AUTOMATED ASSESSMENT OF CHILDREN'S HANDWRITTEN SENTENCE RESPONSES

Jonathan Allan, Tony Allan and Nasser Sherkat

Automated Assessment of Children's Handwritten Sentence Responses

Jonathan Allan, Tony Allan, Nasser Sherkat
School of Computing & Mathematics,
The Nottingham Trent University
Burton Street,
Nottingham,
NG1 4BU.
Tel: (+44)0115-848-2150 Fax: (+44)0115-848-6518
Email: { ja, tja, ns }@doc.ntu.ac.uk.

Abstract

This paper compares two approaches for the recognition and assessment of handwritten sentence style answers to free text response questions. The first uses a conventional approach to handwriting recognition in which a general purpose lexicon is used in an attempt to recognise all words in each of the responses. This has the advantage of producing a recognised response every time but has the disadvantage of introducing many recognition errors into the overall assessment process. Assessment is then performed by comparing the recognised response, formed from the best matching words found in each position, against a set of model answers. The second method employs a specific word assessment technique to evaluate each word in the written response only against a set of keywords derived from the model answers. Using a threshold based confidence measure the system can determine whether or not the recognition is correct at each word position. If the technique is not confident about a recognised word then it will not give a response for that word position. Assessment is then performed by comparing the confidently recognised keywords against the model answers. In both approaches there is the option of rejecting a response and passing it for manual assessment when the recognised response contains only partially correct answers. The use of a questions history is also exploited to help make the assessment more robust. Results show that the Specific Word Assessment Technique with History performs best with an overall assessment accuracy of 100% on a response yield of 33.2%. The other 66.8% of responses were automatically classified as pass to manual marking because the approach was unconfident in marking the responses. The Conventional Lexical Approach with History managed a response yield of 72.0% but with an assessment accuracy of only 41.4%.

Keywords: Automated Assessment, Handwriting Recognition, Confidence.

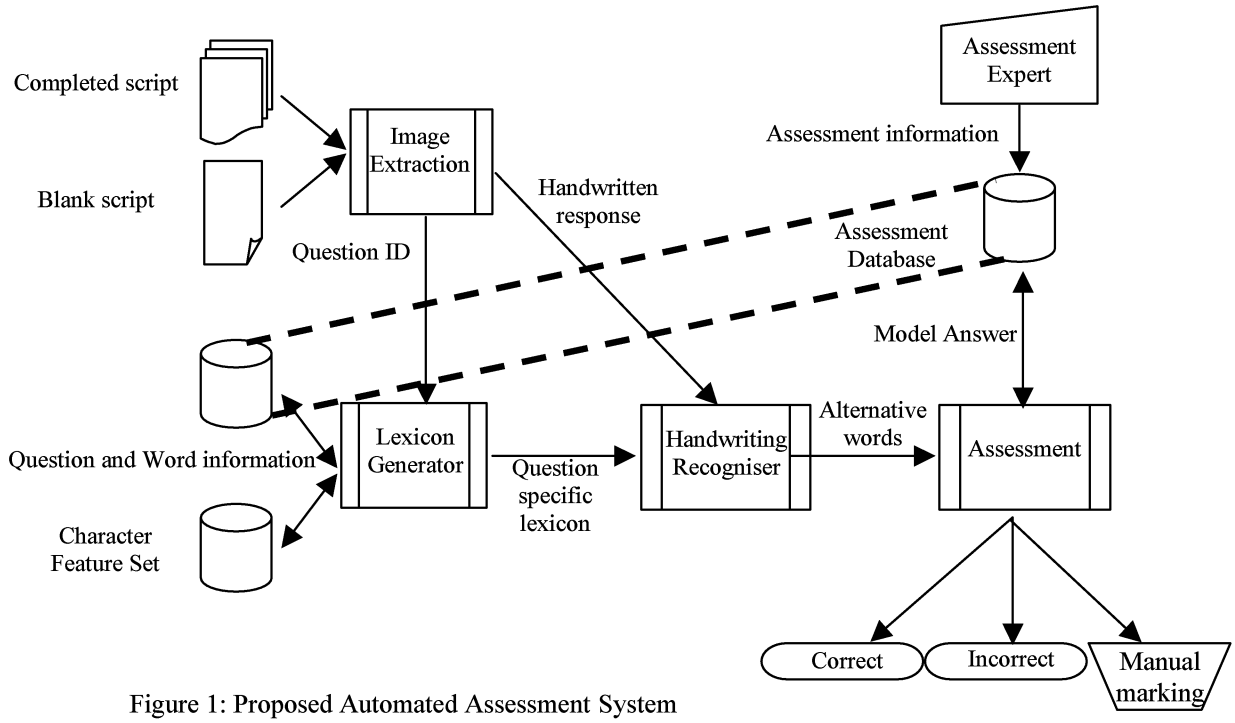


Figure 1: Proposed Automated Assessment System

Introduction

Figure 1 shows a proposed automated assessment system that could automatically assess handwritten responses. An applied system such as this can utilise the knowledge of an 'assessment expert' to produce an assessment database. This database can be then used as a reference for information such as the model answers, history (past answers) and stimulus for a specific question, once the ID of the question is found. The assessment process is dynamically linked to the assessment database as it is significantly related to the question and therefore the process has to be built around each specific question. The system itself has four main processes: Image Extraction, Lexicon Generation, Handwriting Recognition and Assessment.

A semi-automatic Image Extraction process was used in the experiments in this paper, in which 100% of the handwritten responses were processed correctly and passed for recognition. The handwriting recognition lexicons were also manually generated since it was necessary to know the specific question they represented. Therefore, this work focuses on whether or not the assessment methods can overcome the errors introduced into the overall assessment process during the handwriting recognition stage.

The recogniser used in this work is a holistic word recogniser (Evans *et al)*. Instead of segmenting a word image into characters and then trying to recognise each of the characters independently, the holistic approach works by recognising the whole word image. This takes advantage of the shape of the word and how the characters influence other characters around them. However, handwriting recognition still has many inherent difficulties that range from coping with a wide variety of hand writing styles to the complexity of recognising multi-word combinations that cause ambiguity. Hence, the

performance of the current handwriting recognition systems is still far from perfect. For general-purpose applications, it is neither desirable to limit the number of users of the system nor is it possible to know the writing styles of every user who is going to be evaluated by the system. However, constraining the scope of what can be expected within the written responses, in order to improve the recognition rates, is a possibility.

Previous work has already shown that highly accurate assessment of handwritten responses is possible if the constrained nature of the responses is taken into account (Allan *et al*, 01). There it was shown that prior knowledge of the required response can allow contextual bridging to be used to augment the basic word recognition rates in order to increase the recognition confidence; albeit at the expense of a reduction in the response yields. However, in a situation where there is only a single word in the expected response no contextual knowledge can be gained but there is still a need to improve the accuracy of the automatic assessment system. When recognising single words, the lexicon used could be highly ambiguous and therefore the resulting recognition accuracy's will be poor. In such a situation, the confidence of the automatic assessment system will be low. To overcome this problem a Specific Word Assessment Technique (SWAT) was introduced (Allan *et al*, 02) This technique was employed to automatically assess single word responses from the same perspective as a human assessor. SWAT exploits a lexicon that only accommodates the correct answer to a specific question; this takes away the latent ambiguity that is inherent in a more generalised lexicon. This paper show how SWAT can be modified to find keywords within a handwritten sentence in order that they can also be assessed.

Children's Sentence Response Assessment

The automatic assessment of a five-question exercise is to be attempted (see figure 2). The exercise formed part of the 'Progress in English 10' exam paper published by NFER-Nelson. All the questions require a sentence response, however the minimum answer can be simply a single word. Should the child give only a single word response then they will not be penalised for it and the response would be scored accordingly (i.e. a correct response to Q1 could just be '*dragon*'). In preparation for the questions the children had to read a short story (stimulus) in which the answers to the first two questions were explicitly mentioned and in which a contextual link for the last three questions could also be found.

Exercise 5: The Tunnel

Please answer these questions.

1. He was waiting so that he could watch the steam-engine come roaring out of the tunnel.

   This sentence makes the train sound like an animal.

   Which animal?

   It makes the train sound like a lion.

2. The steam-engine shot out of the tunnel, snorting and puffing.

   What was snorted and puffed out by the steam-engine?

   snorting and puffed means that steam is coming out of the train.

3. The railway lines were two straight black serpents disappearing into the tunnel in the hillside.

   How might the railway lines have looked like serpents?

   the railway lines might have looked like serpents because they look as if they never end.

4. A sound like distant thunder issued from the tunnel.

   How might the approaching train have made a rumbling sound like distant thunder?

   The train might have made a distant thunder because some trains are very loud and are even louder if you stand next to them.

5. And then the train had gone, leaving only a plume of smoke to drift lazily over the tall Shisham trees.

   Why was the smoke described as being lazy?

   the smok has been described as being lazy because the wind is slowly blowing the smoke away.

NFER-NELSON

Figure 2: A completed example of Exercise 5 in the Progress in English 10 exam paper published by NFER-Nelson

Model answers for all the questions are produced along with the questions. In the case of questions 1 & 2 the model answers are explicit in that it would be hard for a child to answer the question correctly without writing a model answer. Questions 3, 4 & 5 however are more open ended and the child has the opportunity to show their understanding of the subject. In this case it is down to the human assessor to compare the written response to the model answer and determine whether it was correct or incorrect and mark it accordingly.

Two experiments have been designed that employ different approaches to recognise a handwritten response. The recognised responses in both cases are then assessed using the same assessment criteria. In addition, the use of the questions past response history is investigated to show that the model answer is insufficient and a higher 'real world' knowledge is required to mark the answers.

Conventional Lexical Approach

In this first experiment, a conventional lexicon was generated from the stimulus provided, Fry's 300 most frequent words (Fry *et al)* and all the words that have been written in both the test & training set. The stimulus for the exercise is a short story and the questions themselves. All the written words are used to generate the generalised lexicon as this is not an exercise to test the recognition potential of the system but to provide a baseline measure as to how well the assessment process can deal with errors introduced at the recognition stage. Fry's 300 most frequent words claim to represent 75% of all words used. In this exercise 54% of the words written are from the 300 list. Table 1 shows where the words used in the lexicon originated from in relation to Fry's 300 word list.

	Words in Lexicon
In the written responses but not in Fry's 300	38%
In the written responses and in Fry's 300	54%
Not in the written responses/stimulus but in Fry's 300	6%
Not in written responses or in Fry's 300 but in stimulus	2%

Table 1: This table shows where the words that created the lexicon originated from in relation to Fry's 300 most frequent words.

The size of the lexicon used in this experiment is 1455 words and a low recognition rate is expected as a result of a large lexicon made up largely of small words. In holistic recognition, small words have this effect as the number of unique features within the words is low thus causing high ambiguity between the words in the lexicon.

In the Conventional Lexical Approach (CLA) the word in the lexicon that best matches the word image is used to build a recognised response, which is passed on for automatic assessment. Figure 3 shows an example of how CLA is used to build a recognised response. The written response of "The vibration of the wheels on the tracks" is extracted and every word is

independently passed to the recogniser. A list of best matched words is produced for each word (the top three best matches are shown). The top word match in each case is then used to build a recognised response that is used in the assessment stage. For the response shown, in figure 3, this would be "once vibration eyes The where on five made".

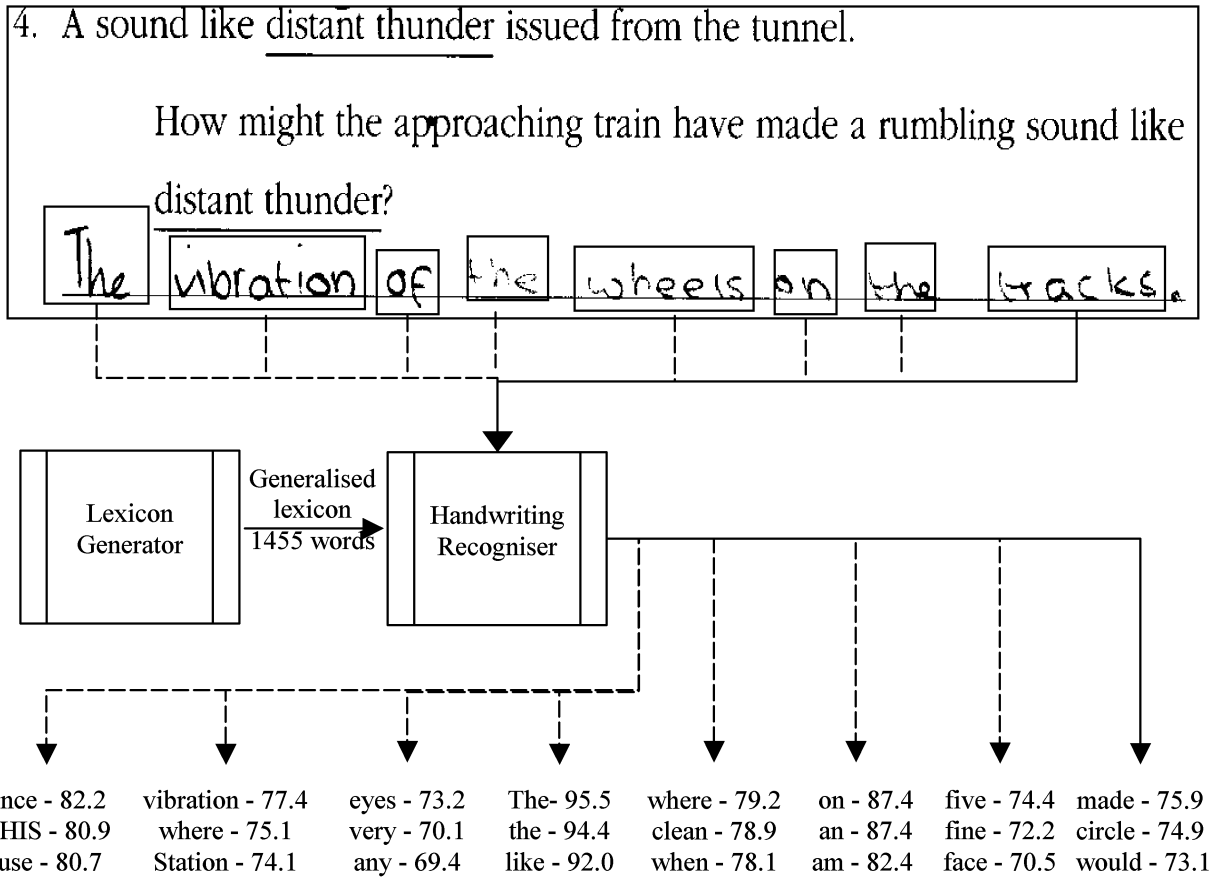


Figure 3: An example of the Conventional Lexical Approach being employed to recognising a handwritten sentence

Specific Word Assessment Technique

SWAT exploits the nature of the question and answer medium by only comparing each word image to the template of the correct answer(s) for that specific question. For Q1 all word positions will be recognised using a lexicon containing only the word *dragon.* Of course, by neglecting any other response, this approach will always generate the correct answer as the recognised response. This localised approach must then use two confidence thresholds per word in each model answer to classify the recognised words either as a keyword (KEY), possible keyword (POS) or not a keyword (NKY). This is achieved using the training set. Each word in the model answer is compared against all the word images in the training set.

From this a frequency density graph can be produced for each keyword based upon the recogniser confidence score (see figure 4). Two data sets are shown on the graph. The solid-line is the frequency density scores for the times when the recogniser is passed a word image that is a keyword, and the dashed-line shows the times when a word image is not a keyword. A high recognition score (higher than T2) implies that the recogniser has achieved a close match between the word image and one of its word target templates. The system can thus confidently classify the word as a keyword. However, if the word has a low score (lower than T1), this means that the recognised word either is a word that is not in the model answer or is illegible and it can therefore be automatically classified as not a keyword. This can be achieved with a high confidence as the ambiguity within the lexicon has been removed. If a word has a score between the two thresholds then, owing to a lack of confidence, the word must be classified as a possible keyword.

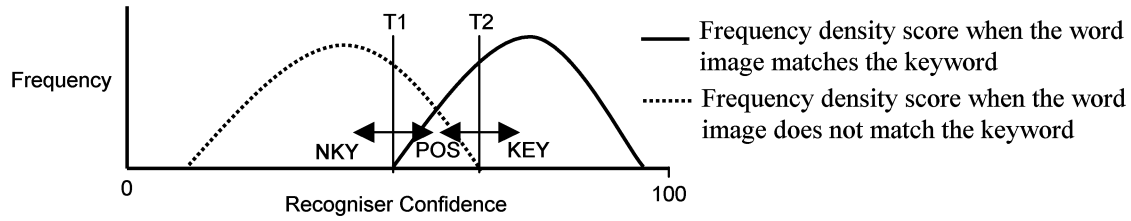


Figure 4: A stylised frequency density graph to obtained the two confidence thresholds for SWAT

Figure 5 shows an example of how SWAT can be applied to build a recognised response for the written response: “The vibration of the wheels on the tracks” using only the model answer as the lexicon.

Using the two thresholds for ‘wheels’ (T1 = 60.9 & T2 = 74.9) and ‘track’ (T1 = 61.8 & T2 = 87.5) the recognised words can be evaluated and classified. The classified response, using the thresholds, would be: ‘*POS POS NKY NKY* **KEY** *NKY NKY POS’*.

Since SWAT is confident that the word images which are classified as NKYs are not a keyword and SWAT is not confident about the word images classified as POS then only the recognised words that are classified as keywords will be passed to the assessment stage. In this example the recognised response would be ‘wheels’, since it is the only word that was classified as a keyword.

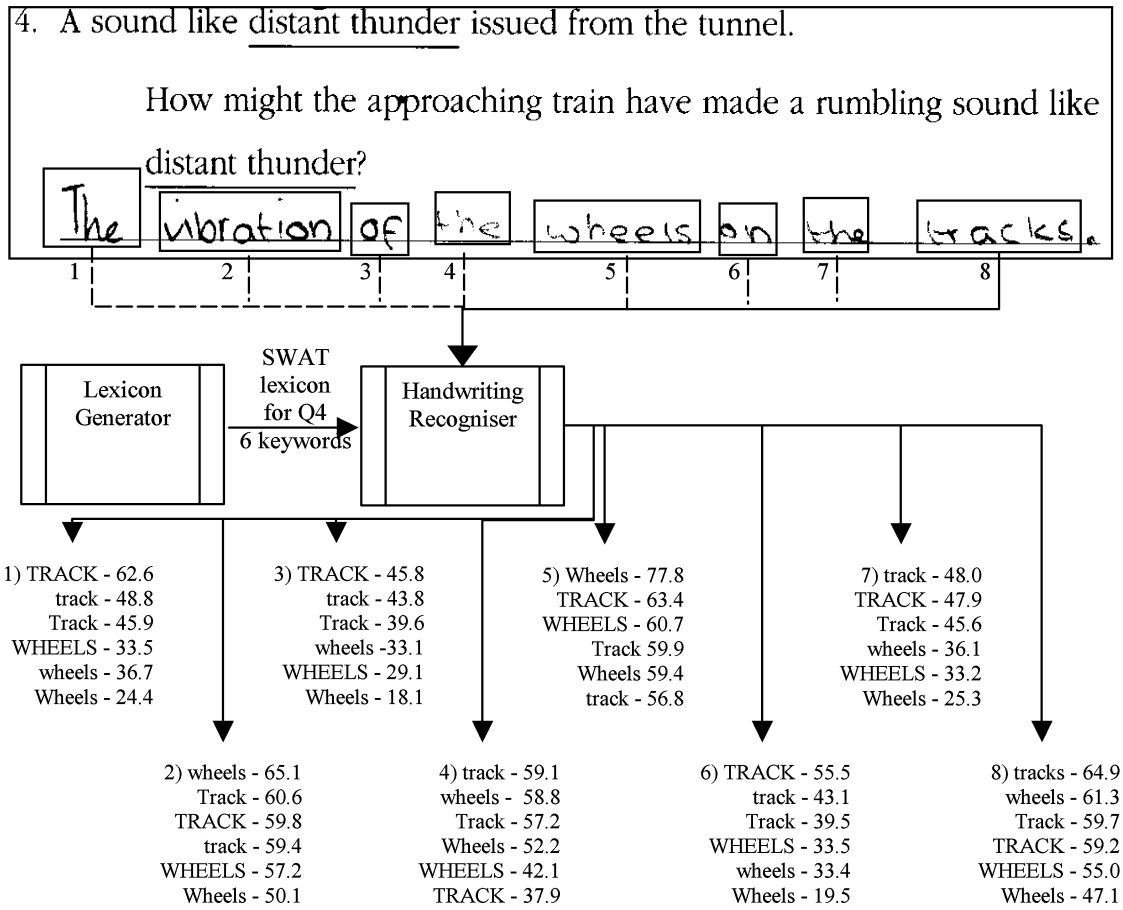


Figure 5: An example of the Specific Word Assessment Technique being employed to recognising a handwritten sentence

Assessment Criteria

The assessment criteria in both experiments is the same. To assess the response, each word in the recognised response is checked against the model answer. The correct answer for question 1 is simply *dragon,* but question 2 has two possible correct answers, *steam* and *smoke.* For questions 3,4 & 5 the model answers can be found in Appendix Ai. If the whole of a model answer is found in any of the word positions and in the correct orientation, then the whole response is scored as correct. If only a partial model answer is found or keywords are found but in the wrong order then the response is passed for manual marking. If this is the case then it must be passed for manual assessment, as the price of assessing a miss-recognised response is too high (i.e. marking a correct response as incorrect). If no keywords are found that relate the recognised response to the model answer then the response can be marked as incorrect. To make an initial comparison in respect of the inclusion of knowledge into the experiments a History set is created from previous correct, and frequently incorrect answers. This set is used to augment the model answers to form new assessment criteria, as shown in Appendix Aii.

Using either criteria, the written example in figures 4 & 5 would be automatically marked as incorrect by CLA because it contains no keywords and passed for manual assessment by SWAT as a partial answer was found.

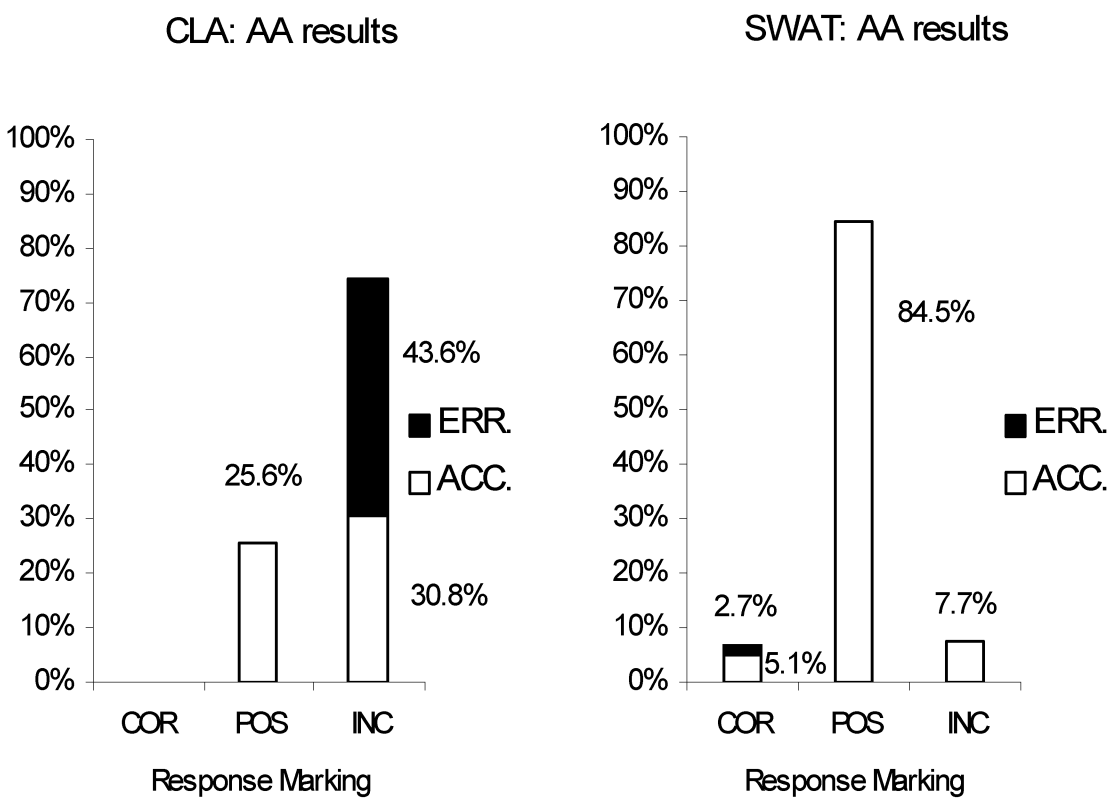
Results

26 children aged between nine & ten, completed five questions as part of Exercise 5 in the Progress in English 10 exam in June 2000 published by NFER-Nelson. Two data sets were randomly selected to form a test set and a training set, 13 writers in each. The training set was also used as the history set. The test set contained 65 written responses (592 word images). Using a 1455 word lexicon, with all the written words held within it, CLA and CLA with history (CLAH) achieved a word recognition rate of only 33%. This compares to SWAT with History (SWATH) where 63.9% of the words were classified as keywords/non-keywords with an accuracy of 97.1%. The remaining 36% of words were classified as possible keywords. The assessment results of the responses can be seen in the next two sections, where both approaches have been applied. First the recognised responses are assessed without history and then with the history incorporated in the assessment criteria.

Without History

Figures 6 & 7 show the results of the CLA & SWAT assessment approaches. It can be seen that 25.6% of the responses recognised using CLA were sent for manual assessment because a partial model answer was found. The remaining 74.4% have been automatically assessed as incorrect answers. 58.6% of these are actually correct answers that have been miss-recognised and erroneously assessed. This error rate is primarily due to the low word recognition rate. In contrast SWAT automatically assessed 15.5% of all response rejecting 84.5% for manual assessment. 2.7% of the responses where incorrectly marked as correct. This was a result of two children giving an incorrect response to question two which included the phrase ‘steam engine’. SWAT confidently assessed ‘steam’ as being a keyword therefore the response was automatically marked as correct.

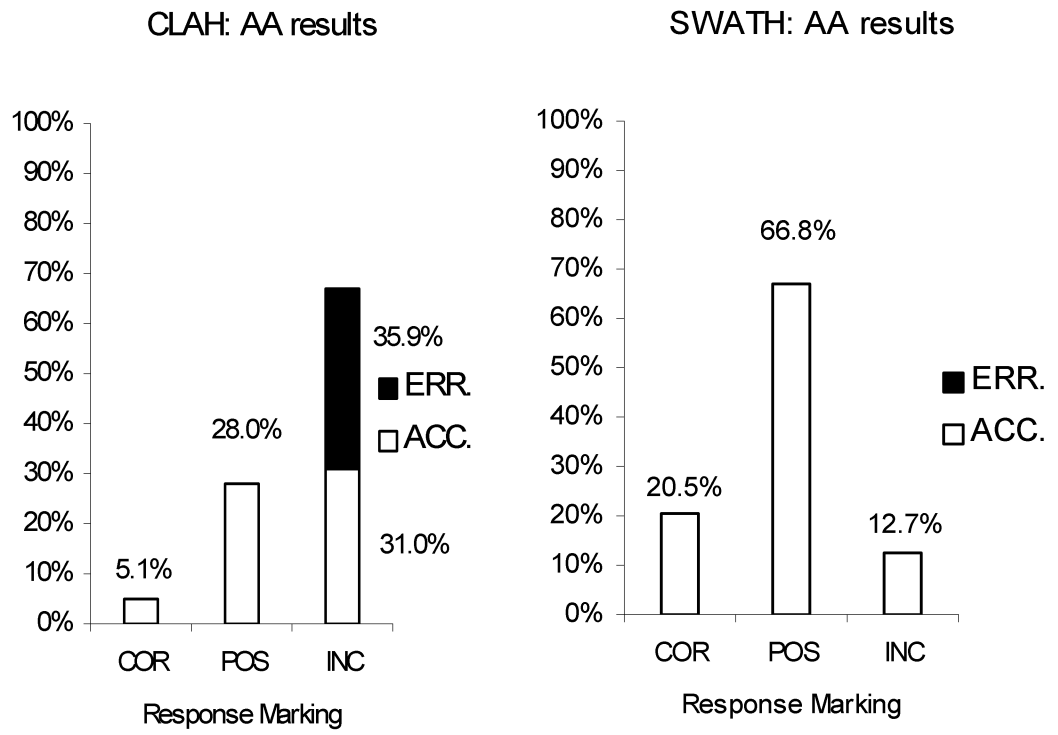
SWAT achieved a higher assessment accuracy than CLA, however the number of assessed responses was low as a result of the assessment criteria being too limited. This shows that human assessors most use ‘common sense’ or additional knowledge to score a written response against a model answer.



Figures 6 & 7: Graphs to show the results of automatically assessing the handwritten responses using the CLA & SWAT approaches respectively

With History

As in the previous experiments the graphs for the automatic assessment of the responses using the two approaches can be shown, see figures 8 & 9. With the inclusion of new assessment criteria the accuracy of both methods has increased. However the number of responses automatically assessed by CLAH has decreased whilst it has increased using SWATH. In the case of CLAH, this is a result of recognising more partial model answers as there are more keywords in the assessment criteria that can be matched to recognised words. However SWATH can confidently assess more responses because it has a high keyword/non-keyword classification rate and is therefore able to identify more complete model answers. When the history data is added to the assessment criteria, the accuracy of the SWATH increases to 100%. This was due to the approach being able to assess the response 'steam engine' in question 2 as incorrect because the new assessment criteria now classifies the phrase 'steam engine' as incorrect. By comparing these results to those shown in figures 6 & 7, the addition of the history is shown to make up for the lack of 'common sense' therefore making it possible to automatically score more responses.



Figures 8 & 9: Graphs to show the results of automatically assessing the handwritten responses using the CLAH & SWATH approaches respectively

A summary of the number of responses that are automatically assessed and the assessment accuracy of each approach is given in table 2.

	Responses Automatically Assessed (%)	Assessment Accuracy (%)
CLA	74.4	41.4
CLAH	72.0	46.3
SWAT	15.5	82.8
SWATH	33.2	100

Table 2: A summary of the Assessment Accuracy and % of responses assessed for all approaches

Conclusion

In this paper, two methods of assessing children's handwritten sentence responses have been compared. The conventional lexical approach, using a 1455 word lexicon, provided high assessment yields of 74.4%. However, this approach incurred a large number of errors resulting in a response accuracy of just 41.4%. This increased slightly to 46.3% when the history was introduced but at the expense of the response yield. This is direct result of the poor recognition rate (33%) when using a generalised lexicon. SWAT on the other hand has a very high keyword/non-keyword classification rate (97.1%) and thus had a higher response assessment accuracy (82.8%). However, this again was at the expense of the total number of responses automatically assessed (15.5%). SWATH assessed 17.7% more responses than SWAT, automatically marking 33% of the responses with an accuracy of 100%. This

is close to being commercially viable, however a large scale trail of SWATH is required to determine if these results can be sustained and therefore be a viable solution to ease the burden of marking traditional handwritten responses.

References

Allan, J. Allen, T. Sherkat, N. **Automated Assessment: It Assessment Jim But Not As We Know It**. The sixth International Conference on Document Analysis and Recognition (ICDAR '01), Seattle, Sept 10th 2001 pp 926 – 930.

Allan, J. Allen, T, Sherkat, N. **Confident Assessment of Children's Handwritten Responses.** Accepted for presentation at the 8th International Workshop on Frontiers in Handwriting Recognition. Niagara on the Lake, 2002.

Evans, R. Sherkat, N. Whitrow, R. **Holistic recognition of static handwriting using structural features,** *(1999). Document Image Processing and Multimedia (DIPM'99), IEE Colloquium 99/041*

Fry, E., Kress, J., and Fountoukidis, D. (2000). **The Reading Teacher's Book of Lists**. Paramus, New Jersey: Prentice Hall.

Appendix Ai – model answers for Q1-5 without History

Model answer for Q1 –	DRAGON
Model answer for Q2 –	STEAM
	SMOKE
Model answer for Q3 –	LONG BLACK
Model answer for Q4 –	WHEELS TRACK
Model answer for Q5 –	STAYED BEHIND TRAIN GONE

Appendix Aii – model answers for Q1-5 with History

Model answer for Q1 –	DRAGON
History answer for Q1 –	LION (frequent incorrect answer)
Model answer for Q2 –	STEAM SMOKE
History answer for Q2 –	'STEAM ENGINE' (frequent incorrect answer)
Model answer for Q3 –	LONG BLACK
History answer for Q3 –	LOOKED SNAKES
Model answer for Q4 –	WHEELS TRACK
History answer for Q4 –	RATTLING LINES THROUGH TUNNEL ENGINE
Model answer for Q5 –	STAYED BEHIND TRIAN GONE
History answer for Q5 –	STAYING BEHIND TRAIN GONE DID NOT MOVE DIDN'T MOVE STAYED THERE NOT GOING ANYWHERE FLOATS ABOUT FLOATING AIR TRAIN GONE