AUTOMATED ESSAY MARKING FOR CONTENT
~ DOES IT WORK?

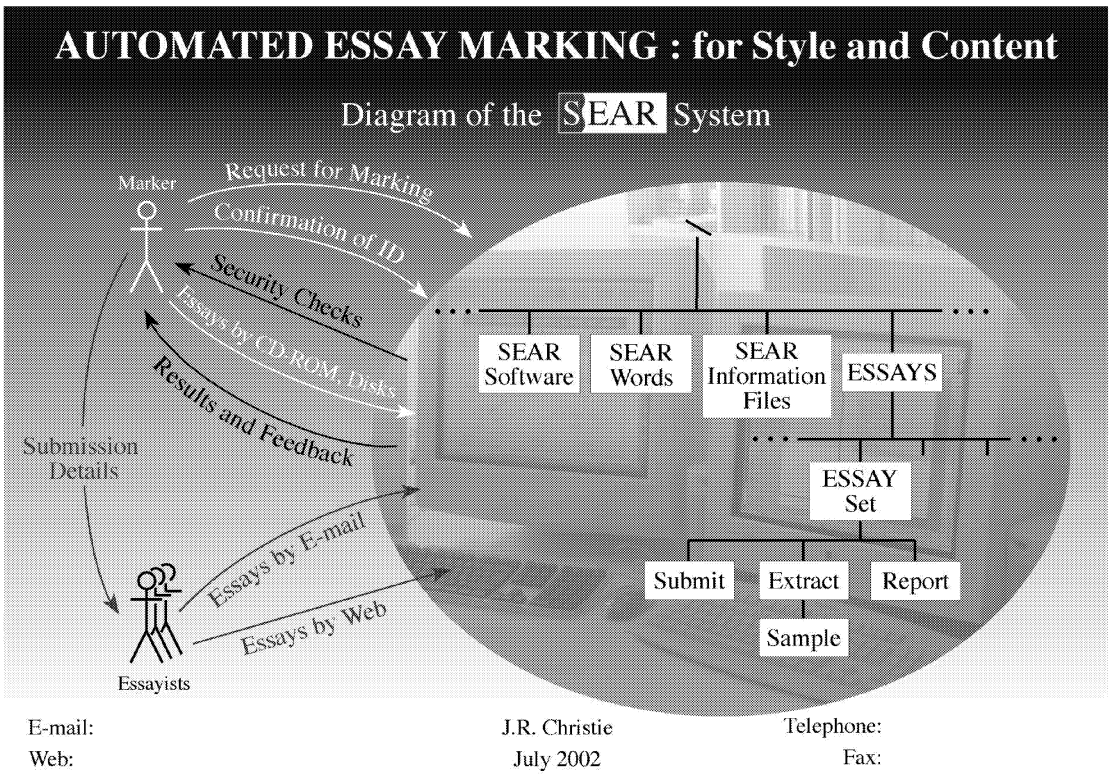
James Christie

Automated Essay Marking for Content ~ Does it Work?

James R Christie
james.christie@btinternet.com http://www.jkp.christie.btinternet.co.uk

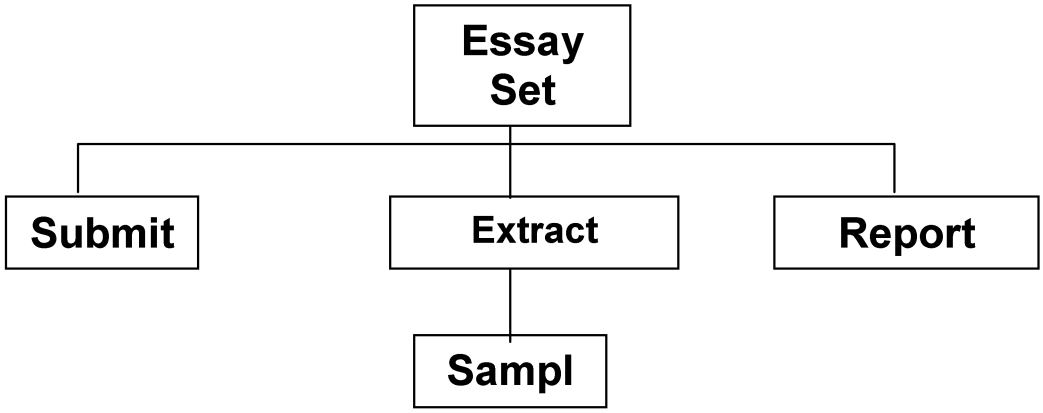
Abstract
The answer is "Yes, but in a manner similar to that of the first powered flights!" Following a very brief overview of the author's **SEAR** system, the *preliminary* results generated by the software will be revealed and then discussed. So far, SEAR marking of essay content is successful for the lowest cognitive level, knowledge. As the software development continues then the marking performance and cognitive level will increase as the software makes gains in sophistication. The potential for added value arising from using SEAR in terms of feedback to the essayist, feedback to the examiner and the detection of possible instances of plagiarism will be indicated. The author will demonstrate how style essay marking could be achieved with SEAR. Possible candidate criteria for the acceptability of automated essay marking will be reviewed. Routes for future research work will be outlined. Some of which may raise a few surprises.

Paper
The author created a software-based system for automated essay marking called **SEAR**. This paper reports on the earlier work published by this author [Christie 1998, 1999, 2003]. This system marks word-processed essays for content and has the capability to mark for **style** as well. SEAR was developed to support the author's PhD research. An overview of the software system is shown below.



SEAR stands for four stages which are, in turn, **S**chema, **E**xtract, **A**ssess and **R**eport modelled on the author's approach to the use of assessment instruments. The Schema stage occurs before the assessment due date [or deadline] and the remaining three stages all occur after the assessment due date.

The **Schema stage** is concerned with the establishment of the marker's identity and assessment details, especially the submission of the essays. Should content marking be required then the content marking schema and a model essay must be prepared at this stage by the examiner. For each assessment a specific essay set directory is created. For each assessment the essays are held in a submit folder within the essay set directory.



The **Extract stage** is a piece of software that pre-processes the essays from the submit folder into the extract folder from which the actual marking is performed. This one extract process is the same for both style marking and content marking. Currently the extract process only works for versions of the Microsoft ™® Word package. Should another word-processing package be used to create the essays then another package specific extract software will have to be created.

The **Assess stage** is where the actual assessment occurs. Marking for style uses a different piece of software from marking for content. Should an examiner seek both marking for style and marking for content then both pieces of assessment software will have to be run, one after the another [sequence order is irrelevant]. Both pieces of assessment software operate only on the extracted essays. The results from the assessment are stored in the essay set's results folder.

Assessing style is still to be field tested ~ the lack of style marked essays sets is the problem here. The mark for style would be produced from a mathematical calculation [a weighted linear function] based on a [common] set of metrics. The weightings are produced from processing a human marked sample taken from the main essay set.

Assessing content is achieved by the matching of the content schema with the essays. This marking is based on key word(s) and the relationship(s) between these key word(s). The image below serves to show what is required from the marker in terms of producing a content schema. This schema was produced by using a very simple text editor. The marker produced content schema is converted into a computer file that is then used in the marking for essay content.

```
Sample of a [short] Schema for Robert Gordon, founder of the Robert Gordon
University. Essay is marked out of 25, while the sum for individual items is 35
~ not a mistake!
**********************************************************************
Mark   Item
 0       Robert Gordon
 3               born 1668 Castlegate Aberdeen
 3       inherited 1680 £1,000
 5               graduated 1689 Marischal College Aberdeen
 3               retired 1720 Aberdeen
 1               died 1731
[max:15]

 1               Father
 2                       Arthur Gordon
 2                               Edinburgh advocate
[max:5]

 1               Mother
 2                       Isabella Gordon
 2                       nee Isabella Menzies
[max:5]

 1               Grandfather
 2                       Robert Gordon
 1                       Cartographer
 3                               Blaeu's Atlas 1654
[max:7]

 1               Trader
 2                       Danzig Baltic
[max:3]

[max:25]
```

The **Report Stage** is not a ‘true stage’ as it deals with the results from the assess stage. Results may be viewed as-is by the use of any simple text editor. Alternatively, these results may be imported into any commonplace spreadsheet package for further processing. The results folder holds the

results produced by the assess stage and also holds performance information on the extract process and assessment processes.

So what about the results? The spread of performance of the computer marking is shown in the table below.

Essay Set	First V Second Markers	1st Human V SEAR
A	0.810** / 0.740**	0.404** / 0.376**
B	0.704** / 0.700**	0.594** / 0.596**
C	0.164 / 0.277	0.302* / 0.394**

Pearson Correlation Coefficient / Spearman Correlation Coefficient
Significance ** = 0.01 * = 0.05

Some explanation is necessary before any analysis of the table is made. The first marker [who was also the question setter] and the second marker are all experienced subject expert lecturers and the computer marking was done by SEAR. The author assumed that the first marker is always absolutely correct, that is marking is 100% accurate. In reality the safety in making this assumption may be false.

The feature common to all essay sets in the table given above is that the assessments were all set to elicit facts or knowledge, or to explain a process. In short, all the essays covered by this one table could be categorised as operating at the lowest level of Bloom's Taxonomy, namely "Knowledge". The author has only had access to this type of essay. Therefore the performance of the SEAR system in marking essay content that represents the higher levels of Bloom's Taxonomy will remain speculative until the author has access to several such essay sets.

Overall this table shows that human marking is still statistically better than automated essay content marking. So there are no surprises here.

However the table is worth a closer inspection. For the first two rows the agreement between first and second markers and the first marker and computer marking are both statistically significant. Thus the computer marking may not be totally rejected. In the last row the computer marking is statistically better than second marker.

Net result is that SEAR can mark essays provided that the essays are biased towards facts. It is possible for the human-human marking to be poorer than human-computer marking. The author has found that the bigger the content schema the poorer the second marker and the computer marking becomes. Likewise with the marks awarded, the bigger the amount of marks awarded then the poorer the second marker and the computer marking becomes. Does this reveal something more about using essays for assessment than it does about human-computer marking? Should there be a limit to the size of the content schema and or a limit to the marks awarded?

What may cause the computer marking to be poorer? Human markers are capable of correcting errors of spelling and errors of grammar, but the computer is not that capable. Human markers are aware of alternative expressions [for example bandwidth is often shortened to BW or bw or b/w, and could be written as “band width”] whereas the computer only marks bandwidth. Human markers may give the benefit-of-the-doubt for confused expressions, the computer will easily “lose the plot”.

It is expected that when improvements are made to the content algorithm then the performance of the SEAR system should increase. Equally the more essay sets that are made available then the better the scope for the fine-tuning of the algorithms used. Again this should result in performance enhancement. Further, the greater the essay sets range across Bloom’s Taxonomy then, again, the better will become the SEAR performance. In passing note that in step with performance improvement in content marking there should be a corresponding increase in style marking performance.

Pure marking is not fully acceptable since merely handing out marks was never acceptable educational practice, except in summative assessments. Added value in essay assessments comes from providing feedback. Currently in the SEAR system the provision of feedback is only very limited in extent. Feedback to students is merely a list of those facts that the essays did not contain. Each essay has to be examined individually. Staff feedback is possible by manually examining the detailed matrix of the results across the essay set. In SEAR the content marking software has the facility to furnish the student with specific meaningful feedback comments if required, but it is not yet activated.

It would be easy to create software that would take the results matrix and produce appropriately formatted feedback to staff and to students.

Unfortunately another form of feedback is sought [but not welcome] nowadays – report on possible plagiarism. Detection of potential plagiarism is also crudely provided for by the SEAR system. By importing the detailed results into a commonplace spreadsheet and then sorting the data, like pairings will appear. Like pairings is the basis of deciding if plagiarism has occurred. The results matrix could be directly processed by software based on mathematical techniques, such as Euclidean Distance or the Horn Statistic, to identify potential like pairings. In both the mathematical techniques mentioned above, the smaller the value produced for the pairing of results then the greater is the potential likelihood that plagiarism has occurred. The author envisages a reporting format of an upper diagonally populated matrix showing the mathematical number produced for each particular pairing as a suitable method of alerting staff to potential wrong doings.

Whatever method of determining potential plagiarism, the author believes that the human marker is the better judge in deciding that any academic mis-conduct has occurred.

Similar to other areas of life on this planet there is always a focus of attention on the acceptance criteria of the new paradigm; precious little interest on the existing paradigm. There is now, and has been for some time, an interest in marking free text, essays for style and essays for content. Perhaps now is the time to establish which criteria, if any, must be present in order to establish if the performance of essay marking software is acceptable. Kaplan [Kaplan 1992, 1998] assembled the following criteria as a [starting] list of possible acceptance criteria.

Kaplan's Suggested Criteria for Automated Assessment Acceptance
Ease of creating a scoring schema
Ability to score on various mark regimes
Ease of identification of non-scoring elements
Ease of modification should scoring errors occur
Consistent and reproducible scoring
Acceptability of results to markers [and essayists]
Defensibility [of the marks awarded]
Accuracy and precision
Coachability avoidance
Cost

These ten criteria are straightforward in terms of desirability. The difficulty is in the construction of suitable objective measure(s) to support each criterion. How does one measure "ease of …"?

To this list the author suggests adding the following criteria.

- Acceptance by a range of stakeholders.
- Performance no worse than human marking.

The first of these additional criterion serves to include stakeholders such as external examiners, professional bodies, funding bodies and so on. Each of these stakeholder bodies has its own agenda for acceptance or non-acceptance of automated assessment in general and automated essay marking in particular. All the stakeholders mentioned above might view automated assessment differently. It would be understandable if professional bodies were to take the view that automated assessment might lower academic standards, while funding bodies would support automated assessment because of its cost saving potential.

The second additional criterion, namely performance, is probably the hardest one to measure. This is because manual marking is known to be susceptible to subjectivity and the very mention of computer operations generates the notion of 100% accuracy and or the notion that such marking may be easily subverted whilst being a systematic robotic operation. The author suggests one measure could be that computer marking is no worse than human marking. Would this then set a low threshold for accepting computer based marking?

Routes for the future

There are many routes that the further development of the SEAR system may take. Each route is independent of any other, but offers the opportunity to help develop other route(s). The major possible routes are ~

- Better content performance,
- Better style performance,
- Increase operational effectiveness,
- Achieving further levels of Bloom's Taxonomy for marking content, and

- Operating in non-English Latin character based language(s).

The first four are not unexpected, especially from what has been raised so far in this paper.

Better content performance could be achieved through the incremental development of the algorithm by identifying a possible performance improvement and consequently designing and testing that increment. If the increment successfully passes through these stages then that increment would be incorporated into the SEAR system. Further development in terms of maximising the use of active and passive voices, coping with spelling and grammar errors, and including non-textual elements are all obvious routes to improved marking performance.

Better style performance will follow the same route as for content. However, here the improvement would be to confirm the identity of a "common metric set" that could be applied to essay sets. Further style improvement would be to include word-processed sourced metrics such as font type, size and colour together with text enhancing features.

Increased operational effectiveness would be the development of better feedback to the essayist and to the marker. Unfortunately increased plagiarism detection would also improve operational effectiveness. Because today the volume of plagiarism is increasing it behoves markers to be very vigilant to deter such behaviour.

Achieving further levels of the Taxonomy is somewhat more problematical, with the problem being in the area of the content schema. Indeed how far can the higher levels of the Taxonomy be represented in the content schema? If the answer is, say, synthesis then what would such a schema look like. It might be that alterations to the current methodology of developing the content schema will have to be generated.

Operating in non-English language(s)

The author has yet to find anyone working in the area of essay marking, or for that matter free text marking, who is using non-English language(s) – but his search goes on. As far as very early experimentation shows, it appears that it should be possible to apply the SEAR software to non-English languages, provided that a standard computer keyboard is used to create the content schema and the essays. Here SEAR marking is applicable for both style marking and content marking.

Finally ~

Automated essay marking for essay content ~ does it [SEAR] work?
This author's answer is "Yes, but in a manner similar to that of the first powered flights!"

SEAR is not very grand, not too elegant and a bit under-powered but gives others the basis on which to develop their own ideas!

References

Christie, JR (1998) *Computer-assisted Assessment of Essays*
2nd Computer Assisted Assessment Conference, 17-18 June, 85-89

Christie, JR (1999) *Automated Essay Marking – for both Style and Content*
3rd Computer Assisted Assessment Conference, 16-17 June, 39-48

Christie, JR (2003) *Automated Marking of Essays for Content*
Automated Free Text Assessment Seminar, SCROLLA Heriot-Watt University, 30th April

Kaplan RM & BA, Wolff S E, Burstein JC, Lu C, Rock DA (1998)
Scoring Essays Automatically Using Surface Features GRE Research, August, 1-13

Kaplan, R M (1992) *Using a Trainable Pattern-Directed Computer Program to Score Natural Language Item Responses* GRE Research Report, April, 1-43