieh, J. Z. (2005). Automatic short answer marking. In *Association for Computational Linguistics: Proceedings of second workshop building educational applications using NLP*, Ann Arbor, Michigan, 9–16.

Sukkarieh, J. Z., Pulman, S. P. and Raikes, N. (2005). Automatic marking of short, free text responses. *Research Matters: A Cambridge Assessment Publication*, **1**, 19–22. http://www.cambridgeassessment.org.uk/research/

EXAMINATIONS RESEARCH

The curious case of the disappearing mathematicians

John F. Bell and Joanne Emery Research Division

It is not unusual for claims to be made that some aspect of education is getting worse. Mathematics is often cited as a particular area of concern. There have been a number of reports about this issue including Roberts (2002), Smith (2004) and the UK Mathematics Foundation (2005). The declining number of A-level mathematicians is often cited as a particular concern, for example, in the *Times Educational Supplement* Gardiner (2006) wrote

> *'the number of A-level Mathematics students has slumped from 85,000 in 1989 to 66,000 in 2001, and (thanks to the misconceived Curriculum 2000 reforms) to just 52,000 in 2004.'*

A simple calculation would suggest that there has been a fall in numbers of the order of 33,000 students taking A-level mathematics, that is, a 39% decline. However, the interpretation of educational statistics is not a predictable 'one-piece jigsaw' but is instead a fairly simple multi-step problem. The first step is to identify the source of the statistics and check that they are comparable. It is not surprising or unreasonable that the source is not given in a newspaper story. However, an inspection of the available statistics would suggest that no identical definition of A-level mathematics students could simultaneously give a number as high as 85,000 in 1989 and as low as 52,000 in 2004. To investigate this problem, we decided to use the Summer Inter-board Statistics which have been compiled for A-level since 1990 in their present form (some earlier figures were obtained for 1989 but these may be a slight undercount).

After identifying a comparable source of statistics, the next issue is to consider the definition of A-level mathematics students. It is reasonable to assume that from the point of view of Higher Education and employment this should be based on the number with passing grades (A-E). This is important because in 1989 30% failed A-level mathematics and this was only 4% in 2004. A change in failure rates is unsurprising given that the introduction of modular A-levels led to candidates dropping mathematics rather than completing the course, obtaining a U and appearing in the statistics. Another relevant factor is the number of 17-year-olds in the population. This varied considerably over the period in