

FOR REFERENCE ONLY

41 0640263 8



ProQuest Number: 10290191

All rights reserved

INFORMATION TO ALL USERS

The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if material had to be removed, a note will indicate the deletion.



ProQuest 10290191

Published by ProQuest LLC (2017). Copyright of the Dissertation is held by the Author.

All rights reserved.

This work is protected against unauthorized copying under Title 17, United States Code
Microform Edition © ProQuest LLC.

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 – 1346

326124

THE NOTTINGHAM TRENT UNIVERSITY LLR	
Shore Loan	PHD/C/05
Rea	ALL

**AUTOMATED ASSESSMENT
OF
HANDWRITTEN SCRIPTS**

JONATHAN ALLAN

**A thesis submitted in partial fulfilment of the
requirements of The Nottingham Trent University
for the degree of Doctor of Philosophy**

June 2005

AUTOMATED ASSESSMENT OF HANDWRITTEN SCRIPTS

Abstract

In this thesis, the automatic assessment of handwritten responses to formal exam questions is introduced as a novel research area. This work highlights difficulty of recognising children's handwriting since the handwritten samples used are often of a poorer quality compared to that given by an adult. The work also shows that the errors that are introduced through recognition of the handwriting can be overcome by using a Specific Word Assessment Technique (SWAT). This technique utilises the nature of the assessment medium to concentrate on scoring responses according to how well the handwritten images match to the actual correct answer. This is in direct contrast to a Conventional Lexical Approach (CLA), which is required to match the handwritten image against all possible answers. In the CLA, the automatic assessment is reliant upon the handwriting recognition stage producing a perfect reconstruction of the written responses before being compared to the model answer.

The performance of both the CLA and the novel SWAT is evaluated when each method is employed to assess a number of different question response styles. In the first instance, a preliminary investigation is carried out using the CLA in order to determine the practicality of automatically assessing highly constrained adult handwritten responses. SWAT is then introduced as an alternative method for automatically assessing children's single word handwritten responses and a retrospective experiment is then carried out, employing SWAT to automatically assess the adult's handwritten responses from the preliminary investigation.

The thesis will show that the generalised CLA is not robust enough to be able to cope with the errors introduced at the recognition stage and therefore the overall automatic assessment system incurs a large inaccuracy. This is shown to its fullest extent when the CLA is employed to automatically assess children's handwritten sentence responses. The CLA automatically assessed 88% of all the responses, but this was at the expense of an overall assessment accuracy of only 37%. The use of a questions history is also exploited in order to give greater assessment accuracy. It is used to help assess the recognised responses from both of the two methods and was compared to baseline results where the history has not been used. Results show that the SWAT with History (SWATH) has the better performance with an overall assessment accuracy of 100%. The high accuracy has been achieved at the expense of the total number of responses assessed, 33%. The approach was not sufficiently confident of 67% of the responses to automatically assess them however the system was able to automatically set aside the responses for human intervention.

The work in this thesis illustrates the potential for automatically assessing handwritten responses using current handwriting recognition systems and provides a basis for future research in the area of automatic assessment of handwritten scripts.

The author would like to take this opportunity to thank the people who have made this thesis possible.

I owe a great debt to my supervisors, Dr Tony Allen and Dr Nasser Sherkat, as without the guidance and encouragement I would not have found the motivation to complete this work. Thank you so much.

I would also like to thank my family and friends, who have kept my head straight and told me some hard truths when needed but also shown me great understanding when I have been at my most focused. I will promise to start to rebuild the bridges that have been worn by my time engrossed in my research.

Thank you.

LIST OF FIGURES	vi
LIST OF TABLES	ix
CHAPTER 1:INTRODUCTION TO AUTOMATIC ASSESSMENT	1
1.1 FORMS OF SUMMATIVE ASSESSMENT	2
1.2 QUESTION STYLES	4
1.3 AUTOMATED ASSESSMENT	8
1.3.1 <i>Optical Mark Recognition Scoring</i>	8
1.3.2 <i>Automatic On-Line Scoring Using Natural Language Processing</i>	9
1.3.3 <i>Traditional Handwritten Exams</i>	11
1.4 SUMMARY OF JUSTIFICATION	12
1.5 OUTLINE OF REPORT	12
CHAPTER 2:LITERATURE SURVEY	15
2.1 INTRODUCTION TO HANDWRITING RECOGNITION	16
2.2 THE PROBLEM OF STYLE VARIATION	17
2.3 GENERAL HANDWRITING RECOGNITION	21
2.3.1 <i>Character Based Recognition</i>	22
2.3.2 <i>Holistic Word Recognition</i>	24
2.3.3 <i>Handwriting Recognition Methods</i>	25
2.4 APPLIED HANDWRITING RECOGNITION	30
2.4.1 <i>Postal Address Interpretation</i>	31
2.4.2 <i>Cheque Amount Verification</i>	33
2.4.3 <i>Signature Verification</i>	34
2.5 POST-PROCESSING FOR IMPROVED RECOGNITION PERFORMANCE	36
2.5.1 <i>Confidence Measures In Speech Recognition</i>	36
2.5.2 <i>Post-Processing Measures In Handwriting Recognition</i>	40
2.6 SUMMARY	43
CHAPTER 3:PRELIMINARY INVESTIGATION TO TEST THE VALIDITY OF AUTOMATICALLY ASSESSING HANDWRITTEN RESPONSES	45
3.1 EXPERIMENTAL PLATFORM	48
3.1.1 <i>Image Extraction</i>	48

3.1.2	<i>Lexicon Generator</i>	51
3.1.3	<i>Handwriting Recogniser</i>	57
3.1.4	<i>Assessment</i>	62
3.2	MANUAL ANALYSIS OF THE TEST SET	65
3.3	RESULTS	67
3.3.1	<i>Recognition Results</i>	68
3.3.2	<i>Assessment Results</i>	70
3.4	SUMMARY	71
CHAPTER 4: NOVEL WORD VERIFICATION-BASED ASSESSMENT OF CHILDREN'S HANDWRITTEN RESPONSES		72
4.1	SINGLE WORD CHILDREN'S RESPONSES	72
4.2	IMAGE EXTRACTION	76
4.3	CONVENTIONAL LEXICAL BASED RECOGNITION & ASSESSMENT APPROACH	79
4.3.1	<i>Raw Assessment Of The Conventional Lexical Approach</i>	81
4.3.2	<i>Results For The Conventional Lexical Approach</i>	82
4.3.3	<i>Threshold Assessment Of The Conventional Lexical Approach</i>	84
4.3.4	<i>Threshold Results For The Conventional Lexical Approach</i>	86
4.4	SPECIFIC WORD ASSESSMENT TECHNIQUE	88
4.4.1	<i>Swat Applied To Children's Single Word Responses</i>	89
4.4.2	<i>Swat Results</i>	92
4.5	SWAT APPLIED TO ADULT'S SINGLE WORD RESPONSES	94
4.5.1	<i>Applying Swat To Adult's Three Word Responses</i>	95
4.5.2	<i>Assessment Results Of The Adult Dataset When Swat Was Applied</i>	99
4.6	CONCLUSION	102
CHAPTER 5: AUTOMATED ASSESSMENT OF CHILDREN'S HANDWRITTEN SENTENCE RESPONSES		104
5.1	CHILDREN'S SENTENCE RESPONSE ASSESSMENT	106
5.2	INVESTIGATION 1 – CLA VS. SWAT TO AUTOMATICALLY ASSESS HANDWRITTEN SENTENCE RESPONSES	111
5.2.1	<i>Conventional Lexical Approach</i>	111
5.2.2	<i>Results - Conventional Lexicon Approach</i>	114
5.2.3	<i>Specific Word Assessment Technique</i>	115
5.2.4	<i>Result - Specific Word Assessment Technique</i>	119

5.3	INVESTIGATION 2 – ADDITION OF CONTEXTUAL WORD BRIDGES	121
5.3.1	<i>CLA Using Contextual Word Bridges</i>	122
5.3.2	<i>Results - CLA Using Contextual Word Bridges</i>	125
5.3.3	<i>SWAT With Contextual Keyword Bridges</i>	126
5.3.4	<i>Results – SWAT With Contextual Keyword Bridges</i>	130
5.4	INVESTIGATION III - ADDITION OF QUESTION RESPONSE HISTORY	131
5.4.1	<i>Without The Use Of History</i>	131
5.4.2	<i>With Question Response History</i>	133
5.5	CONCLUSION	134
	CHAPTER 6: CONCLUSION AND FUTURE WORK	137
6.1	WHAT HAS BEEN ACHIEVED FROM THIS RESEARCH?	137
6.2	HOW FAR CAN AUTOMATED ASSESSMENT GO?	141
6.2.1	<i>Proposed Improvements to the Automatic Assessment System</i>	141
6.2.2	<i>How Robust Is Automatic Assessment?</i>	143
6.2.3	<i>Other Applications</i>	144
	REFERENCES	146
	APPENDIX A: A SAMPLE OF ADULT’S HANDWRITTEN RESPONSES TO MULTIPLE CHOICE QUESTIONS USED IN CHAPTERS 3 AND 4	
	APPENDIX B: A SAMPLE OF CHILDREN’S HANDWRITTEN SINGLE WORD RESPONSES TO NFER-NELSON’S ‘PROGRESS IN ENGLISH 6 EXAM, EXERCISE 2 PART A’ USED IN CHAPTER 4	
	APPENDIX C: MODEL KEYWORD ANSWERS FOR NFER-NELSON’S ‘PROGRESS IN ENGLISH 10, EXERCISE 5 THE TUNNEL’ USED IN CHAPTER 5	
	APPENDIX D: A SAMPLE OF CHILDREN’S HANDWRITTEN SENTENCE RESPONSES TO NFER-NELSON’S ‘PROGRESS IN ENGLISH 10 EXAM PAPER, EXERCISE 5 THE TUNNEL’ USED IN CHAPTER 5	
	APPENDIX E: PUBLICATIONS RESULTING FROM THE THESIS	

LIST OF FIGURES

Figure 1.1:	Example of a multiple-choice question from Progress in English 6, published by NFER-Nelson	4
Figure 1.2:	Example of a short answer question from Progress in English 10, published by NFER-Nelson	5
Figure 1.3:	Example of a problem style question from Mathematics 10, published by NFER-Nelson	6
Figure 1.4:	An example of part of an essay style response given by a computing student in their second year of university	7
Figure 1.5:	An example of an OMR ready answer sheet	8
Figure 2.1:	Example of a style classification of handwriting	17
Figure 2.2:	Example of a case classification of handwriting	17
Figure 2.3:	Examples of children's handwriting	19
Figure 2.4:	Simplistic view of character recognition	22
Figure 2.5:	Example of ambiguity within written words	24
Figure 2.6:	Example of the complexity when segmenting a cursively written word	24
Figure 2.7:	The holistic word recognition method. The extracted features consists of vertical bars, holes and cups	25
Figure 3.1:	An example of a completed test question	45
Figure 3.2:	Proposed Handwriting Recognition and Automatic Assessment System	48
Figure 3.3:	The image extraction process	49
Figure 3.4:	The lower case character set	51
Figure 3.5:	The upper case and initial mixed case character set	52
Figure 3.6:	Example of upper-case zoning	52
Figure 3.7:	Example of mixed case zoning	52

Figure 3.8:	An example of question one completed with blue ink	56
Figure 3.9:	Example of lower-case zoning	58
Figure 3.10:	An example of how the zoning can be corrupted	59
Figure 3.11:	The feature extraction process	60
Figure 3.12:	The image extraction process	60
Figure 3.13:	The feature matching process	61
Figure 3.14:	An example of a completed question	63
Figure 3.15:	A chart depicting the overall response spread	65
Figure 3.16:	An example of an OTHER question due to a numbered response	66
Figure 3.17:	An Example of an Abbreviated Response	66
Figure 3.18:	The manual classification of style	66
Figure 3.19:	Assessment classification results	70
Figure 4.1:	Question page from Progress in English 6 published by NFER-Nelson	73
Figure 4.2a:	Example of child 4's responses to Q1-5	74
Figure 4.2b:	Example of child 2's responses to Q1-5	74
Figure 4.3:	The RGB colour model used by Wing <i>et al</i> to define the quantisation thresholds	76
Figure 4.4:	An example of how a grey pixel is transformed into either a black or white pixel	77
Figure 4.5:	An example of how the quantisation process can fail and how applying a new threshold can overcome the fault	78
Figure 4.6:	The conventional approach to handwriting recognition	79
Figure 4.7:	Example of the conventional recognition results for child 8	82

Figure 4.8:	Graphs to show the Automatic Scoring in the Traditional Approach	83
Figure 4.9:	An idealised example of a distribution graph showing the frequency distribution of the correctly and incorrectly recognised response scores	84
Figure 4.10:	Three examples of written response that have been assessed using the threshold for the question "His ____ is in September"	85
Figure 4.11:	Graph to show the smoothed frequency distribution of the correctly and incorrectly recognised response scores and the value of threshold T	86
Figure 4.12:	Graph to show the response yields after the threshold has been applied	87
Figure 4.13:	The Novel recognition system	89
Figure 4.14:	An idealised example to show the frequency distribution of the correct and incorrect response scores and the positions of the thresholds t_1 and t_2	90
Figure 4.15:	Three examples of written response that have been assessed using the thresholds for the question "His ____ is in September"	91
Figure 4.16:	Example of the novel recognition results for child 8	92
Figure 4.17:	Graph to show the smoothed frequency distribution of the correct and incorrect response scores and the values of thresholds t_1 and t_2	92
Figure 4.18:	Graph to show the response yields after the thresholds have been applied	93
Figure 4.19:	An example of question one from the adult dataset	94
Figure 4.20:	The word position specific thresholds determined for question one	96
Figure 4.21a:	The Hard Assessment Criteria	97
Figure 4.21b:	The Medium Assessment Criteria	98

Figure 4.21c:	The Soft Assessment Criteria	98
Figure 4.22a:	Results of assessing the adult responses using the hard assessment criteria	100
Figure 4.22b:	Results of assessing the adult responses using the Medium assessment criteria	100
Figure 4.22c:	Results of assessing the adult responses using the soft assessment criteria	101
Figure 5.1:	The automatic assessment system	106
Figure 5.2:	A completed example of Exercise 5 in the Progress in English 10 exam paper published by NFER-Nelson	109
Figure 5.3:	An example of the recognition results when using the Conventional Lexical Approach	112
Figure 5.4:	A graph to show the assessment accuracy of the conventional lexicon approach	114
Figure 5.5:	An idealised frequency density graph to obtain the two confidence thresholds for SWAT	115
Figure 5.6:	An example of the recognition results when using the Specific Word Assessment Technique	117
Figure 5.7:	A graph to show the frequency density graph of the responses to Q1, Exercise 5 of the Progress in English 10 exam paper	119
Figure 5.8:	A graph to show the assessment yield and assessment accuracy of the Specific Word Assessment Technique	120
Figure 5.9:	Example of contextual word bridges in a sentence	123
Figure 5.10:	The top three bridged sentences formed for the recognition results shown in figure 5.3	124
Figure 5.11:	A graph to show the assessment accuracy of the conventional lexicon approach with contextual bridges	125
Figure 5.12:	The SWATB bridge set used for question four	126
Figure 5.13:	All possible bridged sentences that can be generated from 2,3 & 4 word written responses	127

Figure 5.14:	A graph to show the assessment accuracy of the conventional lexicon approach with contextual keyword bridges	130
Figure 5.15:	A graph to show the results of automatically assessing the handwritten responses using CLA	132
Figure 5.16:	A graph to show the results of automatically assessing the handwritten responses using SWAT	132
Figure 5.17:	A graph to show the results of automatically assessing the handwritten responses using CLAH	133
Figure 5.18:	A graph to show the results of automatically assessing the handwritten responses using SWATH	133

LIST OF TABLES

Table 2.1:	The state of the art in general handwriting recognition	25
Table 2.2:	The state of the art in PAI	32
Table 2.3:	The state of the art in CAV	33
Table 2.4:	The state of the art in SV	35
Table 2.5:	State of the art in applying confidence measures to speech	37
Table 2.6:	State of the art in applying post-processing to handwriting	40
Table 3.1:	The 210 word lexicon for all the questions	54
Table 3.2:	The 27-word specific lexicon for question one	55
Table 3.3a-c:	9 word lexicons for the first three word positions for question one	56
Table 3.4:	A table to show the possibility of errors in the classified Responses	64
Table 3.5:	Raw recognition results	68
Table 4.1:	The possible classification outcomes of the system	80
Table 4.2:	The word position specific classification for question one of the adult three word responses	97
Table 4.3:	The Assessment of the recognised words for question one	99
Table 5.1:	This table shows where the words that created the lexicon originated from in relation to Fry's 300 most frequent words	111
Table 5.2:	Classification of the first, third and fifth word from the response in figure 5.6	118
Table 5.3:	SWAT thresholds for the model keywords for all five questions	120

Table 5.4:	The high (t_2) and low (t_1) thresholds calculated for SWATB	129
Table 5.5:	A summary of the Assessment Accuracy and % of responses assessed for all approaches	134
Table 6.1:	Raw recognition results from Chapter 3	138

CHAPTER 1: INTRODUCTION TO AUTOMATIC ASSESSMENT

Assessment is a vital tool in education. It helps teachers measure the ability of students and identifies individuals who are struggling, improving or excelling at the assessed subject. Assessment information also provides an indication of the rate of learning within a given class. This can then be compared to national averages such as those given by the National Curriculum. However, assessment is very time consuming and this is particularly so for the marking of formal examinations, where double marking is often used. This process is very labour intensive and therefore costly. However, it is essential in order to sustain the high level of accuracy required for such assessments. The work presented in this thesis aims to address this dilemma by investigating novel automated approaches for the assessment of handwritten responses in order to increase the efficiency of the assessment process whilst maintaining a high assessment accuracy. In the remainder of this chapter, the different forms of assessment and question styles currently in use are discussed along with the advantages that can be gained by employing Computer Aided Assessment (CAA).

There are two forms of assessment, summative and formative. Formative assessment is an iterative process used to evaluate the students understanding of a subject before the opportunity to learn the material has passed [16]. This type of assessment is usually straightforward and relatively quick to process, in order that the assessment information can be applied immediately. Summative assessment, on the other hand, is used at the end of a given period of teaching to provide a measure of the 'sum of knowledge' that the student has learnt [16]. Thus, more complex assessment can be applied since the

assessment information is not required immediately. However, this then requires a lot more manual labour to process all of the assessment information. Work to introduce computerised formative assessment is currently being carried out [97,120], however the work in this thesis will focus on the assessment of formal handwritten summative assessment in order to increase the efficiency of this assessment method.

1.1 Forms of Summative Assessment

There are various forms of summative assessment; a few are listed below:

- Unseen Examination
- Seen Examination
- Essay
- Project

Unseen examinations are the traditional form of summative assessment. Such assessments are used to evaluate the student's knowledge of a given subject. Students are confronted with a number of question formats e.g. essay, problem, short-answer or multiple-choice questions (see section 1.2). The examinations are said to be unseen, as the students have no prior knowledge of the questions that will be given. In addition, the students will also have to complete the exam with no other stimulus to aid them. This is in complete contrast to seen examinations where students have access to additional resources such as case-material before an examination. These resources are then used as the stimulus in the exam, although the students still have no prior information about the questions that will be given. Additionally, in seen examinations, students may use textbooks (new or untainted) to reference their answers during an exam. In this way, they can be given an opportunity to show their ability to construct arguments and informative

answers. The question styles used for this type of assessment tend to require an essay style response.

An essay, in itself, is another traditional form of summative assessment. Instead of the student being asked to complete the task in a controlled and timed environment, as in an exam, they are encouraged to investigate and research the question in their own time. Their response takes the form of a piece of writing and is only restricted by a set word-limit. As times and technology have changed so the medium in which an essay is presented has also changed from being handwritten responses on loose leaf paper to an electronic submission or hardcopy of word processed document. A project, report or coursework is not too different from an essay. However, in addition to the written criteria the student must also carry out a practical investigation.

1.2 Question Styles

The simplest style of question is the *multiple-choice* question (MCQ). MCQs consist of a prompt, either a question or the start of a sentence which the student must then complete (see figure 1.1).

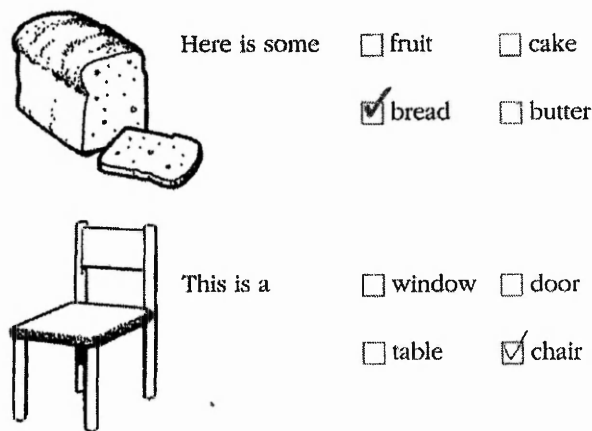


Figure 1.1: Example of a multiple-choice question from Progress in English 6, published by NFER-Nelson

The student has then to choose from a number of alternative answers, one of which is the correct answer. MCQs are objective in that there is a single definitive answer. They are therefore relatively easy to mark thus generating the assessment information quickly. This efficient turn-a-round is very desirable since teaching time is limited and is better spent with the students rather than scoring exam papers. However, it is very difficult to grade a student's understanding of a subject solely from an MCQ exam as they are unable to construct an argument or give their own interpretation of the answer. MCQ's have also been prone to criticism, suggesting that they are too easy and that the students can simply guess the correct answer [16].

Short answer questions require a response from the student that has not been given to them as part of a list of alternatives. A response in the form of a single word, part sentence or full sentence can be given to answer a question and is only limited to the space provided. Figure 1.2 shows a question set for 10 year old children in England. The likelihood of the child guessing the correct answer in this case is considerably smaller than it is with MCQ's. Therefore to give the correct answer the child must have some understanding or, at the very least, be able to remember what has been taught. This form of response can also be used to assess the student's ability to write and spell.

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

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

The steam

Figure 1.2: Example of a short answer question from Progress in English 10, published by NFER-Nelson

The complexity of the assessment criteria has also increased from that of the MCQ format, because short answer questions require the child to give their own interpretation of the answers instead of being given implicit answers. The increased complexity of the assessment criteria will therefore impact on how suitable it is for automatic assessment. Simple assessment criteria, as in the case of the MCQ, may be fairly simple to model but criteria containing a number of possible answers to one question will be more difficult to model.

The *problem* question style usually requires the student to carry out some form of calculation and present the answer for the mark (see figure 1.3). Full marks can be given for the correct answer, but partial marks can also be awarded if the student has derived an incorrect solution but has chosen and shown the correct method.

SECTION A

1

Mrs Jones paid three amounts of money into the bank over the summer months.

In June she paid in £50.28.

In July she paid in £37.50.

In August she paid in £17.40.

$$\begin{array}{r}
 \$ 50.28 \\
 \$ 37.50 \\
 \$ 17.40 \\
 \hline
 105.18
 \end{array}$$

How much money did she pay in altogether?

Answer £ ~~105~~ 105.18

Figure 1.3: Example of a problem style question from Mathematics 10, published by NFER-Nelson

A question that requires more than a simple sentence response is described as an *essay* question (see figure 1.4 on the following page). This style of questioning is exactly the same as setting an essay except that the students are under strict exam conditions. As a consequence, this leads to the students focusing on the content and less upon the neatness of the exercise. This will impact on how well a written response can be automatically recognised and will be discussed again in chapter 2. Essay questions usually focus on ideas and understanding and as a result the assessment becomes very broad. Instead of listing all possible answers, more generalised assessment criteria are used in the form of assessment guidelines. To assess all the possible responses the assessors are required to use their own knowledge to assess whether the responses are correct. A computerised

system would require a level of understanding equivalent to that of a human assessor if it is to assess such responses successfully.

1.

1) The NTU Project guide could consist of distinct areas, that a student must use in his/her project.

- eg.
- (i) Title Page
 - (ii) Introduction
 - (iii) Investigations, methods, outcomes etc
 - (iv) Tables of results, other graphical ~~or~~ representations
 - (v) Conclusions
 - (vi) Bibliography / References

This is by no means a true representation of what the real Project guide contains, but the point being made is that all projects should contain all of these areas! but what the student writes is up to him/her

The OSI 7-layer model has a similar purpose to the NTU Project guide, it defines how many areas or layers there should be, but what is contained in that area is up to the ~~author~~ ^{author} to decide what goes in each layer, in this case the author could be a software developer, a computer networking company, etc. As long as each layer has similar information within it meeting certain guidelines the OSI model is upheld.

Each of these layers in reality contain a

Figure 1.4: An example of part of an essay style response given by a computing student in their second year of university

1.3 Automated Assessment

Automatic Assessment (AA) is not a novel concept. As pattern recognition technology has improved so more of the technology is being utilised for applications such as AA. For instance, Optical Mark Recognition (OMR) is regularly employed to score and assess formal examinations. This section briefly describes the state of the art in AA, and shows that the automated assessment of handwritten scripts is the next logical research step.

1.3.1 Optical Mark Recognition Scoring

OMR was first used to score MCQs in 1963 [146] and is still used today to mark examinations all over the world. OMR works by recognising boxes that have been filled in on an OMR ready answer sheet (see figure 1.5).

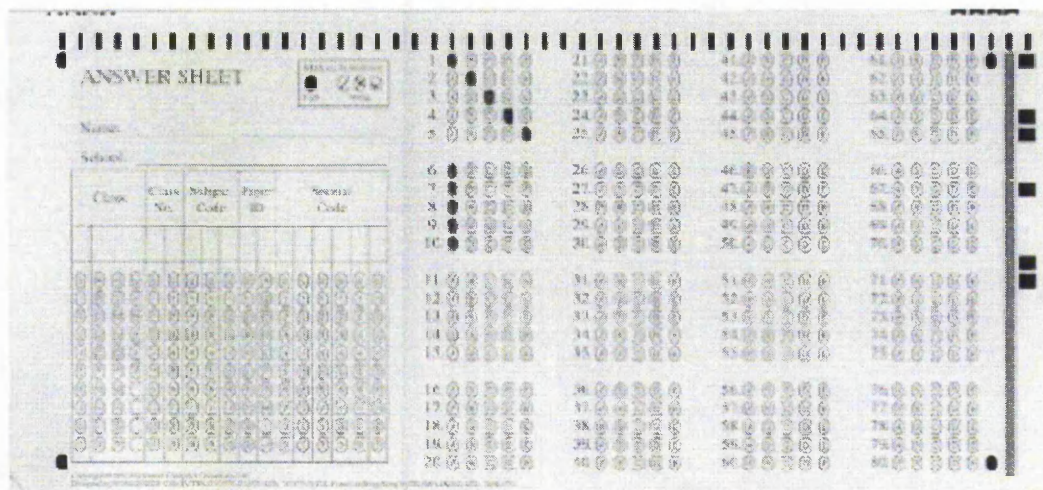


Figure 1.5: An example of an OMR ready answer sheet

Each student is given a question script and an OMR ready answer sheet. They must then read the questions and translate their answer to the sheet by filling in the appropriate box. For example, in question 3 the student has selected the response that requires them to fill

in block C on row 3. The efficiency of this technology has been reported as 99% with an accuracy of 100% [146]; the remaining 1% of the responses having been automatically rejected and passed for manual assessment. Unfortunately, the use of this type of automatic assessment is limited to multiple choice style questions and its use demands that the student be able to efficiently translate their chosen response to a question by marking a separate piece of paper. If the student is not able to carry out this then even if they know the correct answer, they will be penalised. This is unacceptable. Students should not be disadvantaged for the sake of efficient assessment.

1.3.2 Automatic On-Line Scoring using Natural Language Processing

At the forefront of the assessment of on-line testing is the commercial group behind QuestionMarkTM. This software has the ability to generate tests that accept on-line responses to MCQ, problem, and short answer questions. Being text based its recognition accuracy is very good when dealing with the sort of yes/no or single word answers synonymous with MCQ questions.

QuestionMark's performance decreases as the complexity of the response increases however. Research has been done into applying Natural Language Processing (NLP) techniques to the user inputs in order to reduce the ambiguity found in multi-word responses [7,22,23,86] and this has been shown to help improve the assessment.

Burstein *et al* report scoring rates of 97% on computer ready essays and short answer questions using NLP based assessment [22]. Their scoring rate is calculated as the

number of automatically assessed responses that are in agreement with a scoring given by a human assessor. Burstein *et al*'s automatic assessment system scores the responses based on a six-point holistic scale. This is the same method as used in the manual process. The manual process involves two human assessors. Scoring the responses independently. If the two scores differ by more than one point, a third assessor is introduced to resolve the final score. The automatic system takes the place of one of the initial human assessors so that only two human assessors are needed, one to score the responses and one to resolve the responses where the automatic assessment and human assessor are not in agreement. The second human assessor is only required 3% of the time. This level of difference is reported as being comparable to that normally seen when three human assessors are employed.

Despite the reasonable performances of these two systems, there are limitations in using this form of assessment. If the users are incapable of relating their thoughts efficiently via a keyboard or the cost of class testing done on a computer versus the traditional pen & paper is too high, then a move from paper based assessment to on-line testing is impractical. Thus, the automated assessment of handwritten scripts is an obvious progression.

1.3.3 Traditional Handwritten Exams

Traditional handwritten examinations are still the most common form of examination at every level of education and a large proportion of a teacher's time is spent marking these examinations and tests. If these could be assessed automatically, the pressure on teachers would be eased in a similar manner to the way in which OMR and on-line assessment has relieved the burden of assessment for higher education lecturers.

However, the main restriction that arises from this type of application is that it needs to be extremely accurate. The price of a recognition error (i.e., automatically marking as correct a misrecognised response) is much higher than any savings that can be gained from automatically scoring a correctly recognised response. General handwriting recognition systems can correctly recognise only 87% of all responses [25]. Thus, without any post-recognition error detection, 13% of all handwritten responses would be recognised incorrectly. If these recognition errors were to be passed on to the assessment stage an assessment error of 13% would result. This is obviously unacceptable. However, if these misrecognised responses can be detected and filtered out for manual processing before automatic assessment is applied then the assessment accuracy of the system can be maintained. Automatic systems must be able to refuse to assess a response when the probability of making a mistake is deemed to be too high, i.e. the system should only pass on a recognised response to the assessment stage when it is known that the recognition system is highly confident of the recognition being correct.

1.4 Summary of Justification

The aim of this work is to find an economic and reliable solution to the manual assessment of handwritten responses. Traditional human and current on-line assessment are very expensive, therefore the automated assessment of pen and paper exams would be a viable option so long as it can be proven to be as accurate as on-line assessment methods. There are two main reasons to pursue this research. The first is to free a teacher's time so that they can concentrate on teaching, and the second is to prevent the translation errors that can occur using the OMR ready sheets and/or a keyboard.

1.5 Outline of Report

This chapter introduces the rationale for investigating the automated assessment of cursive scripts and shows that applying handwriting recognition to assess handwritten responses is a viable avenue of research. Chapter 2 offers a state of the art review of current handwriting recognition techniques, focusing primarily upon applied handwriting recognition systems and post processing to improve overall system accuracy.

Chapter 3 presents a preliminary investigation using the conventional lexical approach to handwriting recognition. This is used to produce a set of baseline results with the aim of showing that the automatic assessment of handwritten responses is achievable when all of the possible responses are known prior to recognition. Errors introduced by the

recognition processes are shown to be confidently identified, through the use of contextual bridging allowing them to be passed for manual assessment, thereby improving the overall assessment accuracy of the automated system.

In chapter 4 a Specific Word Assessment Technique (SWAT) is developed for the automatic assessment of children's single word responses. The use of contextual bridging cannot be carried over to this style of question responses, as there is only a single word in each of the responses. Therefore SWAT was designed to emulate the assessment reasoning used by a human assessor. Instead of attempting to confidently recognise the response given (as had been the goal in chapter 3) the aim was to assess whether or not the response is correct or not by only evaluating the word image against the correct answer. The approach introduced in this chapter is shown to have a significant advantage over the conventional lexical based assessment as a result of the reduction of ambiguity within the lexicon used. This result is reinforced when SWAT is retrospectively used to assess the data set used in chapter 3.

Chapter 5 introduces a derivative of SWAT that can be used for the process of recognising and assessing children's handwritten single sentence style responses. In this derivative, each question is given a specific lexicon of syntactically ordered keywords derived from a set of model answers. This again minimises the ambiguity found when using larger lexicons and provides a form of confidence measure to check for recognition errors. The chapter then compares the performance of this SWAT derivative against the conventional lexical approach. It then goes on to investigate the benefits of augmenting

SWAT with contextual word bridges as well as using a question's response history to form a more robust assessment solution. Again, in each case, the performance is evaluated against an equivalent conventional lexical approach to determine if there is a performance gain.

Chapter 6 concludes the work and discusses the achievements and possible improvements that could be carried out on the methods produced. This final chapter also suggests possible avenues of research that have emerged throughout this PhD, which could be interesting projects for the future.

CHAPTER 2: LITERATURE SURVEY

How can the automated assessment of a handwritten response be achieved? This is the fundamental question that is to be addressed in this thesis. The task of automatically marking a traditional handwritten exam paper has a definite similarity to the automatic processing of handwritten forms, where it will be shown that it is the recognition of the cursive script that limits the accuracy of such systems. Therefore, before work commences it is necessary to identify the major processes that will be used to build an automated assessment system and to explore the work done by others in the generic research area of handwriting recognition.

A brief review of the problems associated with the recognition of adult and children's handwriting is provided in sections 2.1 & 2.2. Section 2.3 details the principal approaches to handwriting recognition, featuring all aspects of handwriting recognition from Intelligent Character Recognition (ICR) to unconstrained handwriting recognition. This is followed, in section 2.4, with a comprehensive study of the current applications of static handwriting recognition together with an understanding of how the problems involved may impact upon the automatic assessment of handwritten responses. Finally, in section 2.5, current research methods for improving the underlying recognition results are reviewed.

2.1 Introduction to handwriting recognition

There is a huge demand for automatic processing systems that can evaluate and translate vast amounts of paper-based data [125,140]. One example of this has been seen in chapter 1 where OMR methodology is applied to automatically process forms [146]. This form of recognition is highly accurate but is limited to a few applications and prone to user-introduced errors. Another example is the automated processing of handwritten block capital characters found on many forms [43]. Intelligent Character Recognition (ICR) is employed to recognise these characters and, like OMR, it is a commercially viable solution [45,47,76]. The ICR methodology will be discussed in section 2.2.1. The main reason why both of these two techniques are so successful can be attributed to style control. Controlling the structure and design of a form along with how the user provides the relevant information all contributes to constraining the problem of style variation. The companies that create the forms can produce 'computer friendly' forms that have been especially designed so that they can be automatically processed with the minimum of problems. The use of the colour and the layout of the form can be used to identify the form (form registration) and then to locate and extract the handwriting [147]. The producers of the forms can also request that the user write in a specific way (block capitals) and use a specific writing medium i.e. blue or black ink only. This is the only justifiable way to constrain the problem once the form is in the public domain; other constraints such as limiting style variation by limiting the number of users is neither practical nor possible. The inability to handle the vast range of writing styles is the main

restriction that makes the recognition of unconstrained handwriting such a complex task [44,70,92,125,129].

2.2 The Problem of Style Variation

To limit the complexity of style variation it is possible to categorise the types of handwriting to establish the most appropriate methods for recognition. Handwriting can be rudimentarily classified in two ways: style & case. Handwriting style can be further classified into either discrete, cursive or mixed writing (see figure 2.1). Tepar *et al* go further and introduces boxed discrete characters as another style of handwriting [134]. However, this style of handwriting is wholly dependant on whether or not there are predefined boxes to write in. Generally speaking, writers will not limit themselves to write in this manner unless they are prompted to do so. Therefore, boxed discrete characters may be classified as a high-level handwriting style rather than as a fundamental style of handwriting.

discrete - style
cursive - style
 mixed - style

Figure 2.1: Example of a style classification of handwriting

UPPERCASE
Mixed-Case
 lowercase

Figure 2.2: Example of a case classification of handwriting

The second fundamental classification of handwriting is the character-case that the word image has been written in. There are three possible cases: UPPERCASE, lowercase and Mixed-Case (see figure 2.2).

These two classifications of handwriting are independent of each other. Therefore, all handwritten words have both a style and a case. The different examples of styles of handwriting shown in figure 2.1 are all examples of lowercase writing as well as being different style examples. Equally, in figure 2.2, the uppercase words can be classified as discretely written words etc.

Ebadian Dehkordi *et al* have shown that the accuracy of the recognition process can be increased in instances where the style and case of the handwriting can be classified at the pre-processing stage [37]. The way in which this approach improves recognition is two fold. Being able to predetermine the case of a handwritten word cuts down the number of patterns against which it needs to be matched, reducing ambiguity. Secondly, the style of handwriting can be classified, pre-recognition, so that recognition is only attempted for those handwriting styles that have previously been seen by the recogniser. This can be taken further so that the classification process can take into account the style-based performance of a multi-recogniser system such that the style classifier passes the image to the recogniser that is best suited to recognise that style of writing.

In addition to the inherent problems associated with style variation in adult handwriting it should be noted that the majority of handwritten samples requiring assessment are generated from within educational institutes such as schools, colleges and universities. Children as young as 5 years old are required to take formal exams in English and Mathematics. Consequently, there is a need to investigate the impact of children's

handwriting style alongside the problems associated with the recognition of adult's handwriting. An example of children's handwriting can be seen in figure 2.3.

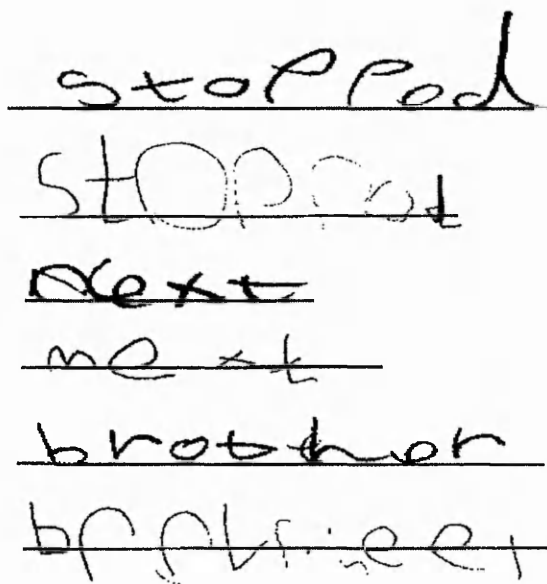


Figure 2.3: Examples of children's handwriting

The handwritten words are responses taken from an exam question written for children aged 6-7 years old. Most of the words are legible even though they are not 'neat'. The children's handwriting is badly formed and misspelt. The formation of words in handwriting is very important to a human reader as it determines how legible the handwriting is. A perfectly formed word will always be legible assuming that the reader has previously learnt the word. However, a badly formed word may still be legible although the reader will have to work harder to understand the words meaning. The reader will often be able to use their experience and the context in which the word has been written (i.e. semantic and syntactic analysis) to better recognise the word. However to a computerised handwriting recognition system, an illegible word is one that is

incorrectly recognised either as a result of the word not being in its lexicon or the recogniser not having been trained on that style of handwriting. For example if the recogniser is unable to find the features present in the word or there are an inadequate amount of features to look for in the first place then the recogniser will be unable to uniquely separate out the word from others in the lexicon.

In the example of the children's handwriting, shown in figure 2.3, it can be seen that most of the characters are formed correctly, but are incorrect in relation to the other characters. For example, the first word reads as 'stoPPed'. All seven characters have been written correctly, although both of the 'P's have been effectively capitalised as they are both above the base line of the word. This transforms the shape of the word and some of the features that are used to recognise the word have been lost. In this case the two vertical bars of the 'P's do not descend below the base line resulting in the loss of two main features known as 'descenders'¹. Thus, children's handwriting will affect the recognition rate of a system that has predominantly been trained and optimised for adult handwriting. This is explored further in chapter 4.

In summary then, it has been shown that handwriting is highly variable even though it is possible to classify style into only nine types formed from just 52 characters. It is the actual written interpretation of the characters, their order and the impact on neighbouring characters that makes handwriting recognition a particularly complex problem.

¹ These features will be discussed in more detail along with the holistic recognition system in chapter 3

The following sections will discuss handwriting recognition in general, its application and how it can be improved through the use of context and confidence measures.

2.3 General Handwriting Recognition

Offline handwriting recognition is the recognition of an image that has been written using a pen on paper and that has to be digitised before recognition. This is in contrast to online recognition where the user writes onto digital paper or by using a digital pen. Online digital images are therefore dynamically generated and the temporal information contained within them is preserved. This temporal information gives an additional dimension to the writing thus enabling more accurate recognition of the images [74,75,105,113,125]. Using the temporal information, commercially viable results have been gained and products are currently on the market that can efficiently recognise online handwriting [125]. It will be shown in the following sections of this chapter that, without this extra dimension, the recognition rates of offline handwriting becomes poorer as it becomes increasingly difficult to find the characters and features in words that are used to distinguish ambiguous images from each other. Therefore offline recognition is still very much an unsolved problem and this work will focus specifically on offline handwriting recognition.

There are two fundamental approaches to recognising offline handwriting:

- Character Based Recognition
- Word Based Recognition

A number of pattern analysis techniques have been employed for both of these approaches. The predominant techniques are Artificial Neural Networks (ANN), Hidden Markov Models (HMM) and Fuzzy Logic along with other rule based classifiers. It is not the author's intention to give a detailed account of how each of these techniques work but to show how they can be applied to handwriting recognition. Please refer to the references for a detailed understanding of ANN [2,8,9,13,17,49,143,145], HMM [12,19,23,53,71,107,112,142] & Fuzzy Logic [39,103] based recognition systems.

2.3.1 Character Based Recognition

Character-based recognition is an approach in which individual characters are segmented from a word image and then recognised separately. The character segments can be evaluated against a lexicon produced from the model feature sets of the 26 characters used in the English written language (this can be increased to 52 characters if the uppercase set is included). A list of most probable (or most confidently) recognised characters can be produced for each character segment. These lists can then be used to create a letter graph that can be traversed and analysed using a dictionary to form a list of possible words (see figure 2.4).

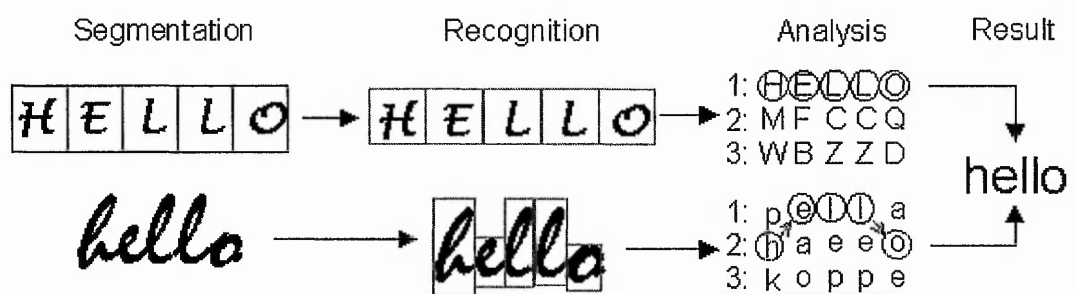


Figure 2.4: Simplistic view of character recognition

This means that a lexicon of just 26 (or 52) characters could be used to recognise every word in the English language. The importance of this is explained further, when the merits and drawbacks of the word recognition approaches are discussed in section 2.3.

Intelligent Character Recognition has been used commercially to recognise deliberately separated characters on forms for the past decade [1,29,46,90,91,146]. The forms indicate that block capital letters must be used and only one character per box. A typical example of this is shown in the top example of figure 2.4. It can be seen that the segmentation of the characters has been made simpler by constraining how the user can respond.

However, character-based recognition is also being researched for recognition of less constrained handwriting [43,48,76,82,122,133]. To do this the segmentation points within the words must first be found before the individual characters can be recognised. This segmentation-based recognition employs similar character recognition processes to that used in ICR, but the recognition is highly dependent upon the segmentation process being 100% accurate. For example, in figure 2.5 and 2.6 the words '*clear*' and '*minimum*' has been written. The word '*clear*' could easily be mistaken for '*dear*' if the segmentation points are incorrectly positioned. Semantic analysis of the whole recognised sentence containing the word '*clear*' would improve the chances of determining whether or not the word is actually '*clear*' or '*dear*' in both word and character-based recognition. Semantic analysis is discussed in section 2.4 along with other post-processing methods, which attempt to improve the recognition. On the other hand, '*minimum*' poses a more complex problem. The best and required result is '*minimum*' segmented into seven characters [81,

82, 83]. However, it is easy to see why it is difficult to know when to stop segmenting and in the final attempt at segmentation there are 16 characters. This is the major weakness in segmentation-based character recognition systems as the ligatures (a ligature is a feature that connects two written characters) are highly ambiguous and, depending upon the writer, highly variable.

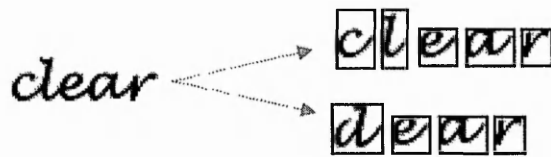


Figure 2.5: Example of ambiguity within written words

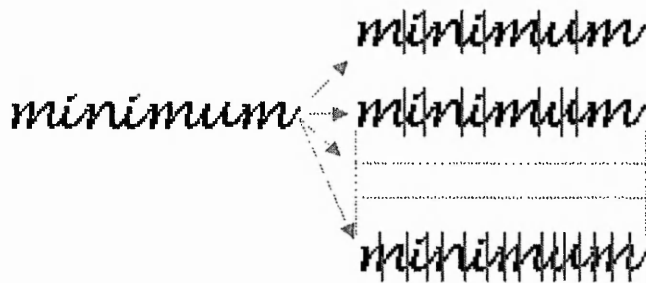


Figure 2.6: Example of the complexity when segmenting a cursively written word

2.3.2 Holistic Word Recognition

Word recognition is the process of recognising the whole word image [51,78,79,102,115,116,131]. Features such as ascenders & descenders as well as the word, length, height and shape of the word are all used to match the image to a set number of word templates held in a lexicon. The lexicon must hold all the possible words

that may be passed for recognition. If an image is passed for recognition then the system will try and recognise that image as one of the word templates in the lexicon and the best match will be produced (see figure 2.7).

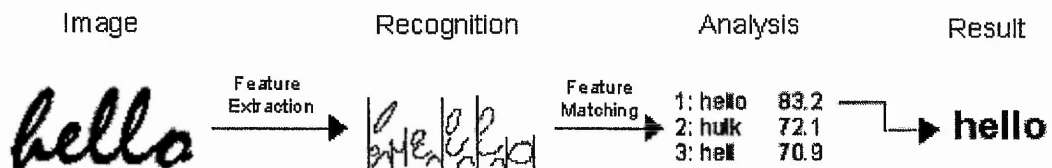


Figure 2.7: The holistic word recognition method. The extracted features consists of vertical bars, holes and cups

2.3.3 Handwriting Recognition Methods

Table 2.1 suggests that the problem of recognising handwritten numerals, characters and words is close to being solved. However, the recognition rates (Rec. Rate) reported are heavily dependent upon specific test conditions such as the size of lexicons (Lex. Size), the style of handwriting and the number of writers tested.

1 st Author	Rec. Rate (%)	Lex. Size	# Images	Rec. Type	Rec. Method
Flemming	~99.5	24	100	Char	Gaussian Classifier
Kavallieratou	98.8	10	2000	Num	Radial Histograms
Bunke	98.4	150	3000	Word	HMM
Kimura	98.0	10	2998	Word	Dynamic
Koerich	98.0	10	4764	Word	HMM
Kovacs	97.0	26	12000	Char	ANN
Chen	96.8	26	2000	Char	HMM
Senior	96.2	10	1016	Word	Hybrid
Garris	92.0	10	1434	Char/Num	ANN
Kimura	91.5	1000	2998	Word	Dynamic
Soon-Man	90.4	100	5297	Word/Num	ANN
Procter	88.8	713	2031	Word	HMM
Senior	88.3	30,000	1016	Word	Hybrid
Vinciarelli	83.6	1334	1016	Word	HMM
Cai	64.6	14	113	Word	Fuzzy
Evans	62.4	200	1987	Word	Fuzzy

Table 2.1: The state of the art in general handwriting recognition

As stated early in this section, many methods have been developed to try and solve handwriting recognition. Methods include: HMMs, ANNs and other classifiers that are trained on particular features to distinguish the characters or words. At the moment, there is no exceptional method that outperforms the rest and all have shown the potential to solve the problem of handwriting recognition. Also, there no a single way in which each method can be applied or combined to attempt to solve the problem.

Kovacs presents a novel architecture for character recognition using three ANNs in parallel [76]. Three Multi-Layer Perceptron networks were trained to classify three unique features that related to pixel positions, contour orientation and the bending points. Creating pair-wise vectors as inputs to the networks produced results of 97% accuracy on uppercase characters and 89% on lowercase characters. An alternative to using an ANN is the use of HMM's. Chen describes a method of using different HMM architectures to recognise cursive characters under "real world" conditions [28]. Each handwritten character is segmented into quarters and graphemes are extracted to build a model of the word that can be analysed by the HMM. The authors accomplished this with varying degrees. Recognition rates between 74% and 98% were reported depending on the size of dictionary used. In addition, this method assumed 100% reliable segmentation of the characters from the cursive word image in the first instance. Garris *et al* describe an approach whereby character images are passed though Gabor filters before being recognised by an ANN [48]. Gabor functions reduce random image noise and smooth out irregularities in the image structure by acting as spatially localised low-pass filters. The Gabor filter proves to be effective in developing training sets that have improved

generalisation resulting in an improved recognition rate of 92% for the ten numerals. The three previous examples of character recognition used the lower case character set. A block capital recogniser has been designed by Flemming & Hemmings in which the characters are expressed as “lines and junctions” [43]. These features are measured and identified in terms of their location within the character. The measurements taken are then applied to a Gaussian statistical classifier. The results show that 93% of the test set was classified with only 0.65% error i.e. 99.5% recognition rate on the classified words. The ability to reject low-confidence recognition results becomes even more important when the importance of not passing on a misrecognised handwriting outweighs the need to recognise every image passed for recognition. This is acknowledged further in section 2.3, when handwriting recognition is applied to solve a real world problem such as reading the address on a piece of mail.

In contrast to character recognition, Soon-Man *et al* present an holistic approach for recognising touching numeral pairs as a whole (word) image rather than attempting to segment them [123]. Instead of trying to recognise each segment as one of 10 classes (the numerals 0-9) a lexicon of all possible pairs has to be created prior to recognition. In this case, 100 classes have had to be defined to accommodate all the possible images that could be passed for recognition even if the images are never passed. Choi *et al* employ a Neural Network formed from a series of modular ANNs [30]. Each of the modules is a representation of one of the numeral pairs. Thus the output of each module is only two neurons, a positive and a negative neuron. The image data is passed to all modules and the output neurons fire. If the positive neuron fires then that module has classified the

image as belonging to the specific numeral pair that the module has been trained to identify. If the negative neuron fires then the image is not the specific numeral pair the module is trained to classify. If both fire (or both neuron outputs are equal) then the module is not confident in recognising the numeral pair as either and thus cannot give a result. If all modules fire their negative neuron or both neurons then the numeral pair is rejected as no classification can be given.

Bunke *et al* uses HMMs to recognise cursive handwritten words producing a recognition rate of 98.4% [23]. Using a 150-word lexicon a skeleton graph was produced for each word to identify features in the shape of the word. A HMM was then used to find these features in the handwritten images and recognise the word. All of the written words in the test set could be found in the 150-word lexicon and the writers were given exact instructions as to how to write the words. Even though this experiment was highly constrained, it was concluded that in a situation where the lexicon is limited and the writing could be controlled then a commercially viable solution could be possible.

Procter *et al* also describe a method for recognising the whole word image using Hidden Markov Models (HMMs) [107]. The authors present a method of holistically recognising written words using a target window to extract features from characters within the words. When training each HMM, the authors extract features from words using a window that is only one pixel wide and 128 pixels tall. The training set is ground truthed so that each character that is observed within the window is known. Therefore the features can be used to identify characters. The HMMs are then trained using these features and tested on

2031 cursively written words. A recognition rate of just 49.8% was presented for this experiment. The authors then showed how this could be improved by adding contextual knowledge into the recognition stage. In this instance, the contextual knowledge was taken from the lexicon and instead of relying on the best HMM model to be observed at each point the lexicon was used to govern which HMM could be selected at any given time thus adding context to the recognition. The recognition rate reported was improved to 93.4% when the lexicon was directly used in the recognition stage. Further to this, the authors also applied a post-recognition word rejecter based on the probability of the overall HMM model not being in the lexicon. This probability was calculated using the observed length of the word. Therefore any word that could not be matched to a word in the lexicon would be rejected and flagged to be classified by another process i.e. a manual inspection. The results show that whilst rejecting 19% of the words, a recognition rate of those words that were not rejected was 99.2%. Rejecting words in this way thus showed to reduce the word recognition error by 5.8%. However in this experiment, the HMM was trained on words written by a single source. Therefore it is likely that these results would only be sustainable in a system that is used to test the same writer's handwriting style that the system has been trained on and not previously unseen handwriting styles.

In summary, the major weakness of holistic word recognition is that for every application a large lexicon is needed that holds every possible word that could be written. Therefore there is no possible way to recognise word images that are not in the lexicon. This is in contrast to character recognition based systems where the recognised characters can be

used to recognise any word that can be found in a large dictionary. Potentially this makes character based recognition more efficient and more robust in applications that require a large lexicon. However, the inability to confidently segment a word so limits the performance of such systems that holistic word recognition remains the primary approach in recognising unconstrained offline handwriting [141,144].

2.4 Applied Handwriting Recognition

Handwriting recognition has been shown to have many inherent difficulties that range from coping with a variety of handwriting styles to the complexity of recognising multi-word combinations that cause ambiguity. Constraining the handwriting recognition by means of limiting the lexicon [44] and writer styles [13,88] has been shown to help reduce the errors by rejecting definite misrecognition errors before they are passed on to the subsequent stages.

Unfortunately, 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.

It is envisaged that the constrained nature of assessment structure can be used to improve recognition so that the confident scoring of scripts can be achieved. Assessment is by its

very nature constrained; only a known number of responses are expected to a given question. Thus the number of responses is finite and therefore the size of the lexicon used in recognition can be limited.

Off-line recognition methods such as postal addresses interpretation, cheque amount verification and signature verification are also areas where the benefits of constraining the scope of recognition can be shown to enhance recognition. These are the three main areas of application reported in the state of the art surveys on constrained handwriting recognition [101,140] and they will now be discussed briefly in order to show how the problems of applying handwriting recognition in each of these areas can be related to automated assessment.

2.4.1 Postal Address Interpretation

A Postal Address Interpretation (PAI) system is used to recognise the layout and handwritten address on a piece of mail in order that the system can route the mail automatically. Table 2.2 (on the following page) shows a summary of the state of the art in PAI systems. The main problem faced in PAI is that an address may produce a number of different recognition problems [30,78,130]. These include postcode recognition, touching digit recognition and all the problems of recognising mixed case, uppercase and slanted handwriting, as well as dealing with noisy images [127].

Authors	Error Rates	Recognition Yields	Specific System Information
Srihari [129]	3.2%	40%	Address encoding System.
Kim <i>et al</i> [69]	7% - Home Address 23% - Business Address	100%	1095 real mail address. Address encoding System
Ding <i>et al</i> [36]	9%	100%	1000 real mail address. Japanese Character recognition using slant correction on an existing projection distance classifier.
Brakensiek <i>et al</i> [18]	14%	100%	416 sized lexicon. 2000 images tested using a Hidden Markov Model.
Blumenstein <i>et al</i> [17]	42%	100%	11 writers tested. Upper Case Characters using a Neural Network.

Table 2.2: The state of the art in PAI

Some problems have already been solved whilst others still need more research in order to perfect the techniques. For instance, the recognition of upper-case characters has been essentially solved for 20 years. Fleming *et al* developed a method for recognising block capitals (upper-case characters written in individual boxes) and reported response yields of 93% with only 0.65% error in 1983 [43] (see previous section). However, recognition rates decrease as characters become touching, misaligned and noisy (Blumenstein *et al* [17]). Despite this, highly accurate results are being reported with the use of context to help the recognition of the whole address. Syntactic knowledge held within the structure of the address is exploited. By using cross-referencing between the different lines of the address, it can be encoded (usually with a barcode) for future processing even when the recognition rate is poor (Srihari *et al* [129] and Kim *et al* [69]). The knowledge contained within the responses that are acquired in AA could also be utilised to aid recognition, as

the constraints of assessment usually require a formal structure to the responses that is related to the question set used in chapter 3.

2.4.2 Cheque Amount Verification

On a cheque there are several fields that require handwritten recognition [34,50,66,95,96,99,124]. These include the courtesy amount (digit type), the legal amount (word type), the payee's name and the account holder's signature. Signature verification introduces other, more specific, complexities and is discussed separately in section 2.4.3. Table 2.3 shows the state of the art in recognising legal amounts on cheques.

1 st Authors	Error Rates	Recognition Yields	Specific Recognition Information
Kornai [72]	2.8%	50%	Legal amount
Suen [134]	6%	90%	Legal amount
Tay [136]	16%	100%	Legal amount
Di Lecce [33]	17%	100%	Legal amount

Table 2.3: The state of the art in CAV

Digit recognition has been the main focus of work over the past ten years [101] and when applied to Cheque Amount Verification (CAV) error rates of 6% have been achieved with just 10% of the responses being rejected (Suen *et al* [134]). The recognition of the legal amount has also shown respectable results with error rates of 17% using contextual *a-prior* knowledge [33] and 16% using Hidden Markov Models [136], both producing 100% response yields. However, when the recognition of the legal and courtesy amounts are combined, the error rates are lowered to 2.75% (Kornai *et al* [72]) albeit at the

expense of the response yield, which falls compared to that gained from the recognition of the separate amounts, i.e. from ~80% to 50%. The drop in response yield is a product of the introduction of confidence measures, which are used to evaluate the results from the recognition of the two amounts. These measures reduce the error rate by rejecting all but the most confidently recognised response. Thus, the error rate is reduced at the expense of the response yield. The main advantage that is obtained by combining the two amounts is that the system becomes more robust as it is less reliant upon just one aspect of the cheque.

In a similar vein, accuracy is also paramount in an automatic assessment system. The system must reject a response for manual assessment rather than pass on a misrecognised response for automatic scoring. Therefore, it should also be possible to use confidence measures that sacrifice response yield in order to reduce error rates in such a system (see chapters 4 & 5).

2.4.3 Signature Verification

Off-line Signature Verification (SV) has a slightly different goal to that of the previous two applications [101]. The focus of SV is to match the signature to a known signature and to verify that the writer of the signature is the same writer as that of the template. Therefore there is more emphasis on the False Acceptance Rate (FAR). This rate measures the number of signatures that are recognised as a writer's signature even though they were actually written by another writer i.e. a forgery. False Reject Rates (FRR) are

also recorded. These show the number of correct signatures that have been classified as forgeries. Table 2.4 shows the state of the art in SV. Both rates show that this is a very difficult problem. In particular they show that the FRR needs to be brought down in order to match the number of actual forgeries found in manual bank cheque verification e.g. less than 0.03% [59]. Indeed, rather than improving efficiency, any misrecognition of correct signatures may actually increase the workload.

Authors	False Acceptation Rates	False Reject Rates
Baltzaks <i>et al</i> [9]	9.8%	3%
Fang <i>et al</i> [41]	16.4%	18.1%
Al-abbas [2]	0 – 40%	10%
Herbst [59]	10 – 23%	6%
Wessels <i>et al</i> [145]	15%	-

Table 2.4: The state of the art in SV

This problem is similar to that faced in AA. When marking a response it is possible to mark a response wrongly, i.e. marking a correct answer as incorrect or incorrect answer as correct response would produce a recognition error comparable to that of the false acceptance/rejection rate of the signatures. Therefore, when automatically marking a response, the false reject rates must also be recorded and analysed. This idea about comparing an image against a known template can also be used in automatic assessment. See chapter 3 where a specific word assessment technique is described based around this principle.

2.5 Post-Processing for Improved Recognition Performance

Improving the accuracy of the recognition by analysing the recognition results is vital if a solution to the problem of handwriting recognition is to be found. As seen in the previous section, the context in which the handwriting is recognised can aid the recognition. In this section, a more detailed account of specific methods for improving handwriting recognition using confidence measures is given. First of all there is a review of speech recognition systems to show that confidence measures have been successfully employed to improve the recognition rates there and then a look at how confidence measures have been applied in handwriting recognition. Other post-processing methods for improving handwriting recognition have been designed, the most established being Natural Language Processing (NLP).

2.5.1 Confidence Measures in Speech Recognition

Speech recognition is not too dissimilar to handwriting recognition, for the reason that when a spoken word has been recognised it becomes a *computer-ready* response in the same way handwriting recognition produces a *computer-ready* word responses from a scanned image. Therefore, the post recognition techniques used in speech recognition systems may be applicable to handwriting recognition [11,55,56,65,113,137,142]. Table 2.5, on the following page, shows the state of the art in post processing techniques and confidence measures in speech recognition.

1 st Author	Post-Processing Non	Post-Processing With	Response Rejected	Post-Processing Method
Hazen	71.5%	81.4%	17.1%	Multi-tiered confidence measures employed at the phonetic, utterance & world level
Bernardis	92.9%	95.3%	50%	Confidence measures based around an ANN is employed to improve a HMM speech recogniser
Sen-Segundo	Error rate reduced by 14%		53.2% of errors detected and rejected	Applying an ANN to confidently analyse the speech recognition system results.
Weintraub	54.2%	61.9%	0%	Applied a novel method to combine knowledge sources via an ANN
Chase	63.3%	73% 95%	27% 87%	Use of a confidence annotator to determine the correctness of the recognised response

Table 2.5: State of the art in applying confidence measures to speech

Hazen *et al* presents a paper on how to employ multi-tiered confidence measures at the phonetic, utterance & world level to improve the overall accuracy of a speech recognition system [55]. The confidence measures take into account the instances in which the input pattern is either noisy (corrupt or non-speech inputs) or not in the hypothesis (the speech recognition equivalent of a lexicon). This is extremely important especially when the user is free to give any form of input i.e. in most real world case.

In the example given by Hazen *et al*, the user asks for information about their local weather report via a telephone [56]. The possible ways people can construct the sentence prompted Hazen to calculate a probabilistic confidence score based on many of the input

features and evaluate each response against the hypotheses. Responses that are rejected are those that do not score high enough using a predetermined confidence equation.

The equation consists of a Gaussian density function that has been trained using the raw scores generated after applying a minimum classification error to the recognised features in a training set. This increased the recognition rate to 81% from 71.5% with the loss of only 17% of the responses.

Another method of improving the recognition results is to employ a second 'high powered' recogniser to evaluate the raw results that are gained from the initial recognition stage. Bernardis [11], Sen - Segundo [113] and Weintraub [143] all make use of an ANN to improve the results gained from a HMM based speech recognition system. Bernardis follows a similar approach to that of Hazen, in which a probabilistic confidence score is calculated for each recognised response. However, an ANN was employed to generate the confidence score from the HMM recognition results that had been normalised. Normalisation of the HMM results was necessary as the ANN required a structured input set (i.e. a known number of input patterns). Using the ANN, the recognition rate was improved by 2.4% to 95.3% at the expense of rejecting 50% of the responses. This is an expensive price to pay to improve the recognition rate by only 2.4%, however in terms of reducing the amount of errors that are not detected this is perhaps a reasonable price to pay if a commercially viable system can be produced.

This is further seen in the results published by Chase [26], in which the author utilises a decision tree to automatically decide if a spoken word in a response is recognised correctly or not. At each branch of the decision tree there is a feature vector that evaluates whether or not the feature has been correctly recognised. The tree is organised so that the previous decision branches have more importance in the path through the tree. This means that those features that are deemed to be more important must be placed higher in the tree.

The importance of the features was measured by calculating the reduction in cross-entropy when the feature was used to evaluate a training set. If a feature were able to reject all the incorrectly recognised words then it would have a cross-entropy value of 1, conversely if the same feature allowed all errors to pass then it would gain a value of 0.

When the decision tree was applied to the test set the resulting output was passed through a set of thresholds. The 'softest' threshold that produced the least rejection increased the recognition rate by 10% to 73% with 27% of the responses being rejected. The 'hardest' threshold increased the recognition rate to 95%, however 87% of the responses had to be rejected.

In section 2.4 it has been shown that the context of the applied handwriting recognition task can improve recognition results and in this sub-section it is shown that the use of confidence measures also allows errors to be reduced thus also improving the recognition

results. In the next sub-section it will be shown how confidence measures have been applied to handwriting recognition and the improvements that have been gained.

2.5.2 Post-Processing Measures in Handwriting Recognition

The performance of handwriting recognition systems has been improved by the use of inbuilt context within the applied task. It will be shown in this section that with the use of syntactic and semantic knowledge [14, 15, 20, 31, 35, 85, 88, 110, 111, 117] the ambiguity within handwriting can be minimised and the overall recognition accuracy improved.

Table 2.6 summarises what can be achieved when post-processing has been applied to handwriting recognition.

1 st Author	Post-Processing		Responses Rejected	Post-Processing Method
	Non	With		
Jobbins	61%	71%	0%	Utilising Roget's Thesaurus to identify semantic relations between word pairs to improve confidence
Rose	69.8%	76.7%	0%	Syntactic and semantic knowledge applied to the recognised response
Procter	93.4%	99.2%	19%	Use of static thresholds based on the style of a single writer
Morita	70.6%	88.1%	0%	Word verification of HMM word models using posterior probabilities

Table 2.6: State of the art in applying post-processing to handwriting

Jobbins *et al* utilises Roget's Thesaurus to identify semantic relations between word pairs to improve confidence [63]. Their work describes a technique that identifies semantic relations using a thesaurus. This technique generates a relation weight that provides a measure of the relation between word pairs or bi-grams. The semantic information can be then used to select between multiple recognised responses to produce an enhanced semantic response. Using this approach an increase of 10% was reported on unconstrained handwritten sentences.

Rose *et al* [111] employs syntactic and semantic knowledge that is applied to the recognised response to form a robust recognition system. Instead of a thesaurus being utilised the semantic relations between words is formed by comparing each word's dictionary definitions. The technique then proceeds by assessing the definition of each word in the recogniser's alternative word list and counting the number of words in common with the definition of each of its neighbours across the sentence. Once the whole sentence has been assessed the words with the highest common words are deemed to be the correct words. This produced a 7% improvement on the recognition of handwritten sentences.

Procter *et al* use static thresholds based on the style of a single writer to reject those recognised words to which the systems assign a low confidence [107]. An improvement of almost 6% was reported. However, this was at the expense of 19% of the responses being rejected. Procter justifies this by stating that it is better to detect and reject an unconfident response rather than allowing it to be automatically processed. The

improvement shown was a direct result of rejecting misrecognised responses rather than recognising more responses correctly. The cost of processing a misrecognised response over rejecting a correctly recognised response is much higher and extremely undesirable, this is an important point to re-empathise.

The work of Procter *et al* [107] and Rahman *et al* [40] underlines the fact, also expressed in section 2.4, that rejecting unconfident responses will improve recognition. Therefore, for the automatic assessment of handwritten responses to be successful it must exploit this by the use of the inherent context within assessment or by means of an augmented confidence measure to the handwriting recogniser. Or both.

Morita *et al* [92] uses Hidden Markov Models (HMM) to recognise written words using character models. Morita *et al* go on to show that the word recognition can be improved by introducing a word verifier. The word verifier is based on the computation of the probabilities of correctness of the recognised characters in the list of alternative words (hypotheses) created by the HMM and then re-ranking the list in the hope that the correct word will be promoted to the top. The character models and the word verifier improved the overall recognition by 17.5%, from 70.6% to 88.1%.

Pitrelli *et al* [100] use a Multiple Layered Perceptron (MLP) neural network to recognise characters. This is then augmented with several confidence measures to try and improve the accuracy of the system. Pitrelli *et al* tested eight individual confidence measures, the main three being the raw recognition rate of the MLPs, a derived likelihood ratio and the

estimated posterior probability of the hypotheses. The eight confidence measures individually gave a slight improvement over the raw recognition rate. However, the best overall improvement occurred when Pitrelli *et al* used four additional MLP that each used the eight confidence measures as inputs. Pitrelli *et al* then introduced a fifth MLP to smooth the outputs of the four additional MLPs. The results published show that with the addition of the MLPs and using all eight of the confidence measures the raw recognition could be improved by 9%.

2.6 Summary

This chapter summarises and discusses the state of the art in handwriting recognition. The problems associated with the automatic assessment of handwritten responses are considered when compared to the problems faced in other applications of handwriting recognition. Constraining handwriting recognition, by limiting the size of the lexicons and by using context, can improve the response yield of an automatic recognition system. However, if handwriting recognition is to be used to automatically assess cursively written scripts, the error rates reported here would be too high. Industry demands response yields of 50% with a maximum error rate of 0.5% [128]. One possibility would be to sacrifice response yields to improve error rates. This is acceptable in an AA environment because it is better to reject scripts for manual marking than incorrectly score the scripts.

Chapter 3 outlines a method for carrying out this methodology on multiple choice style questions. It will be shown that the use of contextual knowledge can allow only the most

confidently recognised responses to be automatically assessed. Chapter 4 then describes another method for assessing single word responses when contextual knowledge is not present and chapter 5 shows how this can be extended to the assessment of single sentence responses.

CHAPTER 3: PRELIMINARY INVESTIGATION TO TEST THE VALIDITY OF AUTOMATICALLY ASSESSING HANDWRITTEN RESPONSES

Chapter 2 has shown that there are many handwriting recognition based tasks that could benefit from automation. Postal address interpretation and cheque processing are two areas where automation has reached commercial viability, due principally to the highly constrained nature of the handwritten responses that they have to deal with. The tasks that are more difficult to automate are those that involve the recognition of unconstrained responses that require a level of semantic understanding and recognition performance that is currently unattainable.

In this chapter a preliminary investigation is carried out into the automatic assessment of adult handwritten responses to multiple-choice questions. The style of questions that are to be assessed can be seen in figure 3.1 and a sample of the completed test scripts can be found in appendix A. Each test script contains eight multiple choice style questions [60].

6) In computing, what does FTP stand for:

**File Transfer Protocol
Fixed Text Post
Forced Termination Premise**

Answer

FILE TRANSFER PROTOCOL

Figure 3.1: An example of a completed test question

This style of question was chosen due to its highly constrained nature. The aim of this investigation is to produce a set of baseline experimental results that will give an indication of how well an automatic assessment system can tolerate the errors introduced at the handwriting recognition stage. It can be seen from the example questions that for each question the possible answers are listed below the question itself. Therefore, the scope of the responses that are to be expected has been tightly constrained so that the candidate must give one of the known answers to that specific question. This limits the size of the lexicon that needs to be used when recognising the handwriting and this in turn helps maximise the recognition performance.

Of course an obvious alternative method for automating the assessment of this style of question would be to label each response with the letters A, B & C and then ask the candidate to tick or mark a box corresponding to the answer they believe to be correct. However, that form of assessment methodology cannot be extended to assign a free text response, which is the ultimate aim of this research.

The main restriction that arises from any type of recognition based assessment system is that it needs to be extremely accurate. The price of recovering from a recognition error (i.e. automatically scoring a misrecognised response) is much higher than any savings that can be gained from automatically scoring a correctly recognised response. Therefore automatic assessment systems must be able to refer a response for manual assessment when the probability of making a mistake is deemed to be too high, i.e. the system should only assess a response when it is known that the recognition system is highly confident of the recognition being correct. It is envisaged that the constrained nature of assessment can be used to improve recognition so that

the levels of scoring scripts can be achieved that are similar to the levels of performance reported for other applied applications of handwriting recognition.

The automated assessment methods described in this chapter and subsequent chapters are run on a platform that incorporates existing technologies. The architecture of the system is described in the following section. This includes all of the processes that are required to achieve full automation.

Once the experimental system has been described, a set of experimental results based on the multiple-choice question format (seen in figure 3.1) will be presented. This will serve as a good baseline against which the results of the experiments detailed in chapters 4 and 5 can be compared.

3.1 Experimental Platform

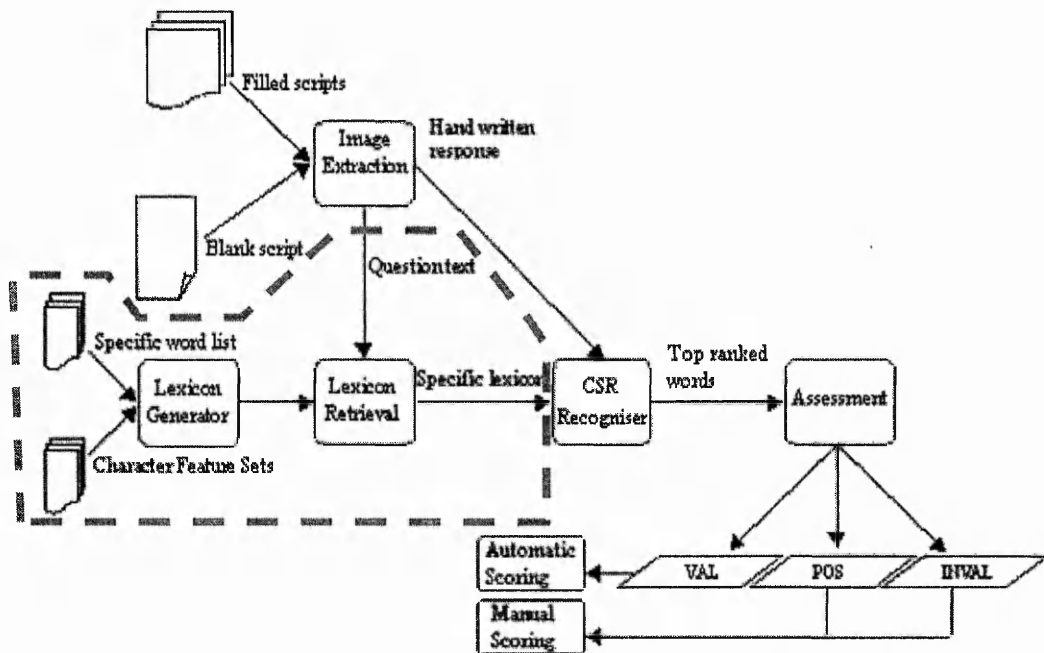


Figure 3.2: Proposed Handwriting Recognition and Automatic Assessment System

Figure 3.2 shows the proposed system. It incorporates existing Image Extraction methodologies, Lexicon Generation/Retrieval techniques and current handwriting recognition technology as well as novel Assessment technologies to automatically score handwritten responses. These areas work independently, as shown in figure 3.2, and are described in sections 3.1.1 to 3.1.4 respectively.

3.1.1 Image Extraction

The use of colour within the test script was felt to be important. Research has shown that data extraction from forms that have a coloured structure is more accurate than that from the more traditional black and white form [137,146]. Thus the script structure and text used in this work was printed in red ink and the students were given

instructions to enter their responses using only blue or black ink. The actual image processing was done using a method devised by Wing *et al* [146]. Wing *et al* also discusses the pro and cons of using colour in the creation of forms and how colour can be used to enhance the image extraction techniques and therefore improve the overall recognition of the system.

The scripts underwent two stages before they were ready for recognition, quantisation and data extraction. The quantisation process reduces the number of colours within the script whilst the data extraction process uses the colour information to directly extract the binarised handwritten responses from the image. Figure 3.3 shows an example of the effect of the quantisation process on one of the completed questions. The text is in red ink and the response was written in blue ink.

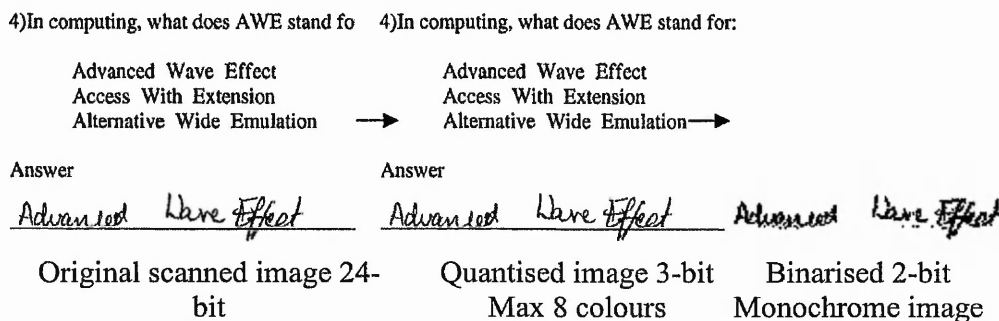


Figure 3.3: The image extraction process

The original image was scanned into the computer using a HP9000C auto feed scanner and was saved as a 24-bit Tiff image. The colour depth of a 24-bit image is ~16.7 million colours and each A4 size uncompressed Tiff image is approximately 13Mb. The quantisation processes reduces the colour depth to a maximum of eight colours. The size of the quantised tiff image was therefore reduced by a factor of 8 and the average size was just 1.5Mb. There were over 100 unique colours in the 24-bit

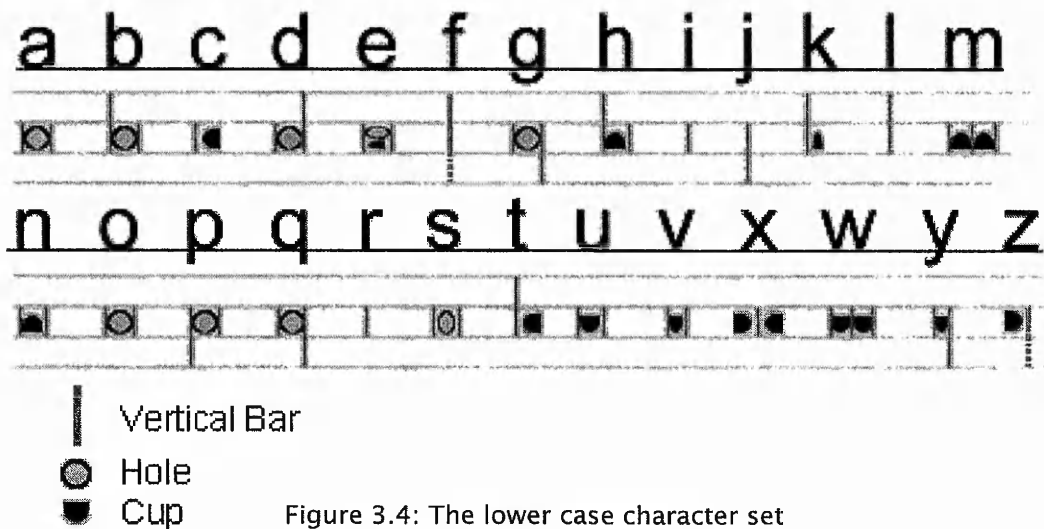
image compared to just 3 in the quantised version, red white and blue. Using the knowledge that the test script comprised of a white background with red text and that the quantised image has an extra blue component it is easy to automatically determine that the response was given in blue ink. Therefore the binarisation processes could turn all blue pixels black and the remaining pixels white thus leaving just the binarised handwritten response.

The binarised image is now ready to be recognised. Each handwritten word is then marked-up and passed separately to the handwriting recogniser. Marking up a word involves identifying all of the pixels that make up a single word from other words and any other erroneous pixels. In a fully automated system this would be done using any of a variety of word segmentation algorithms [82,136]. However, for this work, marking-up has been carried out manually in order to eliminate any errors incurred at this stage in the system, that would further complicate the recognition process and hence affect the automated assessment process.

Of course it is necessary to determine the affects of incorrect word separation and mark-up errors on the overall assessment process but the aim of this research is to develop assessment methods that are tolerant of recognition errors not word segmentation errors. Therefore, throughout the experiments it is assumed that the manual marking up of the word images is perfect and that if it had been carried out by an automatic solution it would also have been achieved with a 100% word mark-up rate.

3.1.2 Lexicon Generator

The holistic word recogniser used in this work was developed in the School of Computing at The Nottingham Trent University by Evans *et al* [39]. This recogniser was chosen due to its availability and proven robustness. However, this work does not try and improve the recogniser but does combine it further with an assessment method to build a strong overall applied recognition/assessment system. The recogniser utilises word zoning and character features to calculate an edit distance score obtained by comparing the feature vector of the images to a set of feature vector templates held within a lexicon. This lexicon of templates is comprised of lower case only words that are created using a character template set based around the three main features that are inherent in the English written language (see figure 3.4). These features are vertical bars, holes and cups. A hole is loop of black pixels that encompass white space and in the lower case character set all of the holes are found in the mid zone. A cup is a line



of pixels that form a concaved shape with an opening that can face in any direction. A vertical bar is simply a column of pixels. These features form the basis of the handwriting recogniser process described in more detail in section 3.1.3.

The character template set is used to build word templates that form the lexicons used by the recogniser. However, the use of a lower case only lexicon was found not to be sufficient to recognise the range of responses found in the test set. Therefore it was necessary to add upper case only and mixed case character sets to the lexicon generation stage. Fig 3.5 shows the upper case (and initial character mixed case) character template set created for this work. It follows a similar format to that used by Evans *et al.*



Figure 3.5: The upper-case and initial mixed case character set

Figures 3.6 and 3.7 show examples of the effects upper case and mixed case writing have on the zoning procedure used by the HVBC recogniser. The zoning process and

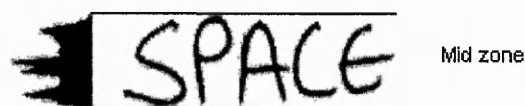


Figure 3.6: Example of upper-case zoning



Figure 3.7: Example of mixed case zoning

subsequent recognition stages are discussed in the following section. It can be seen from figure 3.6 that an upper case only word has just one zone therefore the vertical bars that make up the feature models of each character in an upper-case word must be mid zone only. However, in a mixed-case word the vertical bars of the initial upper case character will be classed as an ascender with a few exceptions. Words such as *All, Of & It* will not produce an upper zone and therefore the vertical bar features of the upper case character must remain as mid-zone bars.

As before the character template sets can then be used to build up a word template for the upper case or mixed case version of each word in the desired lexicon used by the recogniser. This lexicon must include all the words that are expected to be passed for recognition, so that an out of vocabulary word can be recognised and a recognition error will occur. Thus, if a totally generalised holistic handwriting recogniser for the English language were to be hypothesised, then the lexicon for that recogniser must contain every English word. This in practice is impossible, as even humans would struggle to achieve this level of knowledge. Therefore the context of each situation must be considered to determine how large the generalised lexicon must be to include every word that will be expected for that situation. In this instance, what is expected to be written is constrained to just the possible answers to each question on the test script. Therefore the most generalised lexicon created contains just 210 word templates.

This lexicon was formed from the 70 unique words that make up the possible answers to all eight questions along with the three handwriting styles: upper, mixed and lower case for each word.

Table 3.1 lists all of the 210 words expected.

access	connector	extension	jackson	linked	operating	termination
actual	copy	file	java	list	post	text
advanced	data	fixed	joint	machine	premise	the
alternative	development	forced	kernel	mode	protocol	time
asynchronous	device	generally	kit	modulator	quality	transfer
at	domain	geographical	knowledge	natural	query	used
basic	dual	graphical	level	naval	quick	user
blind	dynamic	information	library	network	search	wave
british	effect	interface	limit	of	service	wide
connection	emulation	internet	link	online	system	with
Access	Connector	Extension	Jackson	Linked	Operating	Termination
Actual	Copy	File	Java	List	Post	Text
Advanced	Data	Fixed	Joint	Machine	Premise	The
Alternative	Development	Forced	Kernel	Mode	Protocol	Time
Asynchronous	Device	Generally	Kit	Modulator	Quality	Transfer
At	Domain	Geographical	Knowledge	Natural	Query	Used
Basic	Dual	Graphical	Level	Naval	Quick	User
Blind	Dynamic	Information	Library	Network	Search	Wave
British	Effect	Interface	Limit	Of	Service	Wide
Connection	Emulation	Internet	Link	Online	System	With
ACCESS	CONNECTOR	EXTENSION	JACKSON	LINKED	OPERATING	TERMINATION
ACTUAL	COPY	FILE	JAVA	LIST	POST	TEXT
ADVANCED	DATA	FIXED	JOINT	MACHINE	PREMISE	THE
ALTERNATIVE	DEVELOPMENT	FORCED	KERNEL	MODE	PROTOCOL	TIME
ASYNCHRONOUS	DEVICE	GENERALLY	KIT	MODULATOR	QUALITY	TRANSFER
AT	DOMAIN	GEOGRAPHICAL	KNOWLEDGE	NATURAL	QUERY	USED
BASIC	DUAL	GRAPHICAL	LEVEL	NAVAL	QUICK	USER
BLIND	DYNAMIC	INFORMATION	LIBRARY	NETWORK	SEARCH	WAVE
BRITISH	EFFECT	INTERFACE	LIMIT	OF	SERVICE	WIDE
CONNECTION	EMULATION	INTERNET	LINK	ONLINE	SYSTEM	WITH

Table 3.1: The 210 word lexicon for all the questions

This 210-word lexicon is used so that a direct comparison can be made with the work of Evans *et al*, who evaluated the handwriting recogniser on a 200-word lexicon and achieved a 62% recognition rate.

As already stated, the generalised lexicon contains 210 word templates. Two more specific word lexicons can be produced. The first is a question specific lexicon. For each question the lexicon contains 27 word templates (9 unique words and their case alternatives). Question two is the exception as the word ‘user’ is used twice thus forming a 24-word template lexicon. Table 3.2 lists the 27 word templates that make up the lexicon for the first question.

data	Data	linked	list	List
dynamic	Dynamic	link	library	Library
domain	Domain	level	limit	Limit
DATA	Linked	LINKED	LIST	
DYNAMIC	Link	LINK	LIBRARY	
DOMAIN	Level	LEVEL	LIMIT	

Table 3.2: The 27-word specific lexicon for question one

Question specific lexicons can be used since the response to each question has a defined space on the script in which it can be written. However, the system is dependent upon being able to automatically locate and associate a written response to a specific question with 100% accuracy. Given the large spacing between questions on the test script this level of response segmentation should not be difficult to achieve using existing segmentation techniques.

The most specific lexicon that can be generated is produced when the structure of the responses is taken into account. For each question the writer must choose a response from the list of alternatives given. Since the structure of a valid response is known the lexicon can specifically target each word position. In figure 3.8, a response is shown for the first question.

1) In computing, what does DLL stand for:

Data Linked List
 Dynamic Link Library
 Domain Level Limit

Answer

DYNAMIC LINK LIBRARY

Figure 3.8: An example of question one completed with blue ink

The question has three alternative answers, one of which is correct. Each acronym is three characters long therefore for each character there are also three choices. In the first question, the writer is asked what the acronym DLL stands for. Their choices for the first word position are Data, Dynamic & Domain, hence a word position specific lexicon can be generated containing just the three word choices and their case alternatives. This then produces three separate word-position specific lexicons for the first question each containing 9 words.

Tables 3.3a – 3.3c show the three word-position specific word lists for the first question.

data	linked	list
dynamic	link	library
domain	level	limit
Data	Linked	List
Dynamic	Link	Library
Domain	Level	Limit
DATA	LINKED	LIST
DYNAMIC	LINK	LIBRARY
DOMAIN	LEVEL	LIMIT

Table 3.3a-c: 9 word lexicons for the first three word positions for question one

With the word images extracted and the lexicons generated, the next section describes the holistic handwriting recogniser used in these experiments.

3.1.3 Handwriting Recogniser

The handwriting recognition is carried out using an existing holistic word recogniser designed by Evans *et al* [39]. The holistic recogniser works by recognising the shape of the word from features extracted from the whole word image. Each character is defined in terms of the three features (Holes, Vertical Bars and Cups). Evans *et al* describes this process in detail for the lower-case characters, but it is summarised here for completeness.

The recogniser consists of three main steps; pre-processing, feature extraction and feature set matching.

- i. Pre-Processing

In the pre-processing stage the word image is zoned and the word length and stroke width is estimated. The zoning estimation process involves determining the horizontal pixel density histogram and classifies the word image into a maximum of three zones using four horizontal lines [58,106].

In figure 3.9, the word image 'phone' has been zoned using the horizontal pixel density graph that can be seen to left of the word.

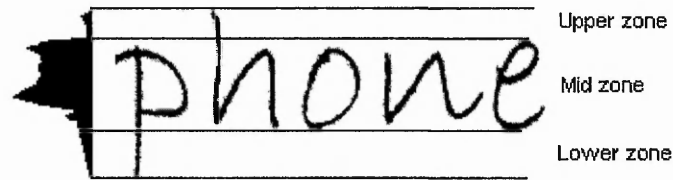


Figure 3.9: Example of lower-case zoning

The mid zone consists of the main body of the word image and if the zoning has been successful then the lower boundary of the mid zone will be the base line of the word image. This base line is calculated using a threshold that is relative to the maximum horizontal pixel density found in the word. Starting from the bottom of the word image and moving upwards, when the horizontal pixel density value exceeds threshold value. Continuing upwards, the base line is set when the pixel count decreases to below the threshold value. The mid zone is defined as the area between these two lines. The lower zone is defined to be from the bottom of pixel histogram to the base line of the word image and the upper zone is defined to be from the top line to the top of the pixel histogram. For a word image to have a lower zone the word must contain a character that has a feature labelled as a 'descender'. Characters such as 'p', 'q', 'g' and 'j' all have descenders. For a word to have an upper zone the word must contain characters that have 'ascenders' for example 't', 'k', 'd' and 'b'. Ascenders and descenders are types of a particular feature, a vertical bar (VB). An ascender is a VB that ascends from the mid zone to the upper zone, likewise a descender is a VB that descends from the mid zone to the lower zone. A VB that remains within the mid zone is labelled as a mid zone bar. For instance the character 'n' should consist of two mid zone bars whereas the character 'h' should have one

ascender and one mid zone bar. The presence of these features in the word image is important to the zoning process and if the features are not prominent in the word image then the zoning process can become corrupt. In figure 3.10 the word phone has been written legibly, however the descender of the character 'p' has not been fully drawn leaving the word image lacking the lower zone. This is crucial because the results of the zoning process is used to extract the remaining features from the word image. The lack of a descender at the beginning of the word image will severely impact on the feature extraction phase.



Figure 3.10: An example of how the zoning can be corrupted

The word length in number of characters is worked out by determining the number of black-to-white crossings in the mid-zone. Knowing the average number of crossings per character, it is possible to estimate the number of characters in the word.

The stroke width is estimated by calculating the distance of each black pixel to its nearest white pixel. Once each black pixel has been evaluated, averaging the values of all the furthest black pixels and then multiplying the result by two gives the estimated stroke width. This information is used in the feature extraction process.

ii. Feature Extraction

This process extracts the features from the word image that are to be matched against the word templates in the lexicon by the feature set matching process. For detailed

information of the techniques used to extract all the features see the work published by Evans *et al* [39]. Each feature is stored in left to right order based on the distance the feature's centre of gravity has from the left-hand edge of the word image. If features overlap horizontally to a significant degree then the features are stored top to bottom. Therefore when the features are extracted from the word image only the features remain and any superfluous information is discarded. Figure 3.11 shows the outcome of a word image before and after the feature extraction.

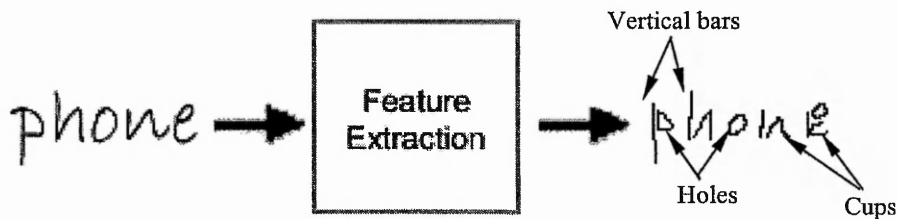


Figure 3.11: The feature extraction process

Holes & Cups are more ambiguous features than vertical bars, especially in cursive script, as handwriting can possess lots of spurious information.

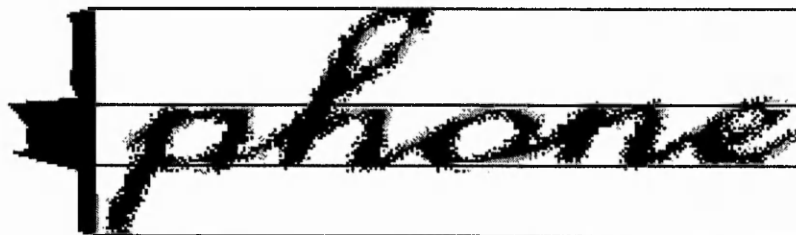


Figure 3.12: The image extraction process

Figure 3.12 shows how loops in characters such as 'h' can introduce spurious holes and how ligatures, the lines that join one character to another, can form additional

cups. Vertical bars are more prominent and less susceptible to corruption, however they can be less prominent in handwriting that is slanted or skewed in some way.

iii. Feature Matching

Using the list of features extracted from the image in step 2, a comparison is made with all of the word templates in the lexicon. For each word template, an edit distance score is calculated that represents how close the word image feature set matches the word template. The closer the match the higher the word template edit distance score a given word template receives. When all of the word template edit distance scores have been calculated the word templates are ordered into a list. The word at the top of the list is the word that has been recognised as the best match to the word image i.e. has the highest word template edit distance score.



Figure 3.13: The feature matching process

Figure 3.13 shows an example of the word image feature set being matched against a word template and the word list that is generated. The actual word template for the word *phone* can be seen in the feature matching process but all of the word templates from the lexicon will be used. The list of the best three word template matches can also be seen along with their respective word template edit distance scores. The edit distance score that is calculated from the feature matching process will be classed as the raw recognition score for the remainder of this thesis.

In this Chapter, only the best-matched words are passed on for assessment. Other methods however, can make use of the alternative word list to implement syntactic analysis if the word is part of a sentence. This is carried out in chapter 5 when part sentences and full sentences are assessed.

3.1.4 Assessment

The assessment relies on the contextual knowledge contained within the required responses. After individual recognition of each word in the three word positions, the best-matched words in each position are concatenated into one response and compared against the three required responses. For example the three target responses for question one are: 'Dynamic Link Library', 'Data Linked List' and 'Domain Level Limit'. A VALID response classification is produced when the recognised response exactly matches one of the required responses. **This classification is irrespective of the correctness of the response.** Recognised responses that do not match any of the three required responses are classified as POSSIBLE or INVALID depending on the number of contextual bridges found within each of the recognised responses. After classification, those responses that have been classified as being VALID can be assessed and scored accordingly. The assessment criterion for question one is simply comparing each VALID response against the correct answer, 'Dynamic Link Library'.

The algorithm, shown below, is employed to classify all of the recognised responses that have the three recognisable word positions (W1, W2 and W3).

Top Word Recognised Response = W1 W2 W3

Classification Rules:

- 1- **IF** W1 is from the same required response as W2
THEN form bridge (W1~W2)
- 2- **IF** W2 is from the same required response as W3
THEN form bridge (W2~W3)
- 3- **IF** W1~W2 **AND** W2~W3
THEN class as **VALID AND END**
- 4- **IF** W1 is from the same required response as W3
THEN form bridge (W1~W3)
- 5- **IF** W1~W2 **OR** W2~W3 **OR** W1 ~ W3
THEN class as **POSSIBLE AND END**
ELSE class as **INVALID AND END**

For example, figure 3.14 shows a response for question six from the test script.

6) In computing, what does FTP stand for:

**File Transfer Protocol
Fixed Text Post
Forced Termination Premise**

Answer

FILE TRANSFER PROTOCOL

Figure 3.14: An example of a completed question

If all the words written in the response shown in figure 3.14, “*File Transfer Protocol*”, were correctly recognised then a recognised response of ‘File Transfer Protocol’ would be produced and classified as **VALID**. However, errors can occur in the recognition stage. If the actual response ‘File Transfer Protocol’ was

misrecognised as ‘Fixed Transfer Protocol’ then a POSSIBLE classification would result. Similarly, if the actual response ‘File Transfer Protocol’ was recognised as ‘File Text Premise’ then this would result in an INVALID response.

It should be noted that neither of these two recognition errors would prove detrimental to the overall assessment process because only the VALID class responses are passed for automatic assessment. The POSSIBLE and INVALID class responses are passed for manual assessment. However, in the event of multiple recognition errors it is possible for the actual responses ‘File Transfer Protocol’ to be misrecognised as one of the other VALID responses i.e. ‘Fixed Text Post’ or ‘Forced Termination Premise’. Under such situations, the misrecognised response would be classified as VALID and would be passed for automatic assessment. This would result in an assessment error because the actual response is correct but either of the two misrecognised response would be marked as incorrect. This is classified as a BAD VALID because the error can only be detected through manual evaluation of the results and not automatically. Table 3.4 shows the possibilities of generating the individual classes when the different word positions have been correctly and incorrectly recognised and the errors that can occur.

	Words Correctly Recognised	Words Incorrectly Recognised	Error
VALID	3	0	None
VALID	0	3	Undetectable
POSSIBLE	2	1	Detectable
POSSIBLE	1	2	Detectable
POSSIBLE	0	3	Detectable
INVALID	1	2	Detectable

Table 3.4: A table to show the possibility of errors in the classified Responses

3.2 Manual Analysis of the Test Set

Fifty first-year computing students completed the test script. These were assessed manually prior to this investigation. The test was not carried out under true examination conditions; however analysing the overall question response spread it can be seen that the questions are suitably difficult as only 53% of the responses were answered correctly (see figure 3.15).

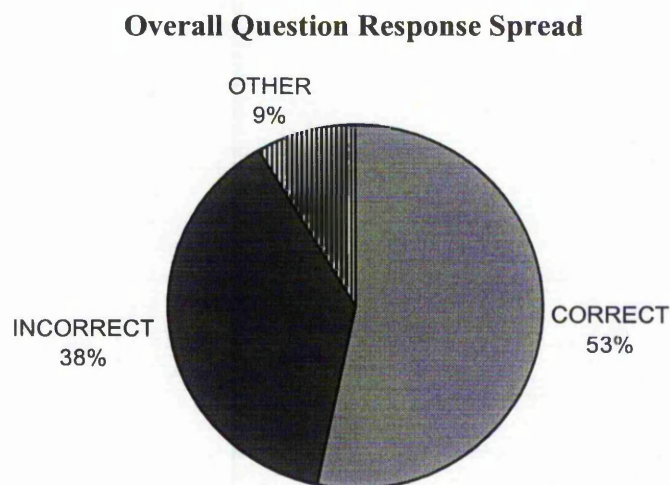


Figure 3.15: A chart depicting the overall response spread

Before processing by the HVBC recogniser was attempted, the number of words comprising each response was manually determined. If the sample contained no response or a numbered response (see fig 3.16) it was manually classified as OTHER before being passed to the recogniser. In principle, these OTHER responses could be automatically pre-classified by the system as it can be programmed to reject answers that are not made up of three independently marked-up word responses. Since manual marking is employed, word segmentation accuracy is 100% and therefore the

rejection rates of the OTHER responses due to ‘numbered’ and ‘no response’ cases is also 100% accurate.

7) In computing, what does QOS stand for:

Query Operating System
Quality Of Service
Quick Online Search

Answer

t

Figure 3.16: An example of an OTHER question due to a numbered response

Writers have, in some cases, opted to abbreviate some of their responses thus giving a word that is not expected and therefore not in the set of templates (Fig 3.17 shows an example of an abbreviated response). These writer errors, unlike the numbered responses, could not be detected by an automated mark up procedure because three independent word images can still be formed. Manual checks have shown that this sort of error amounts to 1% of the total word responses.

8) In computing, what does JDK stand for:

Java Development Kit
Jackson Duel Kernel
Joint Device Knowledge

Answer

Java dev kit

Figure 3.17: An Example of an Abbreviated Response

Further analysis of the test set responses shows that the majority of the handwritten responses were written in mixed case with only a small minority lower case only and

upper case only. Figure 3.18 shows that 60% of the test set responses were mixed case and 16% were upper case thus justifying the need to produce a mixed case and upper case character template set.

Manual Classification of Response Style

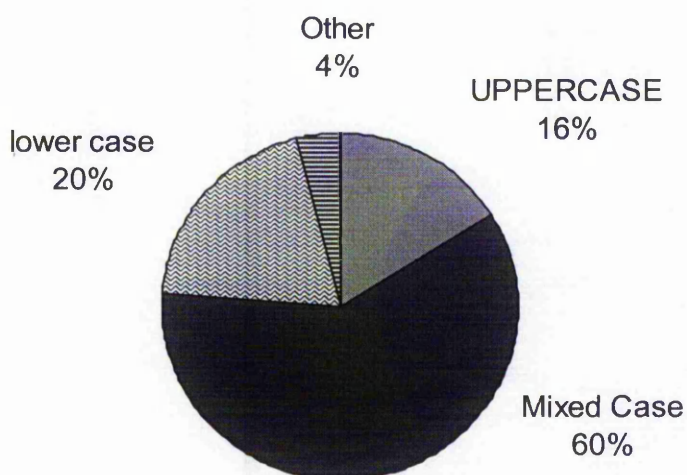


Figure 3.18: The manual classification of style

3.3 Results

The results are reported in two parts. The first part discusses the raw recognition accuracy using the three lexicons, produced in section 3.1.2, and compares them to the accuracy gained by Evans *et al.* The second part discusses the accuracy of the assessment method and whether or not the errors that were introduced by the handwriting recognition stage were dealt with successfully.

3.3.1 Recognition results

Table 3.5 shows the recognition accuracy for each question using the three different sized lexicons.

	Lexicon Size		
	9	27	210
Q1	78.7%	63.1%	40.4%
Q2	93.5%	84.1%	48.6%
Q3	81.1%	68.2%	39.4%
Q4	73.3%	73.3%	38.5%
Q5	84.8%	68.9%	32.6%
Q6	78.5%	68.1%	37.8%
Q7	85.6%	68.9%	41.7%
Q8	90.9%	77.3%	49.2%
Overall	83.3%	71.5%	41.0%

Table 3.5: Raw recognition results

It can be seen that for the 210-word lexicon, the average recognition rate was just 41%. This compares poorly to that of Evans *et al*, who reported an average recognition rate of 63% with a 200-word lexicon. The decrease in recognition accuracy has been attributed to three factors. The first is that the recogniser is highly reliant upon the pixel density. Small words (words with four or less characters) have a low pixel density. Therefore the recogniser has difficulty in extracting features from words such as 'at', 'the', 'time', 'wave' etc. This is the reason that the performance for question five received such a poor recognition rate. The second factor that reduced the average recognition rate is that there is a high ambiguity within the lexicon between a number of word templates [10]. There were a number of words that were not written sufficiently well for the recogniser to differentiate them from other ambiguous word templates. Words such as 'connector' & 'connection' from question three share a very similar orientation of features thus producing ambiguity in the

lexicon. The last factor was attributed to the recognition performance for the upper-case only responses. If these responses were not included in the recognition results the overall recognition rate would increase by 10% to 51% when using the 210-word lexicon. The poor recognition of upper-case words is probably not that surprising as the lack of zoning information produces a larger ambiguity in the upper-case word template portion of the lexicon. This in turn causes difficulty for the recogniser when trying to distinguish between word templates.

However, as expected, constraining the size of the lexicon did increase the recognition rate. The word position specific lexicon increased the overall recognition rate by 40%. This increase is a direct result of minimising the ambiguity within the lexicon thereby eliminating the errors introduced when using the 210-word and 27-word lexicons. However, those responses that contained small words still experienced the errors caused by low pixel density.

3.3.2 Assessment results

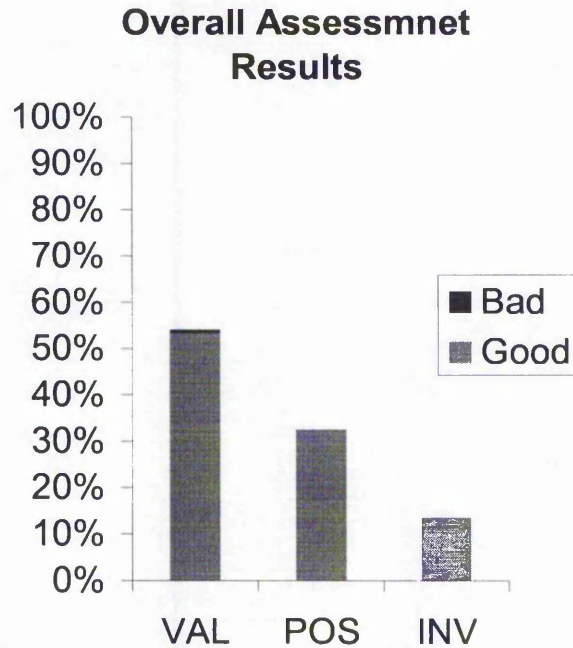


Figure 3.19: Assessment classification results

Based on the findings described in the previous section, it was decided that only the best-matched words from the 9-word template lexicon recognition experiments should be passed for assessment. Figure 3.19 shows the results of the response classification for the recognised responses. It shows that the system classified 54% of all recognised responses as VALID and did this with an accuracy greater than 99%. The 99% of VALID responses were manually classified as GOOD as the system correctly assessed that response. BAD VALIDS occurred when the system assessed a response incorrectly, i.e. a correct responses being assessed as an incorrect answer. The automated assessment of the VALID class responses would therefore also be greater than 99% accurate. This level of accuracy and response yield would be deemed commercially viable in the field of Postal Address Interpretation [128]. The

recognition errors present in the POSSIBLE and INVALID class responses would not affect the assessment accuracy of the system as they could be automatically passed for manual marking.

3.4 Summary

The results of this initial investigation shows that automated assessment using handwriting recognition is feasible. The assessment algorithm introduced in this chapter is seen to be capable of adequately coping with the errors introduced by the handwriting recognition stage (see page 63). Constraining the response of the user is shown to help identify responses that contain handwriting recognition errors thus enabling accurate automated assessment. This is indicated by the overall performance of the assessment system where 54% of all responses can be correctly classified and assessed with an accuracy of 99%.

That said, the 210-word lexicon recognition results, shown in table 3.5, tend to indicate that this conventional lexical approach will not be able to scale up to the sort of large lexicon recognition problems required when trying to assess unconstrained handwritten sentence responses without major enhancement of the recogniser.

In the next two chapters, this prediction will be shown to be correct. However, a specific word verification based recognition and assessment method will be described that can overcome the limitations inherent within the conventional lexical based approach. This will be applied to both children's single word handwritten responses and children's free text handwritten sentence responses.

CHAPTER 4: NOVEL WORD VERIFICATION-BASED ASSESSMENT OF CHILDREN'S HANDWRITTEN RESPONSES

The work in chapter 3 has shown that highly accurate assessment of handwritten responses is possible, provided that the constrained nature of the response is taken into account [3,4]. 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 assessment accuracy, albeit at the expense of a reduction in the response yields.

However, in a situation where there is only a single word in the response no improvement can be gained from such contextual knowledge [5]. If this is further complicated by the need to recognise poorly written word responses then it is probable that a conventional lexical approach will not provide an adequate level of assessment accuracy. This chapter will show this to be the case, and will then go on to show that an alternative word verification based recognition and assessment method can provide the necessary minimum assessment error and sufficient assessment response yield required.

4.1 Single Word Children's Responses

Figure 4.1, on the following page, shows an example of a single word response style exercise used to assess children aged between 5 & 6. This exercise forms part of the "Progress in English 6" written exam, produced by the National Foundation for Educational Research & Nelson publishing company (NFER-Nelson). Many of the questions and exercises are multiple-choice requiring only a single word response

from a given list. In this exercise, the list can be seen at the bottom of the exam page. It should again be noted that labelling the possible responses with a single letter or number (as is possible for the questions set in chapter 3) would not be suitable for this level of assessment given the age of the target audience. More samples of the children's handwritten single word responses can be found in appendix B.

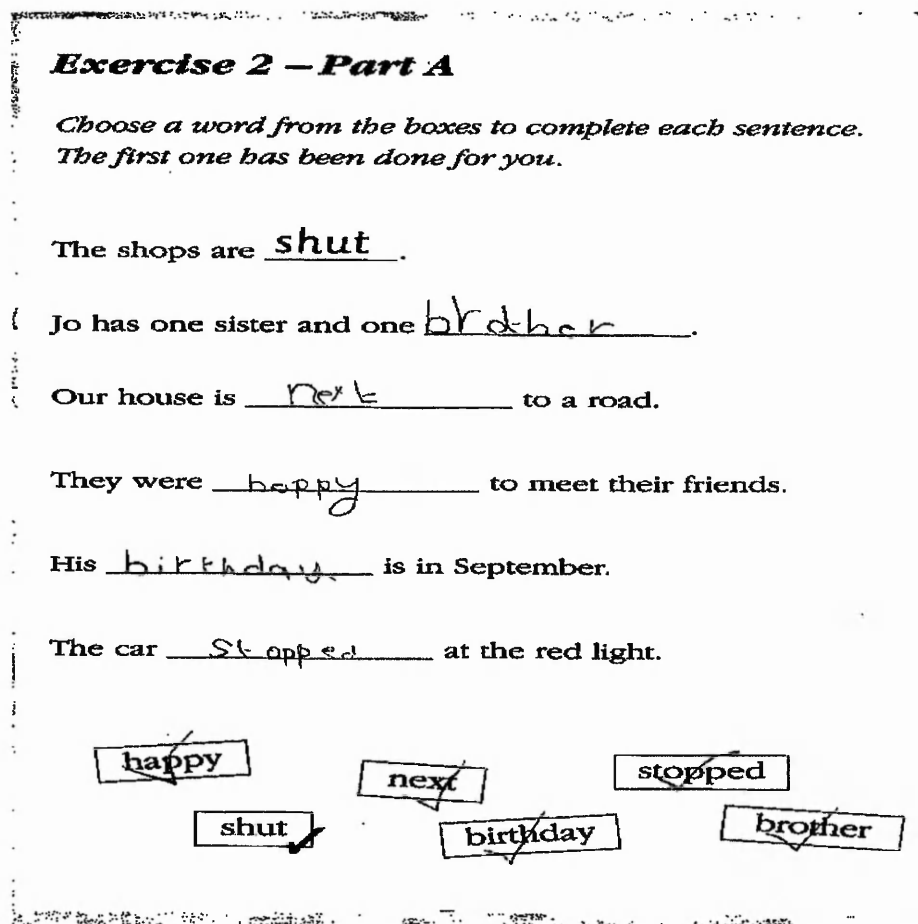


Figure 4.1: Question page from Progress in English 6 published by NFER-Nelson

The actual student responses shown in figure 4.1 also show that the structure of children's handwriting is notably different from that of adult handwriting. Although there are a multitude of individual styles of adult handwriting, the basic word structure is present more often than not, i.e. ascenders ascend, descenders descend and

the shapes of the words are formed. In a child's handwriting these characteristics may or may not be present. The more adept the child is at writing the more structured or 'grown up' the handwriting becomes. On the other hand, the more the child struggles, the more illegible the handwriting. For some children, the acquisition of written expression skills is a difficult and enduring problem [87].

brother
next
happy
stopped
birthday

Figure 4.2a: Example of child 4's responses to Q1-5

badt
bfflf-ee
stoped
jfff
mekt

Figure 4.2b: Example of child 2's responses to Q1-5

Evaluating the data set it is evident that there is a large diversity in the children's handwriting styles. This is shown to great effect in the two samples in figures 4.2a & 4.2b. It can be seen that child 2 shows only a limited comprehension of the exercise and fails to translate their responses in a legible manner. Child 4 on the other hand, understands the exercise and is able to copy and write a legible response from the list of alternatives at the bottom of the page. Characters are formed correctly and spaced neatly and evenly. However, it can also be seen that child 4's writing is of a level where certain characters are incorrectly positioned in relation to the line and its neighbours though legible to a human reader. This characteristic is common in

children's handwriting and can be attributed to the way the children have been taught to write [24,32].

Children are first taught how to form single characters and then taught how to combine them so that they are in proportion and in line to form a word. This developmental approach to handwriting can be seen in the formation of the characters 'r' & 'p' in figure 4.2a, where the height of the letter 'r' is equal to that of the neighbouring 'b' in the word 'brother' and where the letters 'p' in 'happy' and 'stopped' are placed above the line.

Both of these style characteristics were likely to cause problems for the handwriting recogniser used in this work [109]. As documented in chapter 3, the handwriting recogniser relies heavily upon the presence of features such as ascenders and descenders. In cases such as those in figures 4.2a & 4.2b, the zoning process would mistakenly zone the words 'brother', 'happy' & 'stopped'. This would then prevent the system from detecting possible ascenders or descenders in the image pattern leading to eventual misrecognition.

Consequently, for this work, it was found necessary to modify the feature weights in the handwriting recogniser such that it relies less upon the presence of ascender & descender features and more upon other features such as mid-zone bars, holes and cups [39].

4.2 Image Extraction

To extract the binarised image, the same method is used as in the previous chapter. Using Wing *et al*'s quantisation and binarised technique directly on the data set produced some unexpected results. The A4 scripts were scanned in as 24-bit colour Tiff images and passed through the quantisation technique. Over a third of the resulting quantised images were missing the handwritten responses. The reason for this was that all but a few of the completed tests were written in pencil. It was found that the default quantisation method could not differentiate a light grey colour of a pencil from the white script background. This was caused by the thresholds used in this technique being set too low. Figure 4.3 shows diagrammatically the RGB colour model used in the quantisation technique. The eight colours are aligned in relation to the origin (black) and the maximum (white) in terms of their respective RGB intensity. In a 24-bit image each colour component has an 8-bit intensity that ranges from 0 – 255. Therefore when $R=G=B=0$, the pixel is defined to be black and at the maximum, $R=G=B=255$, the pixel is white.

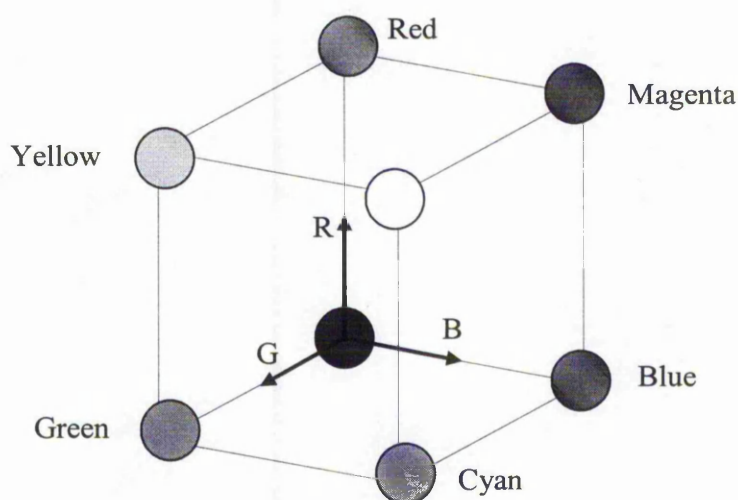


Figure 4.3: The RGB colour model used by Wing *et al* to define the quantisation thresholds

The colour grey is defined where the intensity of the red component is equal to the green and blue components of the pixel tested. Within the colour cube, grey lies on a linear line from the origin (black) to the maximum (white).

In figure 4.4, three grey pixels can be seen: p1, p2 & p3. Pixels p1 & p2 have an equal RGB intensity that is lower than the threshold used in the quantisation technique. Therefore they will both be black pixels following quantisation. Pixel p3 however, will become white as it has a RGB intensity greater than the threshold.

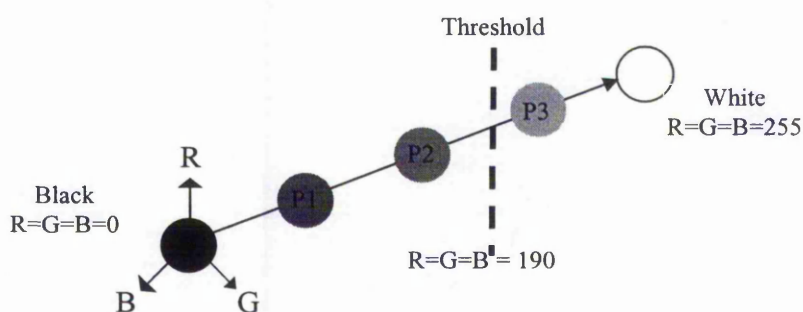


Figure 4.4: An example of how a grey pixel is transformed into either a black or white pixel

Following the experimental practice set out by Wing *et al* an investigation was carried out to find a threshold that would consistently turn the light grey fine pencil responses to black.

The data set consisted of 5 questions that have been completed by 29 children producing 145 handwritten words. Of the 145 words, 56 were lost in the quantisation process. These responses were evaluated and the average pixel intensity was found to be between 195 & 212. Wing *et al* has reported that the highest binarisation threshold

allowable is 215 due to the nature of the light wavelengths used by different scanners. Therefore the image extraction in this chapter used Wing *et al*'s quantisation technique with a black-white threshold of 215.

An unforeseen advantage of increasing the threshold is that the anti-aliasing introduced by the scanner is turned into black pixels along with the true word image pixels. This results in a 'fatter' binarised word image than the original word image. Figure 4.5 shows an example of the 'fattening' of the word images. It can be seen that after the word image has been extracted, the new binarised image is thicker than the original word image. This is an advantage because the feature extraction process relies heavily upon the pixel density of the word image to accurately identify key features in the recognition phase.

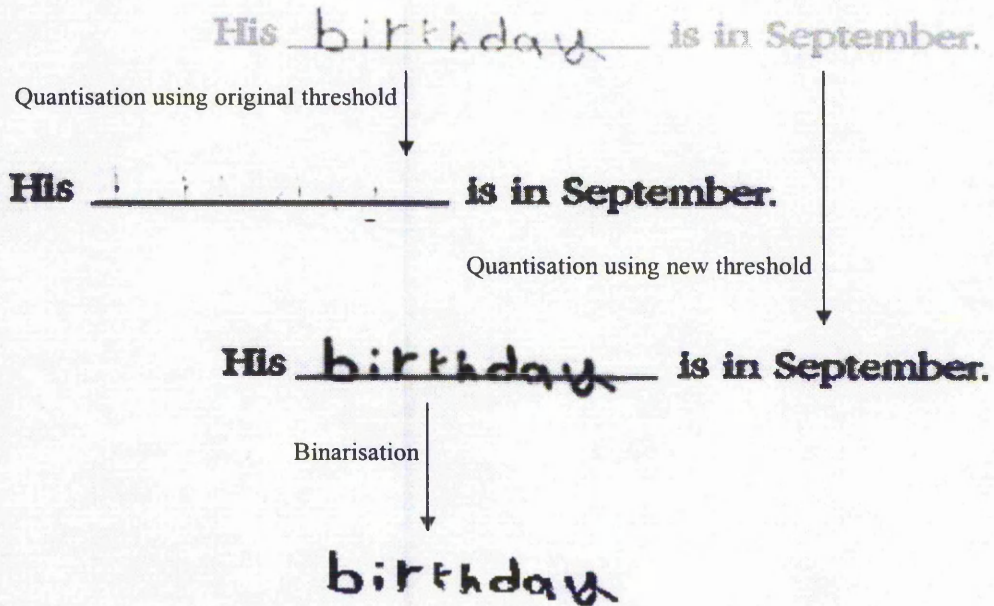


Figure 4.5: An example of how the quantisation process can fail and how applying a new threshold can overcome the fault

It should be noted that both of the recognition and assessment approaches discussed in the rest of this chapter use the binarised images produced by the method described in this section.

4.3 Conventional Lexical Based Recognition & Assessment Approach

In this chapter the automatic assessment of the response format seen in figure 4.1 is attempted. Firstly an attempt to assess the responses using the conventional approach to handwriting recognition is carried out. Secondly, a method of exploiting the constrained nature of assessment is described and compared to the conventional approach is detailed in section 4.4. This novel approach is then applied, in section 4.5, to the adult data set used in chapter 3 so that a comparison of the two methods can be made.

The first method described is the conventional recognition method that is described in chapter 3. A single input image is compared to a list of all possible alternative response templates held within a lexicon. This is shown diagrammatically in figure 4.6.

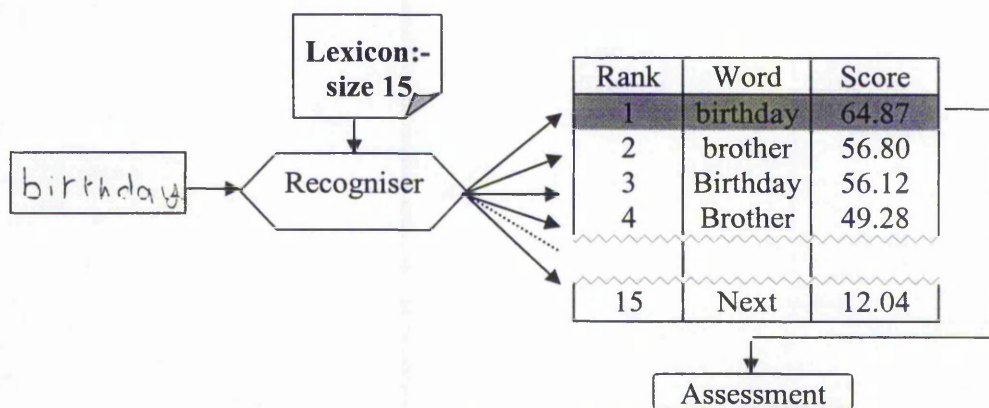


Figure 4.6: The conventional approach to handwriting recognition

In this experiment the lexicon is constrained by what is expected i.e. those alternative answers given at the bottom of the page. The 5 possible answers form the basis of a 15 word template lexicon along with their case alternatives, i.e. lower case – birthday; mixed case – Birthday; upper case – BIRTHDAY etc. After recognition, a single list of alternatives is produced for each response with the alternative words being ranked in order of how close the input image features match the template word. In the conventional approach to handwriting recognition and assessment, the best matched word from the list of alternatives for each question is then passed on for automatic scoring, as described in chapter three.

Scoring is the process of marking and classifying the response as either CORRECT or INCORRECT. Table 4.1 shows the different possible outcomes from the recognition process for this dataset.

Input	Automated		Classification
	Recognition	Scoring	
Correct	Correct	CORRECT	GOOD
	IncorrectA	INCORRECT	BAD
	IncorrectB	INCORRECT	BAD
	IncorrectC	INCORRECT	BAD
	IncorrectD	INCORRECT	BAD
IncorrectA	Correct	CORRECT	BAD
	IncorrectA	INCORRECT	GOOD
	IncorrectB	INCORRECT	BAD
	IncorrectC	INCORRECT	BAD
	IncorrectD	INCORRECT	BAD

Table 4.1: The possible classification outcomes of the system

It can be seen that if a correct handwritten response is the input in this system, then there is the possibility that the recogniser may either recognise it correctly or may

misrecognise it as one of the four incorrect answers. If the recogniser recognises a correct response as correct then the response will also be scored as CORRECT. This can be manually classified as GOOD. However if the recogniser mistakes the correct response for one of the wrong answers then the system will score the response as being INCORRECT. This is manually classified as BAD because it is an unaccountable error within the automatic system and therefore could not be detected without human intervention. In a similar manner, an incorrect handwritten response may be correctly recognised and scored as INCORRECT (classified GOOD) or incorrectly recognised as the correct response (classified BAD) incurring the same unaccountable scoring error. Another recognition error can occur when an incorrect response is incorrectly recognised as one of the other three incorrect answers. In this case, no assessment error has occurred, however the system has allowed a recognition error to be assessed. This is unacceptable for an automatic assessment system, as no confidence could be placed in responses being incorrectly recognised but are still being automatically assessed.

Two variations of automated assessment have been adopted for the conventional approach so that a comparison can be made between it and the novel approach described in the section 4.4.

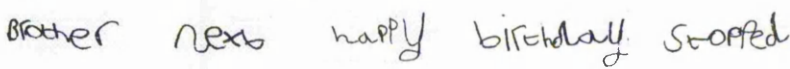
4.3.1 Raw Assessment of the Conventional Lexical Approach

The first conventional lexical approach uses the raw results from the recognition process whereby the top words for each response are simply passed for scoring straight away. Since no classification can be given to the response prior to assessment

all recognised responses are assessed. This is in comparison to the results in chapter 3 where, with the use of contextual bridging, some of the recognised responses could be rejected for manual marking as a result of them not forming valid responses. In this instance, there is no contextual knowledge to assist the assessment. Therefore, the assessment accuracy will be equal to the raw recognition rate.

4.3.2 Results for the Conventional Lexical Approach

As stated in the image extraction section of this chapter, 29 children completed the five questions (shown in figure 4.1) as a part of the Progress in English 6 exam. This gave 145 handwritten responses. A professional human assessor¹ scored 65% of the responses as correct and 35% as incorrect. This was achieved with 100% accuracy. In figure 4.7, an example of the actual written response and the recognition results from the Conventional approach for Child 8 can be seen. This shows the best three matched words for each response. The cost of mis-zoning a word can clearly be seen in figure 4.7, where 'Brother' has been incorrectly recognised as 'next'. This is a direct result of the recogniser finding no ascenders in the word image and therefore classifying the word as mid-zone only. The expense for the recogniser to then reform the ascenders is very high, which can be seen in the recogniser score, and why the mid-zoned word 'next' was been recognised as the best matched word.



Rank	BROTHER	NEXT	HAPPY	BIRTHDAY	STOPPED
1	next 67.3	next 74.0	Next 66.5	Brother 54.6	brother 60.2
2	HAPPY 64.6	HAPPY 65.8	STOPPED 53.6	birthday 53.7	next 58.5
3	happy 64.5	NEXT 64.0	HAPPY 53.3	Stopped 49.1	STOPPED 57.5

Figure 4.7: Example of the conventional recognition results for child 8

¹ The exam papers were pre-scored by an external examiner employed by NFER-Nelson before being used in this work.

Obviously, for this child the conventional approach would only score the response to the second question as correct. The other responses are scored as incorrect even though the actual responses written are correct. These are recognition errors.

Overall, this system achieved a recognition rate of 59% (shown as Good Cor + Good Inc in figure 4.8). This would obviously imply an assessment error rate of 41% if the best-matched words were simply passed for scoring. The actual correct and incorrect responses as determined in the manual scoring is also shown in figure 4.8, where it can be compared to the automatic scoring.

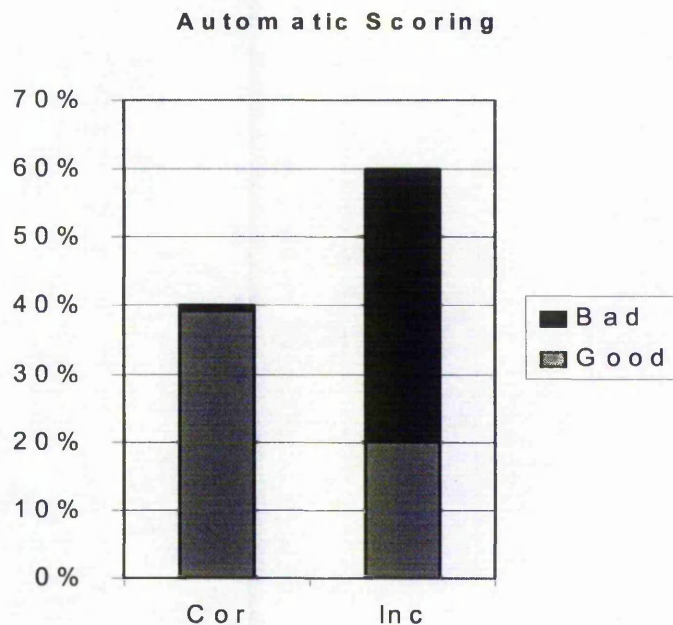


Figure 4.8: Graphs to show the Automatic Scoring in the Traditional Approach

Although the automatic process achieved 99% accuracy for scoring the correct responses, it only achieved 33% accuracy for the incorrect responses. The main reason for this is the inability of the recogniser to cope with handwriting such as that

seen in figure 4.2a & 4.2b. Only 60% of the correct responses and 57% of the incorrect responses were accurately recognised. It is worth noting that if only the responses automatically scored as correct were used for assessment (INCORRECT passed back for manual marking) then this system would produce a reasonable assessment yield (40%) with an assessment accuracy of 99%. However, this level of assessment yield is due to the favourable correct/incorrect ratio (65:35) in the data set used. A less favourable ratio (50:50 etc) would produce a lower assessment yield.

4.3.3 Threshold Assessment of the Conventional Lexical Approach

The second variation of the conventional lexical approach uses a threshold, based on the recognition score, to filter the correctly recognised and incorrectly recognised responses in an effort to minimise the errors produced by the recognition process.

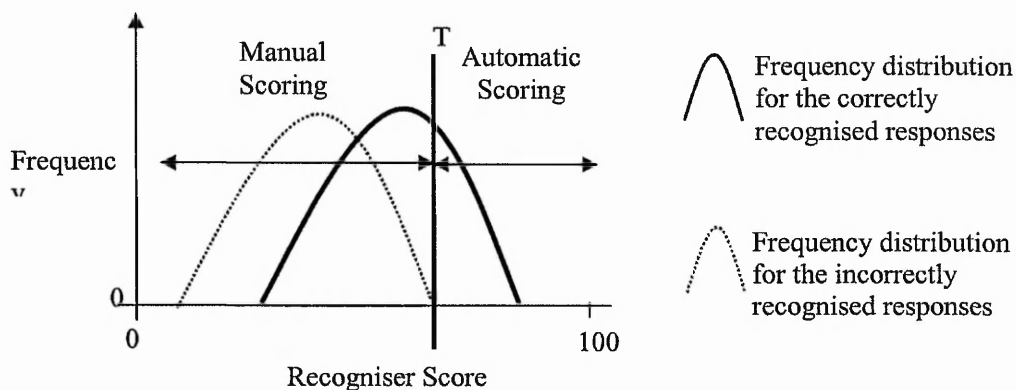


Figure 4.9: An idealised example of a distribution graph showing the frequency distribution of the correctly and incorrectly recognised response scores

Applying a threshold to the associated recognition score of the best matched words is one way in which recognition error can be detected automatically [107]. This threshold can be calculated from a training set by plotting the frequency distribution

graphs of the correctly and incorrectly recognised responses. The threshold (T) is then taken from the point at which there are no further incorrectly recognised responses. An example of this can be seen in figure 4.9. Any best-matched responses with a recogniser score higher than T will be automatically scored and responses with score lower than T will be automatically passed for manual scoring.

Figure 4.10 shows the outcome when three answers are recognised and assessed using the threshold. In the example, each written response has been given in answer to the question 'His _____ is in September'. Using the threshold, the system can automatically classify the first word as confidently recognised as it has achieved a recognition score higher than that of the threshold. This response can thus be automatically assessed. The recognised word can be compared against the correct answer and scored accordingly.

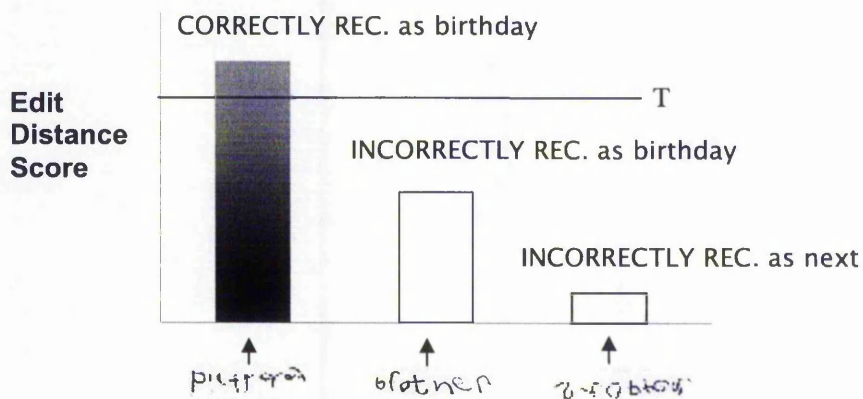


Figure 4.10: Three examples of written response that have been assessed using the threshold for the question "His ____ is in September"

The remaining two responses have scored below the threshold. The system therefore rejects them on the basis that there is no confidence in the recognition result. Using the threshold has allowed the system to recover from the two recognition errors.

Without the threshold both responses would have been automatically assessed with the results that the incorrect response 'brother' would have been scored as correct. Although, 'stopped' was recognised as next and therefore still recognised as an incorrect answer, thus occurring no scoring error, it still is an error and must be taken into account within the systems accuracy.

4.3.4 Threshold Results for the Conventional Lexical Approach

Applying a threshold to the conventional lexical approach involved creating training and test sets. The training set was randomly created using 75% of the dataset with 25% being set aside as an unseen test set. A frequency distribution graph of the correctly and incorrectly recognised response word scores was produced from the training set and a threshold (T) of 81 was determined (see figure 4.11). This threshold was then applied to the test set so That only recognised responses with a recognition score above the threshold were automatically assessed.

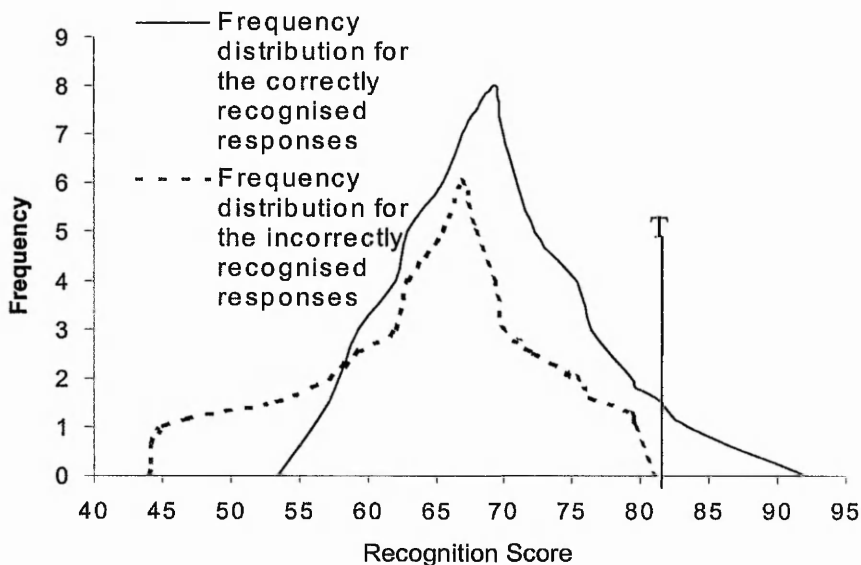


Figure 4.11: Graph to show the smoothed frequency distribution of the correctly and incorrectly recognised response scores and the value of threshold T

From figure 4.12, it can be seen that only 5% of the test set responses could be passed for automatic assessment. However, a scoring accuracy of 100% was achieved with this 5% response yield.

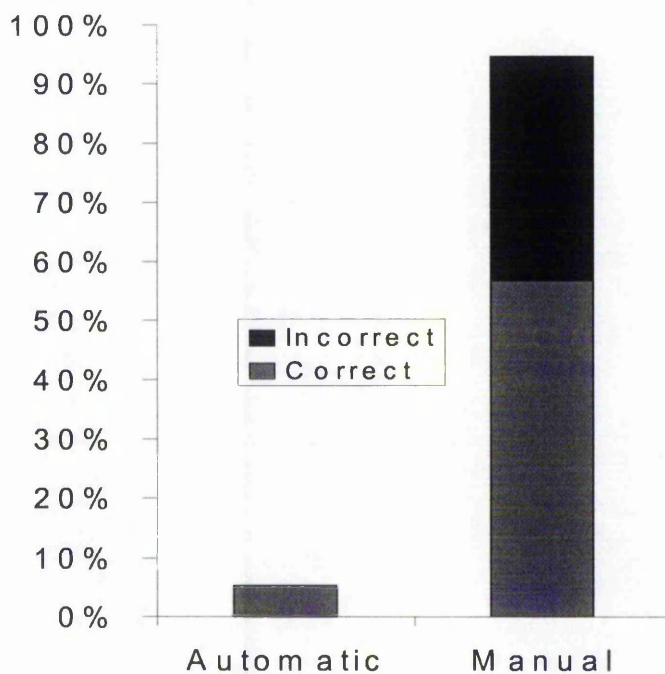


Figure 4.12: Graph to show the response yields after the threshold has been applied

To further illustrate this fact, it can be seen that the application of the threshold to the results shown in figure 4.7 would result in none of the recognised best-matched words being passed for automatic assessment. This means that none of the recognition errors would have been passed for automatic assessment but it also means that neither would be the correctly recognised response 'next'.

In conclusion, the results shown in figures 4.8 and 4.12 show that both variants of the conventional lexical approach have failed to give a satisfactory performance. This is a result of the raw recognition rate being so very poor which is, itself, a direct result of the inconsistency in the children's handwriting. Although the threshold approach produced no assessment errors, due to it being able to overcome the poor recognition rate by only allowing the most confidently matched responses to be automatically assessed, it is too stringent a technique with the result that the overall number of automatically scored responses is too low to justify this method being used to solve the assessment task. The non-thresholded approach on the other hand, is too lenient, with the results that a large number of assessment errors are produced.

4.4 Specific Word Assessment Technique

Specific Word Assessment Technique (SWAT) is a novel approach to handwriting recognition based assessment. It exploits the nature of the question and answer medium by only comparing the input pattern to the template of the correct answer for that specific question. It is known that the reduction in the ambiguity within the lexicon, which results from just having the correct word template, will reduce the number of recognition errors and therefore improve the overall performance of the system. However, this is different to CLA in that the question being asked is not which word in the template set best matches the word image, as in CLA, but how well does the written response match the correct answer template.

4.4.1 SWAT Applied to Children's Single Word Responses

For example, the correct answer to the question "His _____ is in September" is 'birthday'. Therefore, with SWAT, the lexicon used for this question only consists of the feature templates for the words: birthday; Birthday; & BIRTHDAY, see figure 4.13.

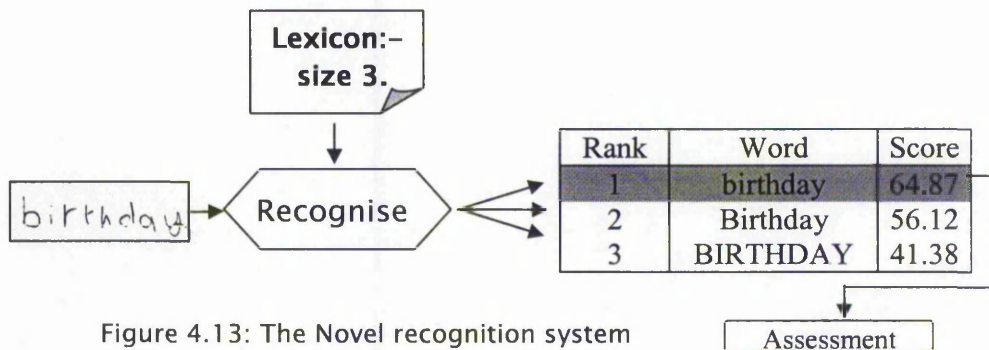


Figure 4.13: The Novel recognition system

Of course, by neglecting the other four alternative responses, this approach will always generate the correct answer as the recognised response. Therefore, the recognition score of the best-matched recognised responses must have thresholds applied to them, in a similar way to the advanced assessment technique used in the conventional approach. A high score implies that the recogniser has achieved a close match between the input pattern and one of the correct word target templates. The system can then confidently score this response as CORRECT. If the recognised response obtains a low score, this means that the recogniser can find few features in the input image that match those in the correct word template. The recognised response is, thus, either a wrong answer or is illegible and it can therefore be automatically scored as INCORRECT. A recognised response that produces a best-match recognition score in between these two thresholds must be automatically passed for manual recognition, as the system is unable to automatically assess the

recognised responses as CORRECT or INCORRECT. These two thresholds now become specific word thresholds, thresholds used for the sole purpose of assessing many word images against a single correct word template. This is in direct contrast to the general recognition threshold employed to strengthen CLA, where one threshold was used to evaluate many word images against many word templates for each question, irrespective of final assessment context.

As in the conventional lexical approach a training set is required to produce the frequency density graph from which the thresholds can be derived. However, this time instead of the scores of the correctly and incorrectly recognised responses being counted the frequency of the best matched words that are actually the correct answer for the question are compiled against those that are not the correct answer. Since this method always produces a correct answer, every incorrect response will be misrecognised. Figure 4.14 shows an example of the graph used to calculate the thresholds used in SWAT.

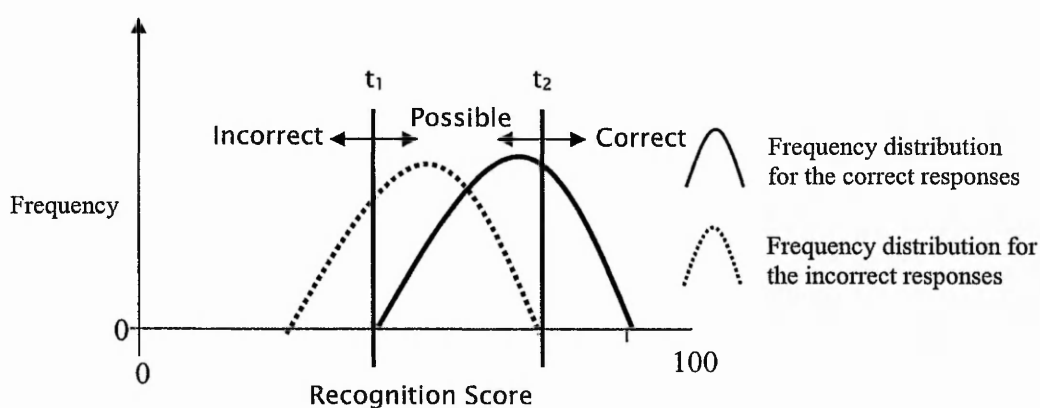


Figure 4.14: An idealised example to show the frequency distribution of the correct and incorrect response scores and the positions of the thresholds t_1 and t_2

Figure 4.15 illustrates how the thresholds are used to classify the three example responses to the question 'His _____ is in September'. It can be seen that the written response 'birthday' achieves a high score that peaks above the upper threshold t_2 , thus allowing it to be automatically scored as correct. The response "Stopped" shares few features with the lexicon templates of 'birthday', 'Birthday' & 'BIRTHDAY' and therefore has a low score allowing it to be automatically scored as incorrect. The response 'brother' on the other hand is a closer match to the 'birthday' template and attains a score that is higher than t_1 but lower than t_2 . This response should not be assessed by the system but must be automatically set aside for manual assessment.

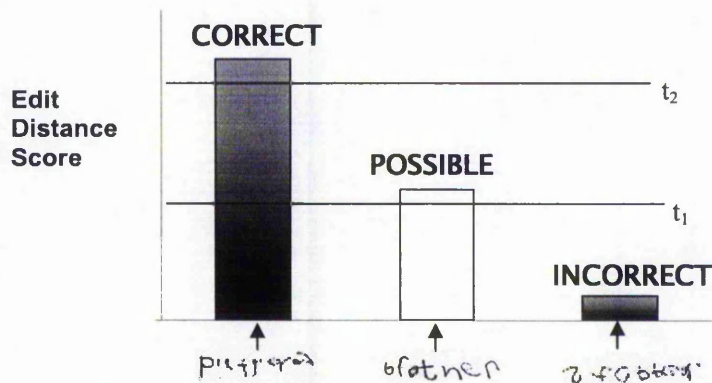


Figure 4.15: Three examples of written response that have been assessed using the thresholds for the question "His _____ is in September"

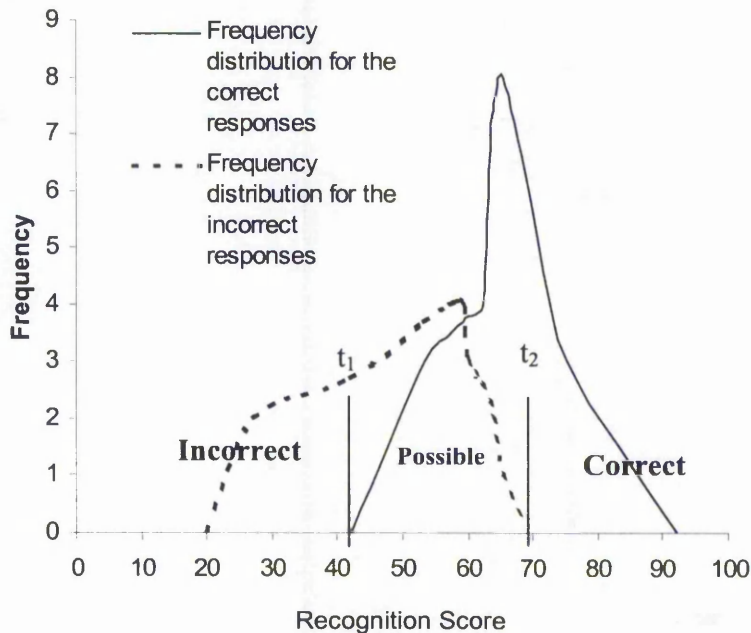
It is worth noting that the thresholded conventional lexical approach can only use a high score threshold because a low score threshold would only identify the definitely incorrectly recognised responses. This would give no indication as to whether the actual response is CORRECT or INCORRECT and since any word that scored below the high threshold is passed for manual scoring the low score threshold is redundant.

4.4.2 SWAT results

The results from the novel approach for Child 8 are shown in figure 5.16. It can be seen that in each list only the correct answer is present. As with the conventional approach, only the best-matched responses are passed for assessment. However in this instance, thresholds have to be applied to the recognised response score in order to classify them as correctly or incorrectly recognised instead of relying on the raw recognition result. Training and test sets used to evaluate SWAT were created in the same proportions as in the traditional approach and the frequency distribution graphs of the correct and incorrect responses were plotted (see figure 4.17).

Rank	BROTHER	NEXT	HAPPY	BIRTHDAY	STOPPED
1	<i>brother</i> 52.7	<i>next</i> 74.0	<i>HAPPY</i> 53.3	<i>Birthday</i> 53.7	<i>STOPPED</i> 57.5
2	BROTHER 51.4	NEXT 64.0	happy 39.7	BIRTHDAY 47.9	stopped 55.0
3	Brother 47.9	Next 47.7	Happy 28.2	birthday 45.6	Stopped 54.2

Figure 4.16: Example of the novel recognition results for child 8

Figure 4.17: Graph to show the smoothed frequency distribution of the correct and incorrect response scores and the values of thresholds t_1 and t_2

From the graph, t_1 and t_2 were determined as 41 and 69 respectively and then used to analyse the test set. The use of the two thresholds produces a higher number of automatically assessed responses than CLA. This is because the ambiguity of the word templates in the lexicon has been removed allowing even poorly written correct responses to be matched to the correct word template without being confused with any other word.

This has a more significant aspect in that SWAT uses the assessment directly to classify the recognition results and therefore classification of each word image is not whether the recogniser has matched it correctly but whether or not the word image is correct or incorrect. Moreover using the two thresholds there is the option to flag responses that cannot be automatically assessed to be passed for manual assessment.

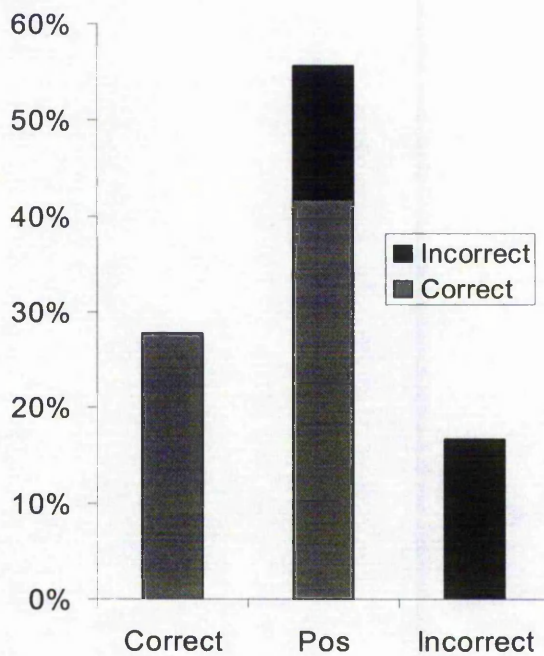


Figure 4.18: Graph to show the response yields after the thresholds have been applied

The three factors, reducing the overall ambiguity, imposing the assessment criteria onto the recognition results and the ability to identify those words that cannot be assessed automatically, makes for an extremely robust system.

Figure 4.18 shows that the novel approach scored 28% of responses as correct and 16% as incorrect without error. This gives an overall response yield of 44% with an accuracy of 100%. The increased response yield is

also a result of using a lower threshold (t_1) such that illegible responses (such as those seen in figure 4.2b), which produced such a mismatch between the image and the correct word template, can confidently be assessed.

4.5 SWAT Applied to Adult's Single Word Responses

Due to the small size of the available dataset that contains children's handwriting samples a single exercise specific threshold was adopted for all five questions in the exercise. To determine whether a question specific threshold would produce better results, SWAT was applied to the dataset containing adult's responses to multiple-choice used in chapter 3. This would also allow a comparison to be made between SWAT and the contextually constrained conventional lexical approach adopted in chapter 3.

The adult dataset contains eight multiple-choice style questions that have been completed by fifty first year computing students. The format of the questions demands that the response to each question contains three words to answer the question correctly.

Using the conventional approach to recognise the dataset with a response position specific lexicon, an assessment accuracy of 99% was achieved with an automated assessment yield of 54%. This is a higher performance than the conventional lexical approach can achieve on the children's responses because the contextual knowledge, present in the adult three word responses, is not present in the children's single word responses.

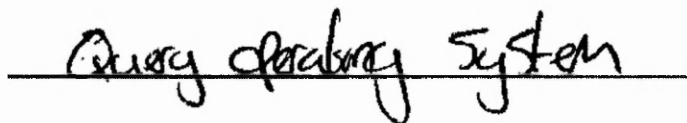
4.5.1 Applying SWAT to Adult's Three Word Responses

Firstly, the dataset was randomly split into training and test sets, again with 75% of the dataset making up the training set and the remaining 25% comprising the test set. However, since the dataset is quite large, this allowed the use of thresholds that were specific to the response word positions within specific questions. This means that instead of using the 9-word template lexicons generated in chapter 3, a new lexicon can be produced that just contains the correct word in the correct word position in each question. For example, the answer to the first question in the adult dataset is 'Dynamic Link Library' (see figure 4.19) therefore three lexicons can be made containing just the case variants of the three individual words that form the correct answer i.e. the lexicon of word position one in question one contains templates of the words 'dynamic', 'Dynamic' & 'DYNAMIC'.

7) In computing, what does QOS stand for:

Query Operating System
Quality Of Service
Quick Online Search

Answer



Query operating system

Figure 4.19: An example of question one from the adult dataset

Using these new lexicons each word position can be recognised to produce the ranked list of how well the recogniser matches the word image to the correct word template.

The recognised response must then have the word position specific thresholds applied to it so that the individual responses can be classified.

Training sets were produced containing 75% of the word images and each set was collated separately so that question and word specific thresholds could be determined. This left a test set containing 360 word responses from 15 writers. For question one the thresholds can be seen in figure 4.20.

$t_1 = 58$	Dynamic	$t_2 = 66$
INC	POS	COR
$t_1 = 52$	Link	$t_2 = 73$
INC	POS	COR
$t_1 = 37$	Library	$t_2 = 56$
INC	POS	COR

Figure 4.20: The word position specific thresholds determined for question one

The low (t_1) thresholds for the words 'Dynamic' and 'Link' were significantly higher than the word 'Library' and higher than those seen in the equivalent low threshold for the children's dataset. This is a result of having specific thresholds for each word and not a general low threshold as needed in the children's test to cover all words. This is a result of not having to accommodate issues such as the misspelling of the word 'Library' (which would be poorly matched to the correct word template thus causing the low threshold) in the score thresholds for the other words. The converse can also be seen with the high threshold for the word 'Link'. This is significantly higher than the other two. The recogniser struggles to correctly match the small words, because finding features is more difficult. Therefore, a high (t_2) threshold is needed to prevent incorrect short words from being mistakenly recognised as the correct answer 'Link'. Using word position specific thresholds allows longer words to benefit from having

specific word position high thresholds (t_2) that are lower thus classifying more recognised responses. Table 4.2 shows the threshold classification of the recognised responses for question one along with the actual assessment results carried out by the human assessor.

Test Writers	Recognition + Threshold Classification for the three word positions for Q1			Actual Assessment
	DYNAMIC	LINK	LIBRARY	
W0	COR	COR	COR	Correct
W1	COR	COR	COR	Correct
W2	POS	INC	POS	Incorrect
W3	INC	POS	POS	Incorrect
W4	COR	COR	COR	Correct
W5	POS	POS	POS	Incorrect
W6	COR	POS	COR	Correct
W7	COR	POS	COR	Correct
W8	COR	COR	COR	Correct
W9	INC	POS	INC	Incorrect
W10	INC	POS	POS	Incorrect
W11	INC	POS	INC	Incorrect
W12	COR	COR	COR	Correct
W13	POS	COR	POS	Correct
W14	INC	POS	INC	Incorrect

Table 4.2: The word position specific classification for question one of the adult three word responses

To assess the recognised responses, three assessment criteria were produced based on how many recognised words gained the correctly recognised classification within a given response. The first of the assessment criteria is the hardest or most demanding criteria and can be seen in figure 4.21a.

IF all three recognised words are classified as **CORRECTLY RECOGNISED**
THEN the recognised response will be automatically scored as **CORRECT**
ELSE IF all three recognised words are classified as **INCORRECTLY RECOGNISED**
THEN the recognised response will be automatically scored as **INCORRECT**
ELSE the response is classified as **POSSIBLE** and must be manually assessed

Figure 4.21a: The Hard Assessment Criteria

By only assessing those recognised responses that have either all words classified as correctly recognised or incorrectly recognised and scoring them appropriately this criteria should produce the least errors as the likelihood of wrongly classifying all three recognised words will be quite small.

The second assessment criterion (figure 4.21b) that will be applied to the adult hand written responses are less demanding and requires at least two words to be classed as either correctly or incorrectly recognised before the responses is rejected.

IF all three all three recognised words are classified as **CORRECTLY RECOGNISED**
OR any two of the recognised words are classified as **CORRECTLY RECOGNISED**
THEN the recognised response will be automatically scored as **CORRECT**
ELSE IF all three recognised words are classified as **INCORRECTLY RECOGNISED**
OR any two of the recognised words are classified as **INCORRECTLY RECOGNISED**
THEN the recognised response will be automatically scored as **INCORRECT**
ELSE the response is classified as **POSSIBLE** and must be manually assessed

Figure 4.21b: The Medium Assessment Criteria

IF all three all three recognised words are classified as **CORRECTLY RECOGNISED**
OR any two of the recognised words are classified as **CORRECTLY RECOGNISED**
OR (any recognised word is classified as **CORRECTLY RECOGNISED**
AND the remaining words are classified as **POSSIBLE**)
THEN the recognised response will be automatically scored as **CORRECT**
ELSE IF all three recognised words are classified as **INCORRECTLY RECOGNISED**
OR any two of the recognised words are classified as **INCORRECTLY RECOGNISED**
OR (any recognised word is classified as **INCORRECTLY RECOGNISED**
AND the remaining words are classified as **POSSIBLE**)
THEN the recognised response will be automatically scored as **INCORRECT**
ELSE the response is classified as **POSSIBLE** and must be manually assessed

Figure 4.21c: The Soft Assessment Criteria

The third assessment criterion (figure 4.21c) is the softest approach in an attempt to assess as many recognised responses as possible without achieving any assessment errors.

4.5.2 Assessment results of the adult dataset when SWAT was applied

As stated earlier, the test set contains 8 multiple-choice questions answered by 15 writers. Only the best matched words from the 3-word position specific lexicons were passed for assessment. Table 4.3 shows the assessment of the test data for question one using the three assessment criteria.

Test Writers	Recognition + Threshold Classification for the three word positions for Q1			Automatic Assessment			Actual Assessment
	DYNAMIC	LINK	LIBRARY	Soft	Medium	Hard	
W0	COR	COR	COR	Correct	Correct	Correct	Correct
W1	COR	COR	COR	Correct	Correct	Correct	Correct
W2	POS	INC	POS	Incorrect	Possible	Possible	Incorrect
W3	INC	POS	POS	Incorrect	Possible	Possible	Incorrect
W4	COR	COR	COR	Correct	Correct	Correct	Correct
W5	POS	POS	POS	Possible	Possible	Possible	Incorrect
W6	COR	POS	COR	Correct	Correct	Possible	Correct
W7	COR	POS	COR	Correct	Correct	Possible	Correct
W8	COR	COR	COR	Correct	Correct	Correct	Correct
W9	INC	POS	INC	Incorrect	Incorrect	Possible	Incorrect
W10	INC	POS	POS	Incorrect	Possible	Possible	Incorrect
W11	INC	POS	INC	Incorrect	Incorrect	Possible	Incorrect
W12	COR	COR	COR	Correct	Correct	Correct	Correct
W13	POS	COR	POS	Correct	Possible	Possible	Correct
W14	INC	POS	INC	Incorrect	Incorrect	Possible	Incorrect

Table 4.3: The Assessment of the recognised words for question one

Eight writers correctly answered the question and seven answered incorrectly. The soft assessment criteria performed the best on the question one set where it was successfully used to automatically assess 14 out of the 15 correctly. The remaining response was passed for manual assessment as all three recognised words were classified as POSSIBLE. The medium assessment criteria automatically assessed 10 of the 15 responses and passed 5 for manual assessment. The hard assessment criteria only managed to automatically assess 5 responses correctly and the rest were rejected for manual assessment. For question one, none of the three assessment criteria

produced an assessment error. Figures 4.22a-4.22c (on the following pages) show the overall results for all eight questions.

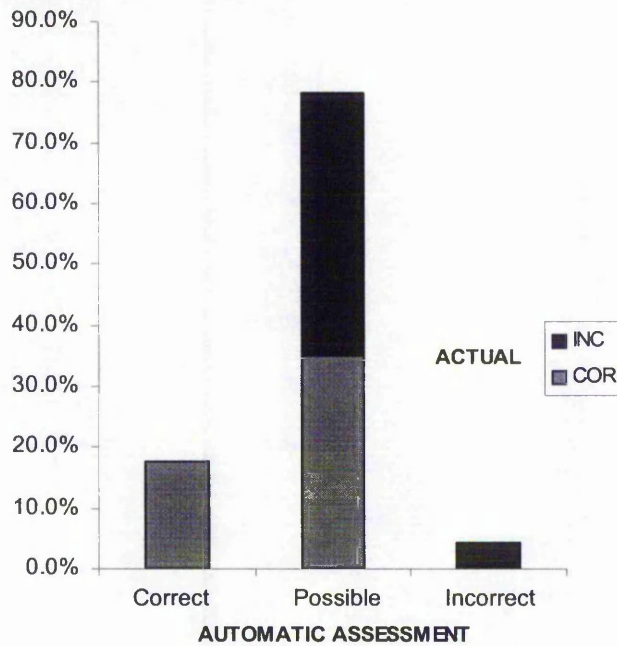


Figure 4.22a: Results of assessing the adult responses using the hard assessment criteria

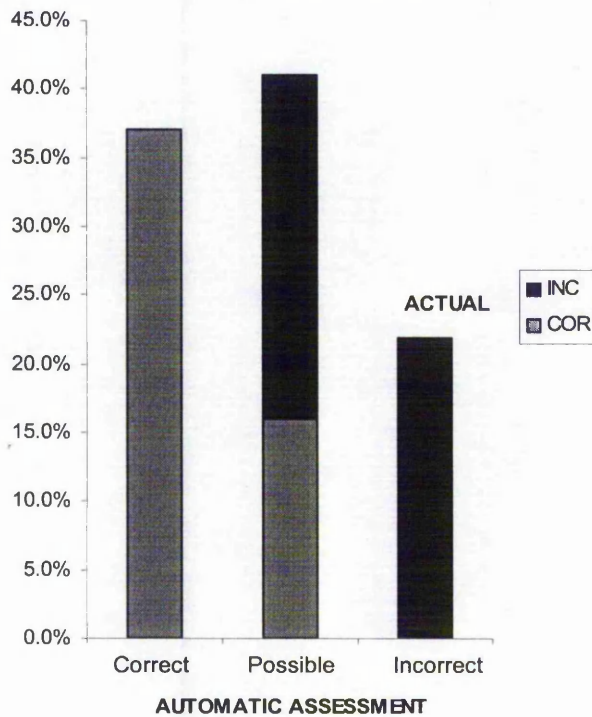


Figure 4.22b: Results of assessing the adult responses using the Medium assessment criteria

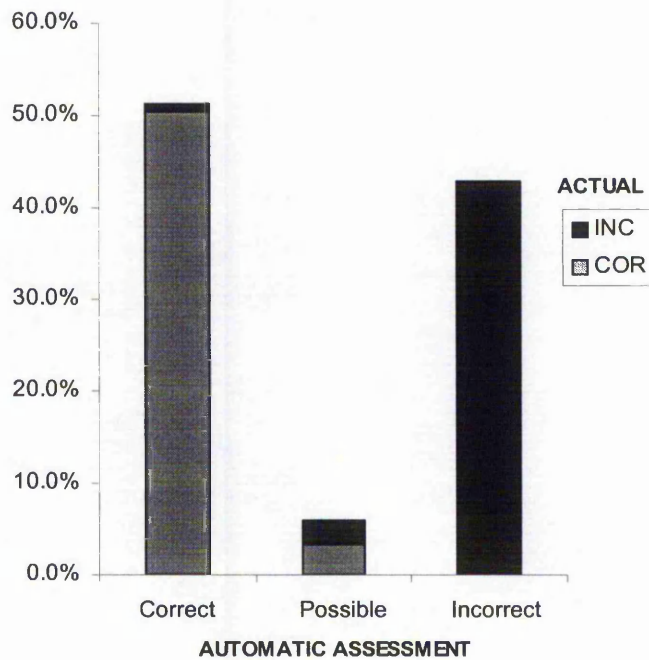


Figure 4.22c: Results of assessing the adult responses using the soft assessment criteria

Figure 3.18 (in section 3.3.2) shows the results of the response classification when contextual knowledge enhanced conventional lexical approach was used to assess the adults handwritten responses. It shows that the system classified 54% of all responses as VALID and did this with an accuracy greater than 99%. In this section, these results will be compared with the results gained from using SWAT to assess the same data.

The best performing assessment criteria over all eight questions was the medium approach, automatically assessing 59% of the responses with no assessment error. The hard approach automatically assessed 22% of the responses with no error and the soft approach assessed 93% but 2% were incorrectly classified thus producing assessment errors. These results show the potential of SWAT. The hard approach was too extreme

thus not assessing enough responses. The soft approach worked extremely well at achieving a high assessment response yield but sacrificed the assessment accuracy slightly to enable this. The medium approach is good compromise as it out performed the contextual knowledge enhanced conventional lexical approach by achieving 100% assessment accuracy compared to the 99% assessment accuracy achieved using CLA and by achieving a higher responses yield, 58% compared to 54%.

4.6 Conclusion

This chapter has introduced a novel approach for the automated assessment of handwritten single word responses. The Specific Word Assessment Technique (SWAT) performed very well when compared to a conventional lexical based approach. Yields of 44% were gained with 100% assessment accuracy using static thresholds on children's single word handwritten responses. This compares favourably to the conventional lexical approach where the low recognition rate of 59% severely compromised the assessment response yield.

The explanation as to why SWAT is more accurate becomes apparent when the frequency density graphs are examined. In the novel approach the response images are only checked against the correct answer. This reduced the ambiguity within the lexicon and increased the separation between the incorrect and correct response edit distance scores thereby increasing the response yield. Since the scope of recognition was narrowed using SWAT, a lower threshold could also be used to identify actual incorrect responses thus increasing the response yields further.

Retrospectively, applying SWAT to the adult responses to the multiple-choice questions, used in chapter 3, has also shown that a word verification based recognition and assessment method can outperform a contextual knowledge enhanced recognition and assessment system.

In the next chapter it will be shown that with the addition of contextual knowledge, children's handwritten sentence responses can be confidently assessed using SWAT.

CHAPTER 5: AUTOMATED ASSESSMENT OF CHILDREN'S HANDWRITTEN SENTENCE RESPONSES

This chapter compares a number of approaches for automatically assessing handwritten sentence style responses based around the two main approaches used in the previous two chapters

The first approach to be tested is the Conventional Lexicon Approach (CLA) to the automated assessment of children's handwritten sentence responses. As described in chapter 3, CLA is an approach that generalises the problem of handwriting recognition such that it is potentially able to recognise all possible responses and as a result it will give a recognised response every time. The second method employs the Specific Word Assessment Technique (SWAT) to evaluate each word in the written response against only a model answer. This technique is described in detail in chapter 4. If the technique is not confident about a word that has been recognised then SWAT will not assess the response but will reject it for manual assessment. In this chapter it will be shown that these two approaches can be tailored to automatically assess a handwritten sentence response. In addition, two further methods will be added to CLA and SWAT to determine whether a more robust automatic assessment methodology can be formed.

The first augmentation to the approaches is to apply a Natural Language Processing (NLP) technique to the recognition results before assessment is carried out. NLP can be applied to sentence responses as they have semantic and syntactic knowledge inherent within them. It has been shown in chapter 3 how contextual knowledge can

be used to evaluate the recognised responses more intelligently thus improving the automatic assessment accuracy. The NLP technique used to augment the two main approaches employs contextual word bridges to help interpret the recognition results. This additional method is applied to CLA and SWAT slightly differently as it is dependant upon the lexicon used in each approach. For CLA, contextual word bridges are used to produce the best-fit sentences from the raw recognition results that are produced using a generalised lexicon. For SWAT, the best-fit sentence cannot be found as the lexicon contains only the correct answer therefore keyword pairs are used along with their specific word position order to identify contextual keyword bridges. The use of contextual bridges is discussed further for CLA in section 5.3.1 and for SWAT, in section 5.3.2. The use of bridges will hopefully minimise the amount of recognition errors being passed over for assessment by providing a syntactic structure that will reduce the ambiguity in the raw recognition results.

The second additional method to the system is the use of a question's response history (past answers) to improve the contextual knowledge held within the lexicon so that a more robust assessment can be carried out. A question's response history can be used to assess the recognised responses for both of the two main approaches. This will be shown to be particularly useful for the SWAT approach as the model answer alone will be seen to be insufficient to assess the responses as accurately as required. This method is discussed in more detail in section 5.4.

5.1 Children's Sentence Response Assessment

Figure 5.1 shows the modified automated assessment system (previously seen in chapter 3). It now incorporates the knowledge of an 'assessment expert' in the form of an assessment database. This database is 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. In both CLA and SWAT the assessment process is dynamically linked to the assessment database as each assessment is related to the question being assessed 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.

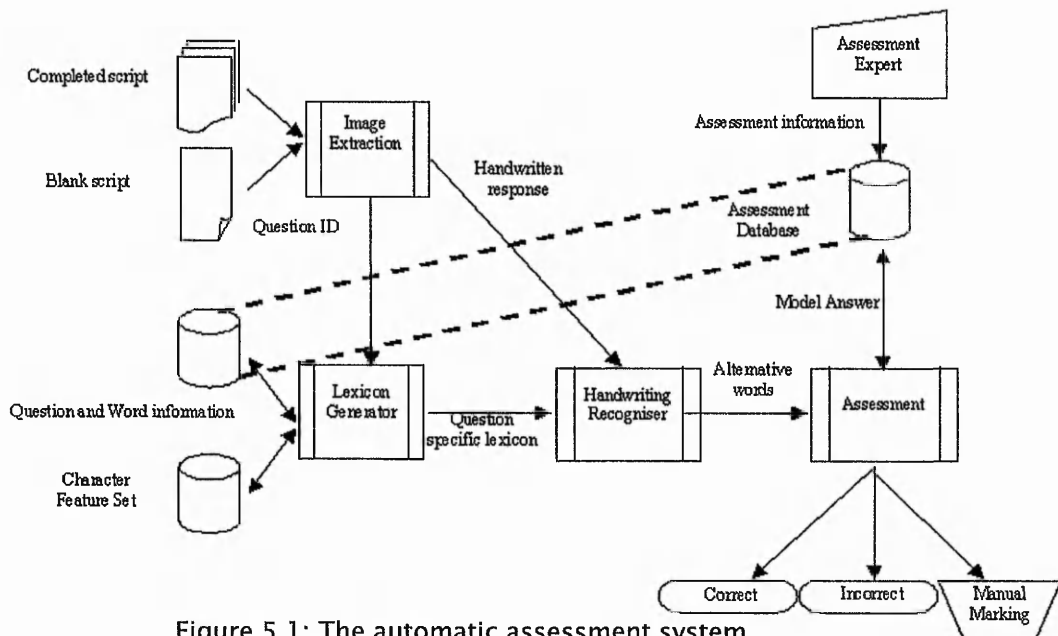


Figure 5.1: The automatic assessment system

A semi-automatic Image Extraction process was used in the experiments in this chapter, by which means 100% of the handwritten responses were extracted correctly and passed for recognition. The handwriting recognition lexicons were also manually generated since it was necessary to relate the lexicons to the specific question they

represented. Therefore, the work presented here focuses on whether or not the assessment methods covered in the following sections can overcome the errors introduced into the overall assessment process during the handwriting recognition stage.

As explained in chapter 3, the recogniser used in this work is a holistic word recogniser (Evans *et al*). It is obvious from the results given so far that the performance of this handwriting recognition system is still far from perfect. For an automatic assessment system this is problematic because 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 assessed by the system. However, constraining the scope of what can be expected within the written responses, in order to improve the recognition rates has been shown, in chapters 3 and 4, to improve the assessment accuracy of the system.

In chapter 3, work has already shown that highly accurate assessment of adult multi-word handwritten responses is possible if the constrained nature of the responses is taken into account. 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. The results presented in chapter 4, on the other hand, showed that when recognising single words, a generalised lexicon could be highly ambiguous resulting in poor recognition accuracy when using CLA. This in turn impacted on the assessment accuracy to the point where it is unfeasible to use CLA as part of an overall recognition and assessment.

To overcome this problem, SWAT was introduced in chapter 4 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 reduces the latent ambiguity that is inherent in a generalised lexicon, leading to increased recognition rates, which, in turn, produced acceptable levels of assessment yield and accuracy.

In this chapter, the automatic assessment of a five-question sentence style response exercise is to be attempted (see figure 5.2 on the following page). The exercise formed part of the 'Progress in English 10' exam paper published by NFER-Nelson. All the questions require a single 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*). Additional samples of the children's handwritten single word responses can be found in appendix D.

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



Figure 5.2: 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 by the assessment body along with the questions. In the case of questions 1 & 2 the model answers are explicitly given in the stimulus thus 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 evaluate the written response against the model answer and determine whether they are semantically equivalent, and therefore either correct or incorrect, rather than just relying on finding an exact model answer. The human assessor in this case marked 62% of the responses as correct and 38% as incorrect answers. This was achieved with 100% assessment accuracy with 100% assessment yield using the assessment criteria provided (see appendix C).

In this chapter three investigations are carried out. *Investigation I – CLA vs. SWAT to automatically assess handwritten sentence responses*, will be detailed in section 5.2. This will serve as a preliminary investigation to evaluate how accurately CLA & SWAT can automatically assess the responses. It is envisaged that since CLA has performed poorly when used in the previous chapters, employing CLA on a more complex task will result in an even worse performance. SWAT, on the other hand, has been shown to perform well in both of the previous tasks and therefore, so long as the lexicon is suitably constrained, SWAT may still achieve a high level of assessment yield and accuracy. *Investigation II – Addition of contextual word bridges*, will focus on incorporating word bridges into both CLA and SWAT in order to help minimise the errors introduced at the recognition stage. *Investigation III – Addition of question response history*, will examine the use of a question's response history with the aim of increasing the number of responses assessed by being able to distinguish more correct and incorrect responses as a result of using an improved model answer lexicon.

5.2 Investigation 1 – CLA vs. SWAT to automatically assess handwritten sentence responses

5.2.1 Conventional Lexical Approach

In the first experiment of investigation I, a conventional lexicon was generated from the stimulus provided, Fry et al's 300 most frequent words [42] and all of the words that have been written in the data set. The stimulus for the exercise consists of a short story and the questions themselves. The short story is provided as part of the exam paper and the child must read it prior to answering the questions in exercise 5. The story gives context to the questions in exercise five. The answers to the questions are explicitly given in the story or they are implied therefore the question tests a child's ability to understand the question being asked and extract the relevant information. All of the handwritten words contained in the responses were used to generate the generalised lexicon as this is not an exercise to test the generalisation potential of the recognition 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 by children up to the age of 10 years old. In this exercise 54% of the words written or read were included in Fry's 300-word list. Table 5.1 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 or stimulus but not in Fry's 300	38%
In the written responses or stimulus and in Fry's 300	54%
Not in the written responses or stimulus but in Fry's 300	8%

Table 5.1: 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 template set used in this experiment is 1455 words long. This is comprised of Fry's 300 words plus 185 words that were found in the stimulus and the written responses but not found in Fry's most frequently used words. This produced 485 unique words which, when their case alternatives were added, forms the 1455-word lexicon. A low recognition rate was expected as a result of using a large lexicon made up largely of small words (words that contain less than 5 characters). In holistic recognition, small words generally have low recognition rates as the number of unique features within these words is low thus causing high ambiguity between the small words in the lexicon.

In CLA the word in the lexicon that best matches the word image is used to build a recognised response, which is then passed on for automatic assessment.

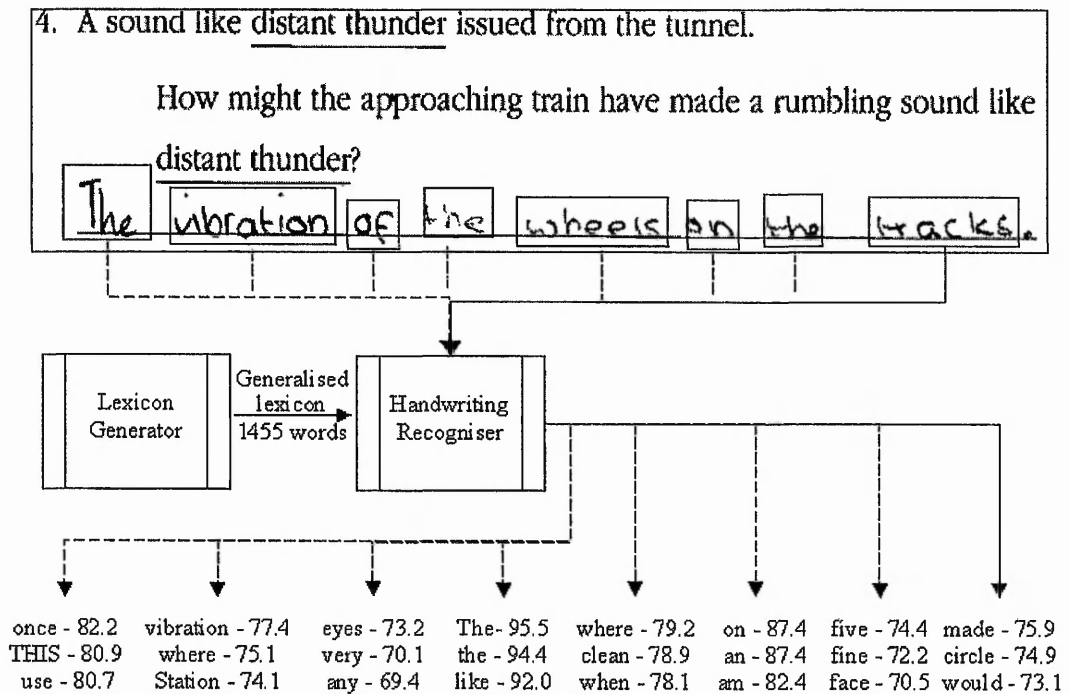


Figure 5.3: An example of the recognition results when using the Conventional Lexical Approach

Figure 5.3 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. Each word is compared to all 1455 word templates in the generalised lexicon and a list of best-matched words is produced for each word (the top three best matches are shown). The best match in each case is then used to build a recognised response that is used in the assessment stage. For example, in figure 5.3 the best match response would be “once vibration eyes The where on five made”.

To assess the response, each word in the recognised response is checked against the model answer. The model answers are given by NFER-Nelson and are shown in appendix C. The correct answer for question 1 is simply *dragon*, whereas question 2 has two possible single word answers, *steam* or *smoke*. For questions 3,4 & 5 the model answers are multi-word answers. 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 no model answer information is found in the recognised response then the response can be marked as incorrect. 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. This is because the price of assessing a misrecognised response is greater than the cost of manually marking these responses (i.e. marking a correct response as incorrect).

5.2.2 Results – Conventional Lexicon Approach

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 (see section 5.4). The test set contained 65 written responses (592 word images). Using a 1455 word lexicon, with all the written words held within it, CLA achieved an overall word recognition rate of only 33%.

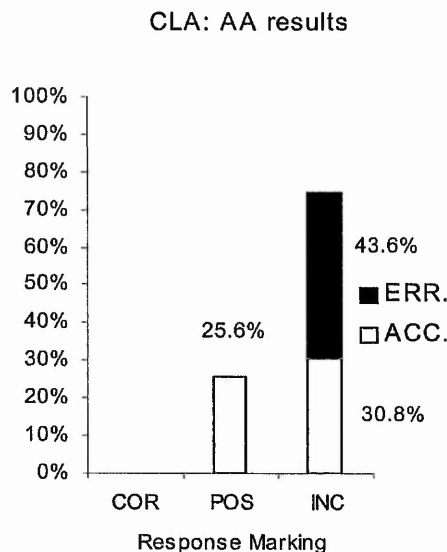


Figure 5.4: A graph to show the assessment accuracy of the conventional lexicon approach

Figure 5.4 shows the automatic assessment results using CLA. It can be seen that CLA has tended towards marking all of the responses as incorrect answers. This is a direct result of the low recognition rate. Few of the best-matched recognised responses for those handwritten responses that were correct contained the required model keyword answer resulting in an incorrect assessment classification. CLA

apparently scored 74.4% of the responses automatically. However, 30.8% of the 74% were accurately assessed but 43.6% were correctly answered responses that were automatically assessed as incorrect answers. Since all of these responses were automatically classified as incorrect the system achieved an overall assessment accuracy of just 41.4%. The achievement is therefore a false one as the results prove that no confidence can be placed in the assessment results when CLA is used to automatically assess handwritten responses.

5.2.3 Specific Word Assessment Technique

SWAT exploits the nature of the question and answer medium by only comparing the input pattern to the template of the correct answer 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.

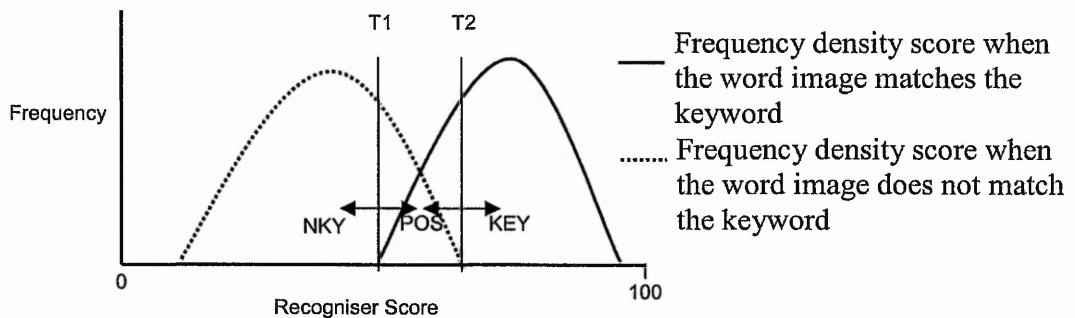


Figure 5.5: An idealised frequency density graph to obtain the two confidence thresholds for SWAT

This localised approach must then use two confidence thresholds per word in each model answer to classify the top ranked recognised words as either a keyword (KEY), possible keyword (POS) or not a keyword (NKY) (see figure 5.5).

Two data sets are shown on the graph. The solid-line represents the frequency density scores for the times when the recogniser is passed a word image that is a keyword. The dashed-line shows the recognition scores for the times when a word image is not a keyword.

This is a similar process to that used in chapter 4 where the confidence thresholds were used to classify the single handwritten word as CORRECT, INCORRECT or POSSIBLE. There is a difference in this method to that used in chapter 4 because the classifications alone cannot determine whether a response has been correctly answered in this case. All the word classifications must be combined to form a sentence style classification response that can be assessed.

The thresholds for each question were determined by using a training set that contained data from 13 different children who completed all five questions. Each word template, built from the model answer keywords, was compared against all of the word images in the training set. From this a frequency density graph was produced for each model answer keyword based upon the recogniser score.

A high recognition score (higher than t_2) implies that the recogniser has achieved a close match between the word image and its model keyword target template. The system can thus confidently classify the word as a keyword. However, if the word has

a low score (lower than t_1), this means that the word image shares few features with the model keyword answer. The recognised word is therefore either a wrong answer or is illegible. Such a word can therefore be automatically classified as not a keyword. This can be achieved with a high confidence because the response image is only compared with that of the correct model answer images.

Figure 5.6 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 keyword answers as the lexicon. The model keyword answer for question four is 'wheels track'. The model answers for all questions can be found in appendix C.

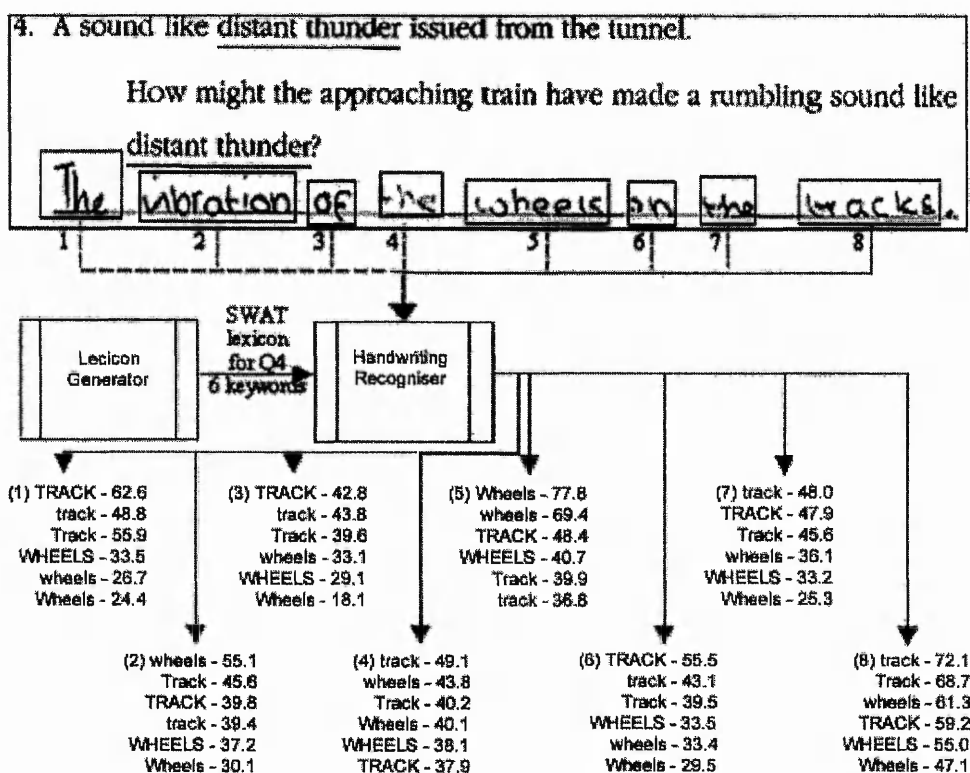


Figure 5.6: An example of the recognition results when using the Specific Word Assessment Technique

Using the two thresholds for 'wheels' ($t_1 = 60.9$ & $t_2 = 74.9$) and 'track' ($t_1 = 61.8$ & $t_2 = 87.5$) the recognised words can be evaluated and classified. Using these thresholds, the classified results for the response shown in figure 5.5 would be:

Keyword classification:- 'POS NKY NKY NKY **KEY** NKY NKY POS'

	Word position 1 - <i>The</i>			Word position 3 - <i>of</i>			Word position 5 - <i>wheels</i>		
Classification	POS			NKY			KEY		
Best-matched list	TRACK	62.6	POS	TRACK	42.8	NKY	Wheels	77.8	KEY
	track	48.8	NKY	track	43.8	NKY	wheels	69.4	POS
	Track	45.9	NKY	Track	39.6	NKY	TRACK	48.4	NKY
	WHEELS	33.5	NKY	wheels	33.1	NKY	WHEELS	40.7	NKY
	wheels	26.7	NKY	WHEELS	29.1	NKY	Track	39.9	NKY
	Wheels	24.4	NKY	Wheels	18.1	NKY	track	36.8	NKY

Table 5.2: Classification of the first, third and fifth word from the response in figure 5.6

Table 5.2 shows how this classification was achieved for the word positions 1,3, and 5 from the recognition results seen in figure 5.6. As can be seen from figure 5.7, the highest possible classification is always chosen from each list of classifications. This is because these results are the most confident results possible using the thresholds. When it comes to assessing the classified response, a complete correct answer must contain all the keywords from the model keyword answer and the keywords must be in the same order that they appear in the model keyword answer. If the keywords found match completely a correct answer then the response will be automatically scored as correct.

For example, if the recognition score for the best matched word in position eight of the response, shown in figure 5.6, had been the 'tracks' keyword threshold (t_2) then the recognised response would have been: POS NKY NKY NKY **KEY** NKY NKY **KEY**. This would have been automatically scored as correct because it has found the model keyword answer.

If all the words are classified as NKY then the response can automatically be scored as an incorrect answer because there no link has been made to the model keyword answer therefore implying that the response is incorrect. A response that is passed for manual assessment is one that contains only a partial model keyword answer (not all keywords have been classified as KEY) and/or contains possible keywords (POS).

For example: NKY NKY NKY NKY **KEY** NKY NKY **POS** – partial answer

or: NKY NKY NKY NKY **POS** NKY NKY **POS** – possible answer.

5.2.4 Result – Specific Word Assessment Technique

As stated earlier, the testing of SWAT firstly involved creating training and test sets. 50% of the data was randomly removed and used for the training set whilst the remaining 50% was set aside as the unseen test set. The frequency distribution graphs of the correct and incorrect responses in the training set were then plotted (figure 5.7 shows the graph for Q1 as an example).

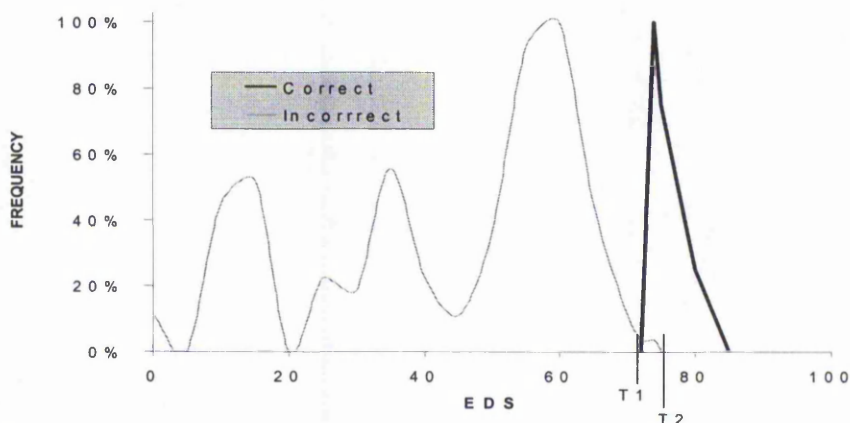


Figure 5.7: A graph to show the frequency density graph of the responses to Q1, Exercise 5 of the Progress in English 10 exam paper

The thresholds T_1 and T_2 are then defined as:

T_1 – the recognition score below which no correct response is scored

T_2 – the recognition score above which no incorrect response is scored

From the graphs, the values for t_1 and t_2 for question one were determined as 72 & 74 for the model keyword answer, 'dragon'. Similar values were gained for all five questions and these are shown in table 5.3.

Question	Model Keyword	Low Threshold (t_1)	High Threshold (t_2)
1	Dragon	72.2	74.6
2	Steam	69.1	73.9
	Smoke	71.2	81.2
3	Long	62.3	81.9
	Black	62.2	73.9
4	Wheel	60.9	77.9
	Track	61.8	87.5
5	Stayed	71.6	80.1
	Behind	64.0	75.5
	Train	71.0	82.9
	Gone	67.0	78.9

Table 5.3: SWAT thresholds for the model keywords for all five questions

SWAT: AA results

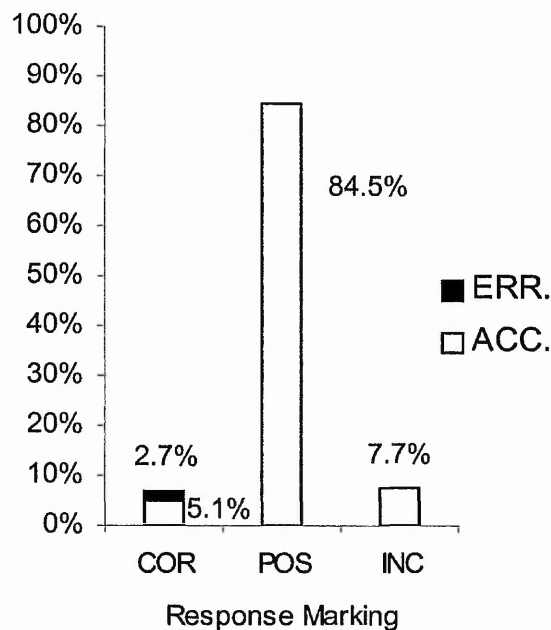


Figure 5.8: A graph to show the assessment yield and assessment accuracy of the Specific Word Assessment Technique

These thresholds were then used to assess the responses in the test set. The test set comprised of responses from 13 writers and the results of the automatic assessment can be seen in figure 5.8.

By comparing figure 5.8 and 5.4 (page 114), it can be seen that CLA has automatically assessed more responses than SWAT. However the overall assessment accuracy of CLA is very low (41.4%) and is easily out performed by SWAT (82.8%). This has confirmed the results from the previous chapter where SWAT was shown to be more suitable for automatically assessing handwritten responses than CLA. However, having said that, the amount of responses that have been automatically assessed by SWAT is far too low to provide a viable solution. Additional methods are therefore needed to support SWAT in order to increase the assessment yield whilst, at the same time, maintaining the assessment accuracy. In the next two sections it will be shown that contextual word bridges and a question's response history can be used with SWAT to produce an assessment system that is potentially a commercially viable solution

5.3 Investigation 2 – Addition Of Contextual Word Bridges

A sentence is a multi-word response, therefore contextual knowledge in the form of contextual word bridges can be used to try and improve the automatic assessment system. In chapter 3 it is shown that by employing contextual bridges between known keywords the automatic assessment accuracy achieved using CLA alone can be improved. In this section, it will be shown that with the use of contextual bridges

between word pairs the automatic assessment system employed in the previous section can be improved. Therefore, in this section a comparison is made between CLA that employs contextual word bridges (CLAB) and SWAT which makes use of contextual keyword bridges (SWATB).

5.3.1 CLA Using Contextual Word Bridges

The conventional approach to handwriting recognition is to use a lexicon that is made up from all the possible words that could be found in any response, as seen in the previous section. This usually requires a large lexicon, which is highly ambiguous. As a result, recognition rates are generally low. However, in this work the recognised responses have also to be assessed rather than just recognised and by exploiting the assessment medium further with the use of word-pair bridges it may be possible to improve the recognition by promoting the best matched contextual bridges instead of relying on the top ranked word.

The recogniser designed by Evans *et al* has been shown to rank the correct word in the top 5 places 82% and this increase to 96% for the top 50 rank word alternatives [39]. Therefore in this experiment, the 1455-word lexicon detailed in section 5.2.1 will be used again to recognise the sentence responses. However, instead of passing just the best-matched recognised response for assessment (i.e. responses made up of all the best-matched words), a search of the top 50 best-matched words in each word position will be carried out to find the highest contextual word bridges and the best-matched sentence that can be made up using the highest contextual word bridges. Therefore, a search of the top 50 ranked words for contextual word bridges based on

the stimulus should promote the correct words as they form the bridges and thus a more acceptable recognised response.

A contextual word bridge is simply a link between pair of words. Figure 5.9 shows the response given in figure 5.3 for the 4th question and shows that there are seven contextual word bridges present in the eight-word sentence.

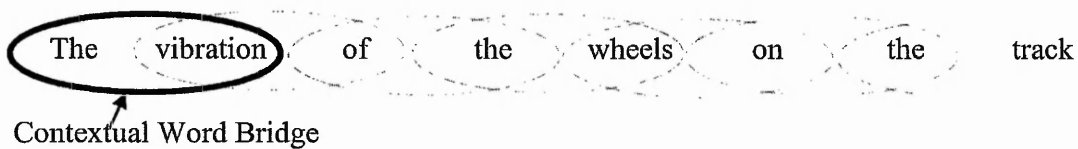


Figure 5.9: Example of contextual word bridges in a sentence

Firstly a list of the contextual bridges has to be produced. In chapter 3 the bridges were formed using the actual possible answers as they were the only expected responses. In this case however the list must be made up of general pairs of words.

In investigation 1 the template word list used was based upon the stimulus provided and the rest of the words that were missing from the actual written responses. Therefore the word pairs have also been taken from the stimulus. A manual evaluation of the text was carried out and all of the unique word pair was recorded. Also, as some of the words from the template set were taken from written responses, the written responses were also examined and any missing word pair was added. The missing word pairs made up 24% of the total contextual bridges used.

Using the recognition lists from the first investigation the top 50 ranking words for the first word in the sentence were explored for contextual bridges with the neighbouring

word's list of recognised words. This continues for all words and all ranked alternatives for each word. When a bridge is formed the raw recognition score of both words is summed. The formation of contextual word bridges that link more than one word pair can form a recognised sentence. The ranked output results of passing the recognition results from figure 5.3 can be seen in figure 5.10.

the (74.2) vibration (77.4) of (40.1) the (95.5) wheels (77.8) on (87.4) the (65.3) circle (74.9) 590
the (74.2) vibration (77.4) of (40.1) the (95.5) track (63.4) on (87.4) the (65.3) circle (74.9) 575
the (74.2) vibration (77.4) of (40.1) the (95.5) trees (45.2) on (87.4) the (65.3) circle (74.9) 557

Figure 5.10: The top three bridged sentences formed for the recognition results shown in figure 5.3

It is clear that the resulting contextual word bridge constrained responses from CLAB are more syntactically correct than the best matched response from CLA. However, the highest ranked response sentence, '*the vibration of the wheels on the circle*' is still not completely correct because only a partial answer was found. This recognised response would, therefore, have been passed for manual assessment using the same assessment method used in investigation 1.

5.3.2 Results – CLA With Contextual Word Bridges

By comparing figure 5.11 with figure 5.4 (page 114) it can be seen that the overall results of the assessment accuracy for CLAB are worse than for CLA.

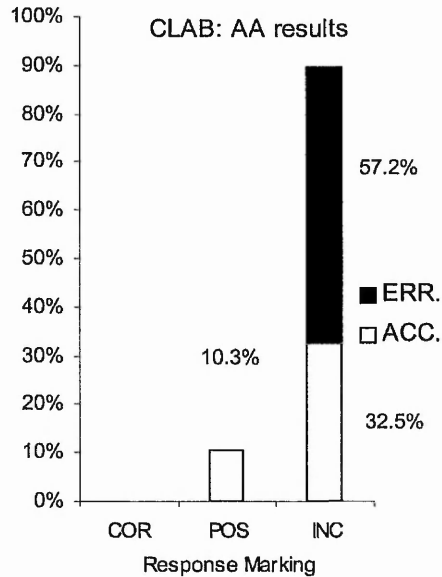


Figure 5.11: A graph to show the assessment accuracy of the conventional lexicon approach with contextual bridges

This is a result of a very poor recognition accuracy (33%). The lexicon contained 1455 words and only the top 50 words were searched for contextual word bridges. As a consequence of the poor recognition rate the actual written words only appear 73% of the time in the 50-word alternative list. Therefore the best-matched sentences that resulted were forced into the incorrect class because the contextual word bridge constrained sentences that contained the keywords were lower down in the list of the best recognised sentences and were not promoted enough.

5.3.3 SWAT With Contextual Keyword Bridges

As it stands, the contextual word bridges employed in CLAB cannot be used to search the recognition list from SWAT because only the keywords are evaluated against all the written words. This means that no other words are available in the recognition list to form bridges between the keywords.

To overcome this problem wildcards are introduced so that the model keyword answer can be found in a contextual keyword bridged response. In CLAB it was shown that the bridges could be used to promote poorly recognised words so that a more comprehensible sentence could be formed rather than just relying on the best-matched sentence. The use of wild cards in SWATB allows the searching of all word positions for the best complete model keyword answer that has the correct contextual keyword bridges structure. This contextual keyword bridged response is then passed for automatic assessment. However, it should be obvious that no extra improvement can be gained for questions one and two, as they only require a single word answer.

X	X
X	WHEELS
X	TRACK
WHEELS	X
WHEELS	TRACK
TRACK	X

Figure 5.12: The SWATB bridge set used for question four

The template word set used for SWATB is identical to that used in SWAT, however a set of contextual keyword bridges has been added. Figure 5.12 shows the set derived from the assessment criteria for question four. The word pair TRACK+WHEELS is

not included in the set because the order of the keywords is also taken into account when generating the set.

Using the contextual keyword bridges, the best-matched model keyword answer can be generated from the recognition results from SWAT. Like CLAB, a ranked list of the best-matched model keyword answers is produced on the basis of the combined recognition score for each word. Wildcards have been given a token value recognition score of zero as they are only used to

aid in the reconstruction of the response and do not contribute to the overall assessment of the response.

Of course, the use of wildcards means that all possible bridges are found for each response, as each position will contain only the recognition results for the keywords *wheels* and *track*. Therefore, if the response only contained two words the length of the list of best-matched keyword answers formed from the list of recognised keyword alternatives would be equal to that of the six contextual keyword bridges in the corpus containing only one

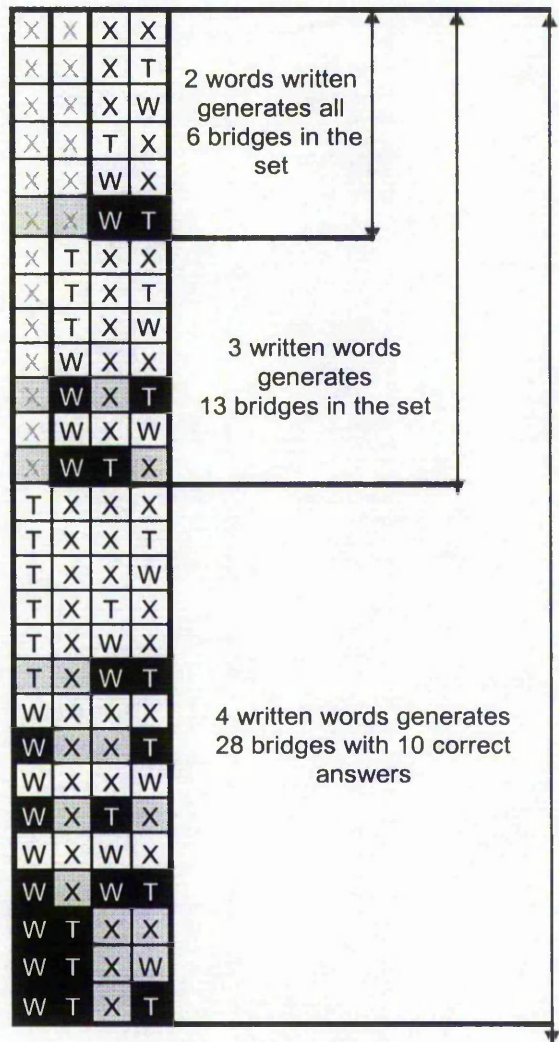


Figure 5.13: All possible bridges sentences that can be generated from 2,3 & 4 word written responses

possible model keyword response (highlighted in figure 5.13). Of course there will also be the response 'X X' which will have a score of zero. This is a necessary contextual keyword bridge because when there are more words in the written response there has to be a way of traversing through the recognised list as there are only keywords and no linking words present.

Figure 5.13 shows all possible bridged sentences that can be generated from 2,3 & 4-word written responses. The correct model keyword answers are highlighted. The remaining sentences are only partial model key word answers or contain both keywords separated by a wildcard but in the wrong orientation.

As an example, using the corpus on the four word handwritten response "Wheels on the track", should result in the best bridged sentence being "*W X X T*" where the first and last words have been matched highly against the 'Wheels' & 'track' word templates found in the lexicon. However it is possible that the best bridged sentence could be either "*W T X T*" or "*W X W T*". This is because the words 'on' and 'the' will be matched against the 'wheels' and 'track' word templates and they will gain a recognition score. However, the score for these words should be significantly low and should be therefore easily identified as non-keywords using the SWATB thresholds.

If the child had written the four-word response "wheels on the track" the score for all 28 possible bridged responses will be calculated and the highest scoring response will be passed for assessment.

In order to assess these new responses the original SWAT thresholds, used to determine the keyword classification, were recalculated from the best contextual keyword bridged responses formed from the training set (see table 5.3). By comparing table 5.4 with table 5.3 (page 120) it can be seen that the thresholds for Q3 – 5 have been altered.

Question	Model Keyword answer	t_1	t_2
1	Dragon	72.2	74.6
2	Steam	69.1	73.9
	Smoke	71.2	81.2
3	Long	65.3	77.5
	Black	63.1	72.1
4	Wheels	63.7	72.5
	Track	65.8	71.3
5	Stayed	71.6	78.2
	Behind	66.2	75.5
	Train	72.3	74.6
	Gone	67.0	78.9

Table 5.4: The high (t_2) and low (t_1) thresholds calculated for SWATB

This is a result of the way that the bridges have been used to build the best contextual keyword bridged sentence response. This reduces the occurrences of keywords being mistaken for non-keywords and vice versa. On the face of it, narrowing the gap between the thresholds should allow the possibility of automatically assessing more responses because the ambiguity between correct and incorrect keyword identification has been reduced. For example the eight word handwritten response shown in figure 5.6 (page 117) “*the vibration of the wheels on the track*”, produces a best bridged sentence of “*W T X X W T X T*”.

Using the original SWAT thresholds this response could not be automatically assessed because the actual written word ‘track’ could only be identified as a possible keyword. In SWAT this response was passed for manual assessment. Under SWATB

the system is able to automatically assess this response as correct because the threshold t_2 is lower which enables the last written word to be identified as a keyword.

5.3.4 Results – SWAT With Contextual Keyword Bridges

Figure 5.14 shows the assessment results when SWATB is applied to the test set data. Compared to the results of SWAT shown in figure 5.8 (page 120) it can be seen that the response yield has increased along with the assessment accuracy, 20.9% and 87.1% respectively. This is a result of converting more possible class responses into correct class responses due to the alteration in the thresholds allowed when using contextual keyword bridges to constrain the recognised response. That said, the assessment yield and assessment accuracy are still too low to provide a commercially viable solution.

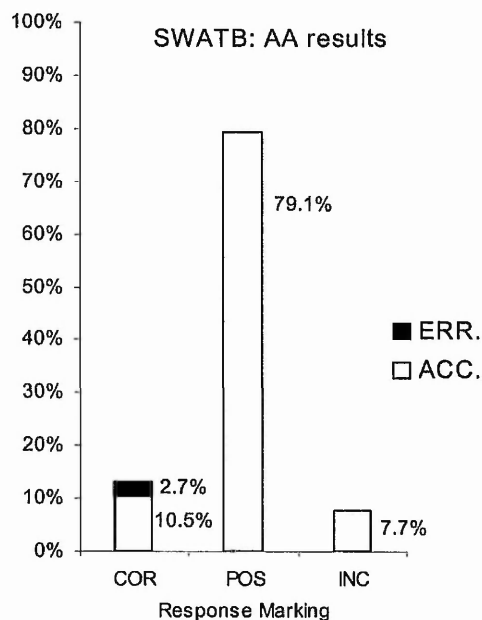


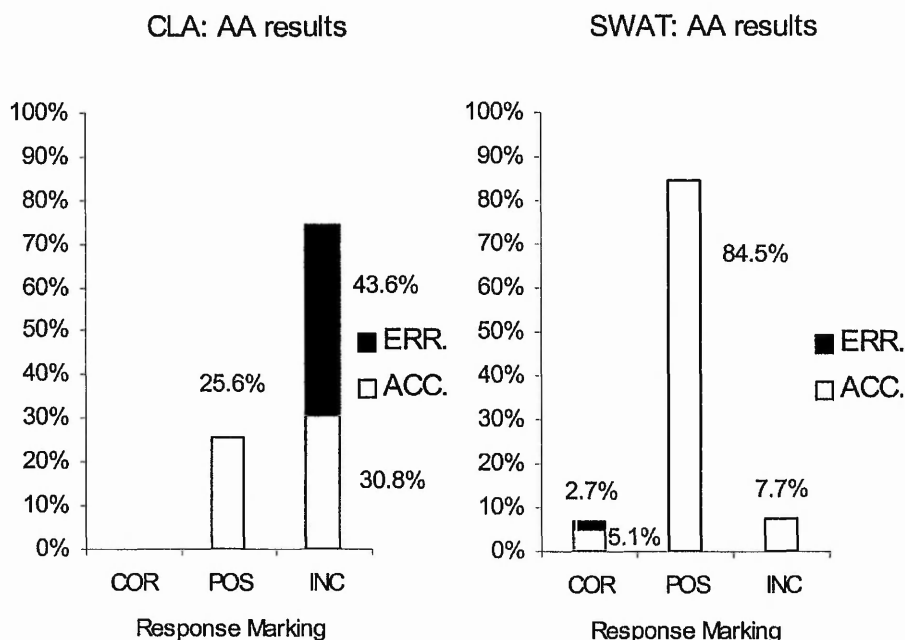
Figure 5.14: A graph to show the assessment accuracy of the SWAT with contextual keyword bridges

5.4 Investigation III – Addition of Question Response History

The use of a questions history has also been investigated to show that author supplied model answers alone provide insufficient assessment criteria for an automatic assessment system. Additional assessment knowledge, obtained from previously marked responses, is necessary in order to mark all the handwritten response more accurately. The training set used to calculate the thresholds used in SWAT & SWATB was also used to provide the question response history. The training set was manually analysed and all the correct answers found. These were used to form new assessment criteria. The new assessment criteria thus contain the model keyword answers for the correct answers found in the training set as well as the author supplied model answers. The criteria also included frequently incorrect answers that could be used to better differentiate an incorrect answer from a correct one. The model answers used in this investigation can be found in appendix C.

5.4.1 Without The Use Of History

To recap on the CLA and SWAT assessment results without the use of a question's response history, the results given in sections 5.2 and 5.3 are repeated here (see figures 5.15 and 5.16 on the following page).



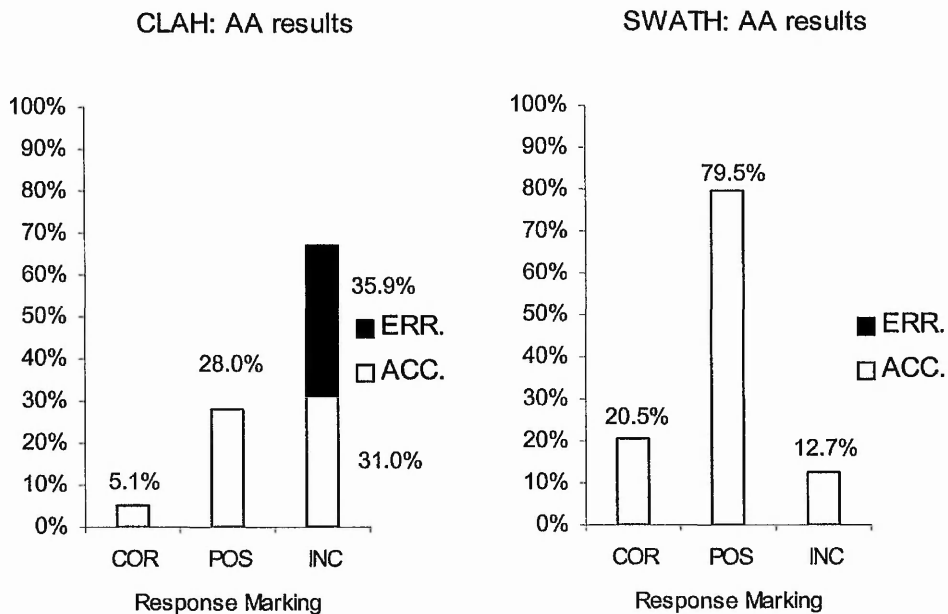
Figures 5.15 & 5.16: Graphs to show the results of automatically assessing the handwritten responses using the CLA & SWAT approaches respectively

Figures 5.15 & 5.16 show the results of the CLA & SWAT assessment approaches respectively. It can be seen that 25.6% of the responses recognised using CLA were all 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 misrecognised and erroneously assessed. This error rate in the responses automatically assessed using CLA was a result of the extremely low word recognition rate obtained when using the large 1455 word lexicon. In contrast, SWAT automatically assessed 15.5% of all response rejecting 84.5% for manual assessment, as it was not confident in marking them. 2.7% of the responses were incorrectly assessed 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 high assessment

accuracy but the number of responses was low as a result of the assessment criteria being too limited. The human assessment of these scripts achieved 100% assessment yield with 100% assessment accuracy using the same assessment models. This indicates that the human assessors must have used 'common sense' or additional assessment knowledge to score each handwritten response against the model answers provided. The use of a question's response history should provide an automated assessment system with some of this additional knowledge thereby allowing it to make assessment decisions that are closer to those of the original human assessor.

5.4.2 With Question Response History

Figures 5.17 and 5.18 show the results of applying CLA and SWAT with history to the assessment of the test set data.



Figures 5.17 & 5.18: Graphs to show the results of automatically assessing the handwritten responses using the CLAH & SWATH approaches

By comparing figures 5.17 and 5.18 (CLAH & SWATH) to figures 5.15 and 5.16 (CLA & SWAT) it can be seen that with the inclusion of new assessment criteria (more model keyword answers) the accuracy of both methods has increased. However, the number of responses (assessment yield) automatically assessed by CLAH has decreased whilst under SWATH it has increased. This is a result of CLAH recognising more of the actual responses as partial model answers rather than incorrect ones because there are more possible keywords recognised. SWATH was also able to be more confident in its assessment of responses because of the wider assessment criteria. When the history data is added to the assessment criteria, the accuracy of the SWATH system also increases to 100%. This is due to the approach being able to assess responses such as 'steam engine puffs and snorts' in question two as incorrect. Without history this response is marked as correct because 'steam' is identified and classified as a keyword. The addition of the history made up for the lack of 'common sense' therefore making it possible to disallow 'steam' when directly connected to 'engine'. Consequently SWATH is able to automatically assess more responses more accurately.

5.5 Conclusion

A summary of the number of responses that are automatically assessed and the assessment accuracy of each approach is given in table 5.5.

	Responses Automatically Assessed (%)	Assessment Accuracy (%)
CLA	74.4	41.4
CLAB	87.7	37.1
CLAH	72.0	46.3
SWAT	15.5	82.8
SWATB	20.9	87.1
SWATH	33.2	100

Table 5.5: A summary of the Assessment Accuracy and % of responses assessed for all approaches

In this chapter, two methods for assessing children's handwritten sentence responses have been compared. The conventional lexical approach, using a 1455 word lexicon, apparently provided assessment yields as high as 74.4%. However, this approach incurred a large number of errors resulting in a response accuracy of just 41.4%. When contextual word bridges were employed the accuracy actually decreased as most of the responses were classified as incorrect answers. This produced a false improvement in the amount of responses that could be automatically assessed. However the accuracy did increase slightly to 46.3% when the history was introduced but at the expense of the response yield. The results for CLA and its subsequent augmentations suffered as a direct result of the poor overall word recognition rate (33%) when using a generalised lexicon to recognise the children's handwriting.

SWAT, on the other hand, has a very high keyword classification rate (97.1%) and had a higher response assessment accuracy (82.8%). The keyword classification rate was calculated as the number of times SWAT correctly identified a keyword against the times when the keyword was present but was not classified as a keyword. Both the assessment accuracy and the assessment yield were increased by the use of contextual keyword bridges and further increased with the addition of the question's response history. SWATH's keyword classification rate decreased slightly to 94.3%. This can be attributed to the increase in ambiguity of the larger lexicon. However, since the assessment criteria contained extra knowledge more keywords were actually found thus producing a higher assessment yield. SWATH assessed 17.7% more responses than SWAT, automatically marking 33% of the responses with an assessment accuracy of 100%. The main reason why SWATH was so accurate was that with the inclusion of more assessment knowledge, in the form of the question response history,

the approach was able to not only assess more responses but also overcome more of the errors in the recognition stage.

The addition of the contextual keyword bridges formed from the question response history and a large-scale trial of SWATH is required to determine if these results can be sustained such that they provide a viable solution to ease the burden of marking handwritten sentence responses.

CHAPTER 6: CONCLUSION AND FUTURE WORK

In this thesis, the automatic assessment of handwritten responses has been introduced as a new research area. This work shows that some of the errors that are introduced through recognition of the handwriting can be overcome using the contextual constraints that are inherent in assessment. These errors are further minimised by using the Specific Word Assessment Technique (SWAT), which utilises the nature of the assessment medium to concentrate on scoring responses according to how well they match the correct answer at the recognition level. This is in direct contrast to the Conventional Lexical Approach (CLA) where the assessment is reliant on the handwriting recognition stage producing strong candidates for the written responses before being assessed.

6.1 What Has Been Achieved From This Research?

The results of the initial investigation in chapter 3 show that automated assessment using CLA is feasible so long as the context within the response can be exploited. Knowledge of the expected response is shown to help identify responses that contain handwriting recognition errors thus enabling accurate automated assessment. This is indicated by the overall performance of the assessment system where 54% of all responses were correctly classified with an assessment accuracy of 99% using question specific 9-word lexicons.

That said, the more generalised 210-word lexicon results shown in table 6.1 (this is a summary of table 3.5 on page 68) indicate that this conventional lexical approach

would not be able to scale up to the sort of large lexicon recognition problems required when trying to assess unconstrained handwritten sentence responses. The CLAB and CLAH results presented in chapters 5 support this conclusion.

	Lexicon Size		
	9	27	210
Overall	83.3%	71.5%	41.0%

Table 6.1: Raw recognition results from Chapter 3

Following on from this preliminary investigation, a novel approach for the automated assessment of children's handwritten single word responses was introduced. This technique was born out of necessity because the lack of context in the single word responses meant that the contextual bridging algorithm used in chapter 3 could not be used to improve the assessment accuracy. The Specific Word Assessment Technique (SWAT) performed very well when compared to a conventional lexical based approach. Yields of 44% were gained with 100% assessment accuracy using static thresholds on children's handwritten responses. This compares favourably to the conventional lexical approach where the low recognition rate of 59% severely compromised the assessment accuracy (41% error).

The explanation as to why SWAT was more accurate than CLA became apparent when the frequency density graph was examined (see figure 4.17 on page 92). In SWAT the response images were only checked against the correct answer. This reduced the ambiguity within the lexicon and increased the separation between the incorrect and correct response recognition scores thereby reducing assessment error. Since the scope of recognition was narrowed using SWAT, a lower threshold could also be used to identify actual incorrect responses thus increasing the response yields further.

Retrospective application of SWAT to the adult multiple-choice question responses used in chapter 3, also showed that the SWAT verification based recognition and assessment method could outperform the contextual knowledge enhanced CLA.

In the final chapter, two methods for assessing children's handwritten sentence responses were compared. The conventional lexical approach, using a 1455 word lexicon, apparently provided assessment yields as high as 74.4%. However, this approach incurred a large number of errors resulting in an assessment accuracy of just 41.4%. When contextual word bridges were employed to augment the conventional lexical approach, the assessment accuracy actually decreased (37.1%) as most of the responses were classified as incorrect answers. This produced a false improvement in the amount of responses (87.7%) that could be automatically assessed. However the accuracy did increase slightly to 46.3% when the question response history was introduced but this was at the expense of the response yield (72.0%). The results for the conventional lexical approach and its subsequent augmentations suffered as a direct result of the poor overall word recognition rate (33%).

SWAT, on the other hand, has a very high keyword classification rate (97.1%) and had a higher assessment accuracy (82.8%). Both the assessment accuracy and the assessment yield were increased by use of contextual keyword bridges and further increased with the addition of the question's response history. SWATH's keyword classification rate decreased slightly to 94.3%. This was attributed to the increase in ambiguity of the larger template set. However, since the assessment criteria contained extra knowledge more keywords were correctly classified than with SWAT thus

producing a higher assessment yield. SWATH assessed 17.7% more responses than SWAT, automatically marking 33% of the responses with an assessment accuracy of 100%. SWATH was more accurate than SWAT because the extra assessment knowledge enabled better reconstruction of the keyword response and thus minimised the possibility of incorrect responses being automatically assessed as correct.

The addition of contextual keyword bridges formed from the question response history and a large-scale trial of SWATH is required to determine if these results can be sustained to provide a commercially viable solution to ease the burden of marking handwritten sentence responses. In addition, the techniques developed in this report have only been shown to be applicable in highly constrained areas of automated assessment. Future work needs to be done to address the issues that will arise when the automatic assessment of more complex question styles is carried out. Increasing the generalisation of the SWAT method may cause the keyword classification rate to decrease with its resulting impact on the assessment performance as a whole.

In summary, the work in this thesis has introduced the novel research area of automatic assessment of handwritten responses. It has demonstrated that current handwriting recognition technology can be used for automated assessment of handwritten responses and has shown that the constrained nature of assessment can be exploited to reduce recognition errors.

6.2 Future Work: How far can Automated Assessment go?

In this thesis, the proposed automatic assessment system produced encouraging results proving that, at the very least, this area of research is worth further investigation. In this section the opportunity for further research into automatic assessment of handwritten responses is discussed. Firstly a discussion of possible improvements to the automatic assessment approach designed in this thesis is given. Then a theoretical method is proposed for calculating how confident an automatic assessment system can be based on its accuracy at recognising model keyword answers. This concludes with a brief account of the how principles behind the Specific Word Assessment Technique could be employed in other applied areas of handwriting recognition.

6.2.1 Proposed Improvements to the Automatic Assessment System

The most obvious improvement to the automatic assessment system used in this thesis would stem from improving the accuracy of the handwriting recogniser itself or by using a multiple recogniser system. Pre-classification of responses, based on the style of the handwriting, could be used to determine which recogniser is most suitable to recognising the given style of handwriting. In this work, the recogniser used is more suited to recognising long words that contain more features than smaller words that contain a low number of features. A word length based pre-classification could be used to only allow long words to be passed to this recogniser thereby minimising the recognition errors. A suitable small word recogniser would then need to be developed.

Improving the recogniser itself could be achieved by exploiting the assessment medium further. By using a predetermined response for the children to copy before the exam starts or by using the information already present (in the form of their name and school) the recogniser could be dynamically trained to a given child's unique style of handwriting. Improving the recognition stage would result in less recognition errors being incurred, which would inherently mean that more responses could be automatically assessed more accurately. Of course, identifying a more advanced recogniser may also improve recognition. Therefore, the principles of SWAT can again be exploited to improve the overall robustness of the system as the work in this thesis is not recogniser dependent.

Another way to improve the overall assessment accuracy would be the introduction of more complex artificial intelligent agents, such as Artificial Neural Networks (ANN), into the automatic assessment process. These could be employed to calculate the thresholds used in SWAT based on more information than just the raw recognition scores of the words. It has been shown that the thresholds are vital and ANN's are good candidates for this type of pattern recognition as more complex information can be evaluated. Extra low-level information in the form of the word lengths, pixel density and centre of gravity of the word image together with high-level information such as the recognition scores of all the keywords, not just the highest scored, could be used to produce better-defined thresholds.

In addition to improving the recognition and assessment accuracy of the proposed system, the automatic assessment of more complex responses could be investigated to

stretch the boundaries of automated assessment. The accurate assessment of essay style responses and sentence responses that require broader assessment criteria than those seen in the investigations in this thesis need to be investigated. In order for this research to meet its ultimate goal the automatic assessment of all handwritten responses must be achieved to the same level as that of a human assessor.

6.2.2 How Robust Is Automatic Assessment?

A paper published as part of this work [4], but not presented in this thesis, shows that the results in chapter 3 can be calculated through the use of a Theoretical Scoring Confidence measure (the full paper can be found in appendix E). Contextual bridging and prior knowledge of the required responses is used to predict the probability of the classification for a given recognition rate. This is defined as the Theoretical Scoring Confidence and was successfully tested on two multiple-choice questions that required different structural responses. Analysis of the results showed that the theoretical rates of assessment accuracy and assessment yield matched the actual results gained from using CLA for the automatic assessment of the adults responses to multiple-choice questions. Currently, the accuracy of the automatic assessment system is calculated by manual analysis of the results gained at the end of an assessment. Use of the Theoretical Scoring Confidence measure would allow the performance of the system to be calculated, prior to the system being implemented.

It is interesting to note that this measure could be adapted to any response format that has more than one word (or recognisable element) in a predetermined format that

allows contextual bridging to be carried out. The recognition of postal address codes, telephone numbers and courtesy amounts on cheques are just a few examples where this measure could also be used to calculate the recognition rate required to meet the industry standard levels of yield and error rate prior to system implementation. The portability of this measure is significant, however, future work could look into whether this approach could also be used to theoretically predict the assessment accuracy and response yield of using SWAT on sentence style responses. This could be based upon the keyword classification rate with the contextual keyword bridges being used as the structural response required by the theoretical scoring confidence algorithm.

6.2.3 Other Applications

The approaches in thesis have been aimed at the automatic assessment of the handwritten responses. However this is not the only area that could possibly benefit from work reported here and two possible areas are Postal Address Interpretation (PAI) and Cheque Amount Verification (CAV). Both applications would probably suffer from not being a more generalised process than that shown in this work as they cannot be constrained as much as the assessment examples used here by virtue of their application area. However, the contextual nature of an address could allow SWAT to be used to verify the postal code of an address that has been confidently recognised or vice versa. That is, the current PAI methods, specifically designed to recognise a postal address (see section 2.4.1), could be as a first pass recogniser in a multistage system where SWAT could be employed as a second pass verification tool with a

lexicon based around the most confidently recognised alternative responses. Using SWAT would therefore improve the confidence of a system as the ambiguity within a generalised lexicon is removed at the SWAT verification stage thus improving both accuracy and yield.

REFERENCES

- [1] Abak, A. Baris, U. Sankur B. The Performance Evaluation of Thresholding Algorithms for Optical Character Recognition. 4th International Conference on Document Analysis and Recognition, 1997, pp 697 - 700
- [2] Al-Abbas. R, A prototype system for off-line signature verification using multilayered feedforward neural networks. Minor thesis, RMIT, Department of Computer Science, Melbourne, March 1994.
- [3] Allan, J. Allen, T, Sherkat, N. Automated Assessment: It's Assessment Jim But not as We Know It. International Conference on Document Analysis and Recognition, Seattle 2001. pp 926 – 930.
- [4] Allan, J. Allen, T, Sherkat, N. Automated Assessment: how Confident are We? 8th International Workshop on Frontiers in Handwriting Recognition. 2002.
- [5] Allan, J. Allen, T, Sherkat, N. Confident Assessment of Children's Handwritten Responses. 8th International Workshop on Frontiers in Handwriting Recognition. 2002.
- [6] Allan, J. Allen, T, Sherkat, N. Automated Assessment of Children's Sentence Responses. Seventh International Computer Aided Assessment Conference, Loughborough, July 2003 pp 3-9
- [7] Allott, N. A Natural Language Processing Framework for Automated Assessment. (1999) PhD thesis, The Nottingham Trent University.
- [8] Baja, R. Chaudhury, S. Signature Verification using multiple neural classifiers. In Proc. Pattern Recognition Vol. 30 No 1, 1997, pp 1 -7
- [9] Baltzakis, H. Papamarkos, N. A new signature verification technique based on a two-stage neural network classifier. Engineering Applications of Artificial Intelligence 14 (2001) 95 –103
- [10] Bellaby, G. Evett, L. Powalka, R. Coping with ambiguity and error in cursive script recognition, 5th International Workshop on Frontiers in Handwriting Recognition. 1996
- [11] Bernardis, G. Bourland, H. Improving Posterior based confidence measures in hybrid HMM/ANN speech recognition systems. In Proc. ICSLP'98
- [12] Bertille, J. Yacoubi, A. Global Postal Code Recognition using Hidden Markov Models. First European Conference on Postal Technology JET POSTE 93, Nantes, France, pp. 129-138, June 273

-
- [13] Blumenstein, M. and Verma, B. A Segmentation Algorithm used in Conjunction with Artificial Neural Networks for the Recognition of Real-World Postal Addresses, *Journal of Intelligent Systems*, 8, 1997, pp 1-13.
- [14] Bouchaffra, D. Krpasundar, E. Srihari, R. Incorporating Divers Information sources in Handwriting Recognition Post-processing. *International Journal of Imaging Systems and Technology*, Vol. 7, 1996, 320-329
- [15] Bouchaffra, D. Krpasundar, E. Srihari, R. Srihari, S. Intergrating Signal and Language Context to Improve Handwritten Phrase Recognition: Alternative Approaches. In Proc. Of the 6th international workshop on artificial intelligence and statistics, 1997, pp 47 - 54
- [16] Box, I. Assessing the Assessment: an Empirical Study of an Information Systems Development Subject. In Proc. Fifth Australasian Computing Education Conference (ACE2003), Adelaide, Australia. *Conferences in Research and Practice in Information Technology*, 20. (2003). Greening, T. and Lister, R., Eds., ACS. 149-158.
- [17] Blumenstein, M. and Verma, B. A Neural Network for Real-world Postal Address recognition. WSC2: 2nd Online World Conference on Soft Computing in Engineering Design and Manufacturing, 1997
- [18] Brakensiek, A. Rottland, J. Wallhoff, F. Rigoll G. Adaptation of an Address Reading System to Local Mail Streams; The proc. 6th international conference on Document Analysis and Recognition, Seattle, 2001 pp 872 – 876
- [19] Bunke, H. Roth, M. Schukat-Talamazzini, E. Off-line Cursive Handwriting Recognition using Hidden Markov Models. In Proc. Pattern Recognition. Vol. 28, No. 9, 1995, pp 1399-1413
- [20] Brammell, N. Connolly, J. Hinde, C. An Investigation into the use of Linguistic context in Automatic cursive script recognition. Seventh International Workshop on Frontiers in Handwriting Recognition September 11-13 2000
- [21] Bunke, H. Roth, M, Schukat-Talamazzini, E. Off-line Cursive Handwriting Recognition using Hidden Markov Models. *Pattern Recognition*, 28 (9), pp. 1399-1413, 1995
- [22] Burstein, J. Leacock, C. Swartz, R. Automated Evaluation of Essays and Short Answers. Fifth International CAA Conference Proceedings. Loughborough (2001).
- [23] Burstein, J. Kaplan, R. Parsing Sentence Fragments in Computer Assisted Test Scoring. 1st International Conference on Teaching and Language Corpora, 1994
- [24] Caesar, T. Gloger, M. Mandler, E. Using Lexical Knowledge for the recognition of poorly written words. 3rd International Conference on Document Analysis and Recognition. 1995, pp 915-918
-

-
- [25] Cai, J. Liu, Z. Off-line unconstrained Hand Written Word Recognition (2000). International Journal of Pattern Recognition and Artificial Intelligence, Vol 14, No3: 259-280.
- [26] Chase, L. Word and Acoustic Confidence Annotation for Large Vocabulary Speech Recognition. Proc. Eurospeech '97, 1997, pp 815-818
- [27] Chaudhuri, A. Chaudhuri, S. Robust detection of skew in document images. In IEEE transactions on image processing, vol. 6, no. 2, 1997
- [28] Chen, C. DeCurtins, J. Word Recognition in Segmentation free Approach to OCR. 2nd International Conference on Document Analysis and Recognition, 1993, pp 573 - 576
- [29] Chin Keong, L. Leedham, G. Empirical Design of a Holistic Verifier for Automatic Sorting of Handwritten Singapore Addresses, 5th International Conference on Document Analysis and Recognition, 1999, pp 733 – 737
- [30] Choi, S. Oh, S. A Segmentation-Free Recognition of Handwritten Touching Numeral Pairs Using Modular Neural Network. IJPRAI 15(6): 949-966 (2001)
- [31] Collier, R. An historical overview of natural language processing systems that learn. Artificial Intelligence Review 8, 1994, pp 17-54
- [32] Cornhill, H. Case-smith, J. Factors that relate to good and poor handwriting. The American journal of occupational therapy, 1996, vol 50, no. 9 pp 732 - 739
- [33] Di Lecce, V. DiMauro, G. Guerriero, A. Impedovo, S. Pirlo, G. Salzo, A. A New Hybrid Approach For Legal Amount Recognition. Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, pp 199-208.
- [34] Dimauro, G. Guerriero, A. Impedovo, S. Pirlo, G. A system for Bankchecks Processing. 2nd International Conference on Document Analysis and Recognition, 1993 pp 454 - 459
- [35] Dimov, D. An Approximate String Matching Method for Handwriting Recognition Post-Processing using a Dictionary. In: S. Impedovo (Ed.), Fundamentals in Handwriting Recognition, NATO ASI Series, Series "F": Computer and System Sciences, Vol. 124, Berlin: Springer-Verlag, 1994, pp. 323-332
- [36] Ding, Y. Okada, M. Kimura, F. Miyake Y. Application of Slant Correction to Handwritten Japanese Address Recognition, 6th International Conference on Document Analysis and Recognition, Seattle, 2001 pp 670-674
- [37] Ebadian Dehkordi, M. Sherkat, N. Allen, T. Case Classification of Off-line Hand-written Words Prior To Recognition. IAPR (2000).
-

-
- [38] El-Nasan, A, Veeramachanei, S. Nagy, G. Word discrimination based on bi-gram Co-occurrences, *Proc. International Conference Document Analysis and Recognition*, 2001.
- [39] Evans, R G. Sherkat, N. Whitrow, R J. Holistic Recognition of Static handwriting Using Structural Features. Document Image Processing and Multimedia (DIPM'99), IEE Colloquium 99/041, Pages 121-124 25th March 1999, London.
- [40] Fairhurst, M. Rahman, A. Generalised approach to the recognition of structurally similar handwritten characters using multiple expert classifiers. In IEE Proc. Visual Signal Process, 1997 Vol. 144, No. 1, pp 15-22
- [41] Fang, B. Wang, Y. Leung, C. Tse, K. Tang, Y. Kwok, P. Wong, Y. Offline Signature Verification by the analysis of Cursive Strokes. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 15, No. 4 (2001) 659 – 673.
- [42] Fry, E., Kress, J. Fountoukidis, D. *The Reading Teacher's Book of Lists*. Paramus, New Jersey: Prentice Hall, 2000.
- [43] Flemming, J. Hemmings, R. A Method of Recognition for Handwritten Block Capitals. *Pattern Recognition Letters* 1 (1983) 457 - 464
- [44] Garris, M. Unconstrained Handprint Recognition using a limited Lexicon. *Proc. SPIE Vol.218 No. 1 Document recognition (1994): 36 – 46*.
- [45] Garris, M. Blue, J. Candela, T. Dimmick, D. Off-line handwriting recognition from forms. *International conference on systems management and cybernetics: Intelligent systems for the 21st century*, 1995, pp 2783-2788
- [46] Garris, M. Dimmick, D. Form Design for High Accuracy Optical Character Recognition. *IEEE transactions on Pattern Analysis and Machine Intelligence*, Vol. 18, No. 6 June 1996
- [47] Garris, M. Wilson, C. Blue, J. Candela, T. Grother, P Janet, S. Wilkinson, R. Massively parallel implementation of character recognition systems. In *Proc SPIE Vol. 1661, 1992, 269 – 280*
- [48] Garris, M. Wilson, C. Wilkinson, R Methods for enhancing neural network handwritten character recognition. In *Proc. ICASSP, 1997, pp 887-890*
- [49] Gori, M. Scarselli, F. Are Multilayer Perceptrons Adequate for Pattern recognition and Verification? Published in *IEEE Transactions on Pattern Analysis and machine Intelligence*, Vol 20, No 11, November 1998.
- [50] Gorski, N. Anisimov, V. Augustin, E. Baret, O. Price, D. Simon, J. A2iA Check Reader: A Family of Bank Check Recognition Systems. 5th
-

- International Conference on Document Analysis and Recognition, 1999, pp 523-526
- [51] Govindaraju, V. Kim, G. Srihari, S. Paradigms in Handwriting Recognition. International Conference on System, Man and Cybernetics, 1997. pp 1498 - 1503
- [52] Grandidier, F. Sabourin, R. Yacoubi, A. Gilloux, M. Suen, C. Influence of Word Length on Handwriting Recognition. 5th International Conference on Document Analysis and Recognition, 1999, pp 777 – 780
- [53] Guillevic, D. Suen, C. HMM word recognition engine. Proceedings International Conference Document Analysis and Recognition, pp. 544-547, 1997
- [54] Harabagiu, S. Patterns of prepositional attachments – where dictionary semantics meets corpus statistics, International journal of pattern recognition and artificial intelligence, vol. 14, no. 6, 2000, pp 809 - 838
- [55] Hazen, T. Bazzi, I. A Comparison and Combination of Methods For OOV Word Detection and Word Confidence Scoring Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP) 2001, May 7 –11 2001, Salt Lake City.
- [56] Hazen, T. Burianek, T. Polifroni, J. Seneff, S. Recognition Confidence Scoring for use in Speech Understanding Systems. In Proc. ISCAASR 2000 Paris 2000.
- [57] Hennig, A. Sherkat, N. Cursive Script Recognition using wildcards and multiple experts. Pattern Analysis and applications Vol. 4., 2001, pp 51-60
- [58] Hennig, A. Sherkat, N. Whitrow, R. Zone-estimation for multiple lines of handwriting using approximation spline functions. Process in handwriting recognition, 1997, pp 63 - 67
- [59] Herbst, B. Coetzer, H On an offline signature verification system. Proceedings of the 9th Annual South African Workshop on Pattern Recognition. pp39-43 (1998).
- [60] Hudson, B. Objective Tests, Chapter 5 of Assessment techniques: an introduction, first published 1973 by Methuen Education Ltd
- [61] Impedovo, S. Frontiers in Handwriting Recognition. Fundamentals in Handwriting Recognition. Ed. Sebastiano Impedovo. New-York: Springer-Verlag, 1994.
- [62] Jiang, H. Han, C. Fan, K. A fast approach to detect and correct skew documents, In Proc. International conference on pattern recognition, 1996, pp 742 - 746

-
- [63] Jobbins, A. Evett, L. Semantic Information from Roget's Thesaurus: Applied to the Correction of Cursive Script Recognition Output. In Proc. Conference on Computing , Speech and Document Processing, 1998, pp 65 - 70
- [64] Kamppari, S. Hazen, T. Word and Phone Level Acoustic Confidence Scoring. Proceedings of the International Conference on Acoustics, Speech and Signal Processing , June 5-9 2000, Istanbul
- [65] Kavallieratou, E. Fakotakis, N. Kokkinakis, G. An unconstrained handwriting recognition system. . International Journal on Document Analysis and Recognition. Volume 4, Number 4, July 2002, Pages: 226 – 242
- [66] Kelland, S. Wesolkowski, S. A Comparison of Research and Production Architectures for Check Reading Systems. 5th International Conference on Document Analysis and Recognition, 1999, pp 99 – 102
- [67] Kilgarriff, A. Putting Frequencies in the dictionary, International Journal of Lexicography, 1996.
- [68] Kilgarriff, A. Using word frequency lists to measure corpus homogeneity and similarity between corpora. In proc. 5th ACL workshop on very large corpora, 1997
- [69] Kim, G. Lee, S. Shin, M. Nam, Y. Analysis of postal address fields for efficient encoding of Korean mail pieces, The proc. 6th international conference on Document Analysis and Recognition, Seattle, 2001 pp 675-679.
- [70] Kimura, F. Shridhar, M. Chen, Z. Improvements of a Lexicon Directed Algorithm for Recognition of Unconstrained Handwritten Words. 2nd International Conference on Document analysis and Recognition, 1993, pp 18 – 22
- [71] Koerich, A. Sabourin, R. Suen, C. Lexicon-driven HMM decoding for large vocabulary handwriting recognition with multiple character models. International Journal on Document Analysis and Recognition. Volume 6, Number 2, October 2003, Pages: 126 – 144
- [72] Kornai, A. An Experimental HMM-based postal OCR system. In: Proceedings of ICASSP'97, IEEE Computer Society Press, Los Alamitos CA, IV, 3177-3180
- [73] Kornai, A., K.M. Mohiuddin, and S.D. Connell, Recognition of Cursive Writing on Personal Checks, ms., 5th International Workshop on Frontiers in Handwriting Recognition, Essex, 1996 pp 373 – 378.
- [74] Kosmala, A. Rottland, Rigoll, G. Improved On-Line Handwriting Recognition Using Context Dependent Hidden Markov Models. 4th International Conference on Document analysis and Recognition. 1997, pp 641-644

-
- [75] Kosmala, A. Rottland, Rigoll, G. Large Vocabulary on-line handwriting recognition with context dependant Hidden Markov Models. *Advances in handwriting recognition*, chapter 2, 1999, pp 67-78
- [76] Kovacs, Z. A novel architecture for high quality hand-printed character recognition. In *Proc. Pattern Recogniton*, Vol 28. No, 11. 1995, pp. 1685-1692
- [77] Lam Suk Wah, L. Suen, C. Guillevic, D. Strathy, N, Cheriet, M. Liu, K. Said, J. Automatic Processing of Information on cheques. *Proc. 1995 IEEE Int. Conf. on Systems, Man, and Cybernetics*, Vancouver, Canada, Oct. 1995, 2353-2358
- [78] Lecolinet, E. Moreau, J. Off-line Recognition of Handwritten Cursive Script for the Automatic Reading of City Names on Real Mail. *International Conference on Pattern Recognition*, pp. 674-676, 1990
- [79] Lecolinet, E. Baret, O. Cursive Word Recognition: Methods and Strategies. In *Proc. NATO/ASI Fundamentals in Handwriting recognition*, 1993
- [80] Lecolinet, E. Cursive script recognition by backward matching. *Advances in handwriting and drawing: A multidisciplinary approach*, 1994, pp 117 – 135
- [81] Liang, S. Ahmadi, M. Shridhar, M. Segmentation of touching characters in printed document recognition. *2nd International conference on document analysis and recognition*, 1993 pp 569 – 572
- [82] Lu Y. Shridhar, M. Character segmentation in handwritten words – an overview. *Pattern Recogniiton*, Vol. 29, No. 1, 1996, pp 77-96
- [83] Lu, Y. On the segmentation of touching characters. *2nd international conference on document analysis and recognition*, 1993, pp 440 - 443
- [84] Liu, C. Nakagawa, M. Prototype Learning Algorithms for nearest neighbour classifier with application to handwritten character recognition, *5th International conference on document analysis and recognition*, 1999, pp 378-381
- [85] Lon-Mu, L. Babad, M, Sun, W. Ki-Kan, C. Adaptive Post-Processing of OCR Text via Knowledge acquisition. *Proceedings of the 19th annual conference on Computer Science*, 1991.
- [86] Lou, B. Foxely, E. STAMS – A Simple Text Assessment Marking System. *The AISB Conference of Computational Linguistics for Speech and Handwriting recognition* (1994).
- [87] Luttinger, L. Gertner, M. Learning Disorder: Written Expression. *eMedicine Journal*, April 30 2001, Volume 2, Number 4.
- [88] Madhvanath, S. krpasundar, V. Govindaraju, V. Syntactic methodology of pruning large lexicons in cursive script recognition. *Pattern Recognition Vol.34* (2001) 37 – 46.
-

-
- [89] Malburg, M. Comparative Evaluation of Technologies for Word Recognition Improvement by Incorporation of Syntactic Information. 4th International Conference on Document Analysis and Recognition, 1997, pp 784 - 787
- [90] Frattale, F. Mascioli, F. Parisi, R. Martinelli, G. RBF Neural Classifier hand-printed character recognition. IEEE 1995 Workshop on Nonlinear Signal and Image Processing, Neos Marmaras, Greece, June 20-22, 1995
- [91] Monger, D. Leedham, G. Downton, A. An Interactive Document Image Description for OCR of Handwritten Forms. 2nd International Conference on Document Analysis and Recognition, 1993, pp 524-527
- [92] Morita, M. Sabourin, R. Bortolozzi, F. Suen, C. A Recognition and Verification Strategy for Handwritten Word Recognition. 7th International Conference on Document Analysis and Recognition, 2003.
- [93] Nadal, C. Legault, R. Suen, C. Complementary algorithms for the recognition of totally unconstrained handwritten numerals. In Proc. 10th international conference on pattern recognition, 1990, pp 443 - 449
- [94] Otsuka, J.L. Rocha, H. An Agent-Based Approach to Support Formative Assessment. International Conference on Computers in Education (ICCE'02) . December 03 - 06, 2002, Auckland, New Zealand p1486 - 1487
- [95] Paik, J. Jung, S. Lee, Y. Multiple Combined Recognition System for Automatic Processing of Credit Card Slip Applications. 2nd International Conference on Document Analysis and Recognition, 1993 pp 520 - 523
- [96] Pain, R. Software tools to Enhance Handwriting Analysis in Bank cheque processing. Thesis submitted to the university of Kent at Canterbury, 1999
- [97] Payne, G. Isaacs, L. Human Motor Development; A life span approach, 4th Ed. Chap. 12 (1999) Mountain View California: Mayfield Publishing Company.
- [98] Pereira, P. Singer, Y. Tishby, N. Beyond N-grams. Proceedings of the Third Workshop on Very Large Corpora, 1995
- [99] Pereira, P. Heutte, L. Bougeois, O. Moreau, J. Plessis, B. Coutellemont, P. Lecourtier Y. Numeral Amount Recognition on Multiple-Bank Checks. In IEEE Proc. International Conference on Pattern Recognition, 1996, pp 165-169
- [100] Pitrelli, J. Perrone, M. Confidence-Scoring Post-Processing for Off-Line Handwritten-Character Recognition Verification, 7th International Conference on Document Analysis and Recognition, 2003, pp 278 - 282
- [101] Plamondon, R. Srihari, S. On-line and Off-line Handwriting Recognition: A Comprehensive Survey. PAMI vol. 22 no. 1 (2000); 63-8
-

-
- [102] Plessis, B. Sicsu, A. Heutte, L. Menu, Lecolinret, E. Debon, O. Moreau, J. A multi-classifier combination strategy for the recognition of handwritten cursive words, 1993, pp 642 – 645
- [103] Powalka, R. Sherkat, N. Whitrow, R. The use of word shape information for cursive script recognition. 4th international workshop on frontiers of handwriting recognition, 1994.
- [104] Powalka, R. Sherkat, N. Evett, L. Whitrow, R. Multiple word segmentation with interactive look-up for cursive script recognition. 2nd International Conference on document analysis and recognition, 1993, pp 196 – 199
- [105] Powalka, R. Sherkat, N. Evett, L. Whitrow, R. Dynamic Cursive Script Recognition: a hybrid approach. Advances in handwriting and drawing: A multidisciplinary approach, 1994, pp 137-154
- [106] Powalka, R. Sherkat, N. Whitrow, R. Zoning invariant holistic recogniser for hybrid recognition of handwriting, 3rd International Conference on document analysis and recognition, 1995, pp 64-67
- [107] Procter, S. Illingworth, J. Mokhtarain, F. Cursive handwriting recognition using hidden markov models and a lexicon driven level building algorithm. In IEE Proc.-Vis. Image Signal Process, Vol. 147, No 4, 2000, pp 332 - 339
- [108] Rahman, A. Fairhurst, M. Hoque, M. Paschalakis, S. A Novel Multiple-Expert Approach to the Recognition of Handwritten Words Extracted From British Cheques. IEE Proc. Vision, Image and Signal Processing, 147(4), pages 340-347, 2000.
- [109] Raza, G. Sherkat, Whitrow, R. Recognition of poor quality words without segmentation. International conference on Systems, Man and Cybernetics 1996, pp 64 –69
- [110] Resnik, P. Semantic Similarity in a taxonomy: an information-based measure and its application to problems of ambiguity in natural language. Journal of Artificial intelligence research 11, 1999 pp 95-130
- [111] Rose, T. Evett, L. Semantic Analysis for Large Vocabulary Cursive Script Recognition. Proceeding International Conference Document Analysis and Recognition, pp. 236-239, 1993
- [112] Schübler, M. Niemann H. A HMM-based System for Recognition of Handwritten Address Words , *in Proc. Sixth Int. Workshop on Frontiers in Handwriting Recognition*, Taejon, Korea, 1998, pp. 505 – 514.
- [113] Sen-Segundo, R. Pellom, B. Hacioglu, K, Ward W. Confidence Measures for spoken dialogue systems. ICASSP'2001, Mayo 5-11, Salt Lake City, Utah, USA

-
- [114] Seni, G. Srihari, R. Nasrbadi, N. Large vocabulary recognition of on-line handwritten cursive words. *Pattern analysis and machine intelligence* vol. 18 No. 7, 1996 pp 757 - 762
- [115] Senior, A. Off-line Handwritten Recognition: A review and experiments. Technical Report CUED/F-INFENG/TR 105, Cambridge University Engineering Department, December 1992.
- [116] Senior, A. Robinson, J An off-line cursive handwriting recognition system. *Pattern analysis and machine intelligence* vol. 20, No. 3, 1998, pp 309 – 321
- [117] Shapiro, S. The CASSIE projects: an approach to natural language competence, *EPIA 89: 4th Portuguese conference on artificial intelligence*, 1989, 362 –380
- [118] Shapiro, S. SnePS: a logic for natural language understanding and commonsense reasoning, Chapter of *Natural language processing and knowledge representation: language for knowledge and knowledge for language*, first published in 1999 by AAAI press.
- [119] Sherkat, N. Allen, T. J. Whole Word Recognition in Facsimile Images, 5th International Conference on Document Analysis and Recognition, 1999 547 – 550.
- [120] Shih, T. Chang, S. Huang , J. A Principled Approach for Formative Web Learning Assessment and Adaptive Tutoring. *First International Conference on Web Information Systems Engineering (WISE'00)-Volume 2*. June 19 - 20, 2000. Hong Kong, China. p 2196 – 2203
- [121] Shinghal, R. A bottom-up and top-down approach to using context in text recognition. In *Proc. International Journal on Man-Machine Studies* 11, 1979, pp 201 -212
- [122] Shridhar, M. Badreldin, A. Recognition of Isolated and simply connected handwritten numerals. *Pattern Recognition* Vol. 19, No 1. 1986, pp 1 – 12
- [123] Soon-man, C. Il-seok, O. A segmentation-free recognition of handwritten touching numeral pairs using modular neural network. *International Journal of Pattern Recognition and Artificial Intelligence* Vol. 15, No. 6 (2001) pp 949-966
- [124] Song, W. Feng, M. Shaowei, X. A Chinese Bank Check Recognition System Based on the Fault Tolerant Technique. 4th International Conference on Document Analysis and Recognition, 1997, pp 1038 – 1042
- [125] Srihari, S. Kim, G. PENMAN: a system for reading handwritten page images *Symposium on document image understanding technology*, 1997, pp 142-153

-
- [126] Srihari, S. Srihari, R. Written language recognition. In NSF-CEC Joint survey on speech and natural language processing, chapter 3, 1994
- [127] Srihari S. Handwritten Address Interpretation: A Task of Many Pattern Recognition Problems. *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 14, No. 4. (2000) 663 –674.
- [128] Srihari S.N, High-performance Reading Machines, *Vision Interface '90*, Halifax, Nova Scotia, Canada, May 1990, 150-155.
- [129] Srihari S.N, High-Performance Reading Machines. In *Proc. IEEE* Vol. 80, No. 7 July 1992.
- [130] Srihari S. N. Y. C. Shin, V. Ramanaprasad, and D. S. Lee, A System to Read Names and Addresses on Tax Forms, *Proc. of IEEE*, 84(7), July 1996, in press.
- [131] Srihari S. N. Y. C. Shin, V. Ramanaprasad, and D. S. Lee, Name and Address Block reader for Tax Form Processing. 4th International Conference on Document Analysis and Recognition, 1995, pp 5-10
- [132] Steinherz, T. Rivlin, E. Intrator, N. Off-line Cursive Script word Recognition – a survey. *International journal on document analysis and Recognition*, 1999, pp 90 – 110
- [133] Suen, C. Nadal, C. Mai, T. Legault, R. Lam, L. Recognition of totally unconstrained handwritten numerals based on the concept of multiple experts. 1st International Workshop on Frontiers in Handwriting Recognition, 1990, pp. 131 - 144
- [134] Suen, C. Y. Liu, K. and Strathy, N.W. Sorting And Recognising Cheques And Financial Documents (1998), *Document Analysis Systems*, pp 173 – 187.
- [135] Tappert, C. "Adaptive on-line Handwriting Recognition", *IEEE 7th International Conference on Pattern Recognition*, pp. 1004-1007, 1984
- [136] Tay, Y.H. Khalid, M. Knerr, S. Lallican, P.M. Viard-Gaudin, C. "Two-cost Stroke Segment Grouping Mechanism for Off-line Hand-written Check-Word Recognition", *AIAI'99*, 1999, Kuala Lumpur
- [137] Tsuruoka, S. Adachi, Y. Yoshikawa, T. A segmentation method for each text line for handwritten unconstrained document using thinning algorithm, *International Workshop on Frontiers in Handwriting Recognition*, 2000, pp 505 - 509
- [138] Ueda K. Extraction of Signature and Seal Imprint from bank checks by Using Color Information. 3rd International Conference on Document Analysis and Recognition, 1995, pp 665-668
-

-
- [139] Vermeulen, P. Barnard, E. Youghong, Y. Fanty, M. Cole, R. A Comparison Of HMM and Neural Network Approaches To Real World Telephone Speech Applications. IEEE international Conference on Neural Networks and Signal Processing, 1995. Pp 796 – 799.
- [140] Verma, B. Blumenstein, M. Kulkarni, S. Recent Achievements in Off-Line handwriting Recognition Systems. Proceedings of the International Conference on Computational Intelligence and Multimedia Applications (ICIMA'98) 1998, Melbourne, pp 27-33.
- [141] Vinciarelli, A. A Survey on Off-line Cursive Script Recognition. IDIAP Research Report 2000.
- [142] Vinciarelli, A. Luettin, J. Off-line cursive script recognition based on continuous density HMM. 7th International workshop on frontiers in handwriting recognition, 2000, pp 493-498
- [143] Weintraub, m. Beaufays, F. Rilin, Z. Konig, Y. Stolcke, A. Neural network Based Measures of Confidence for Word Recognition. Proc. ICASSP '97, Munich, pages 887-890.
- [144] Wesolkowski, S. Cursive Script Recognition: A Survey. Handwriting and Drawing Research: Basic and Applied Issues, 1996. pp 267-284
- [145] Wessels, T. Omlin, C. A Hybrid System For Signature Verification. Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks (IJCNN'00) 2000.
- [146] Woods, J. A comparative Review of Key Recognition Technologies, (1995). Recognition Technology for Data Entry: A Guide and Directory, Chapt 2.B: 14-18.
- [147] Wing, W. Sherkat, N. Allen, T.J. Form Extraction by Colour Information, (2000) In Proc. International Conference on Document Analysis Systems, pp 109 – 120.
- [148] Zimmermann, M. Chappelier, J. Bunke, H. Parsing N-Best Lists of Handwriting Sentences. 7th International Conference on Document Analysis and Recognition, 2003, pp 1- 5


**APPENDIX A - A SAMPLE OF ADULT'S HANDWRITTEN RESPONSES TO
MULTIPLE CHOICE QUESTIONS USED IN CHAPTERS 3 AND 4**

X2-540

Please could you write your answer in the spaces provided and in BLUE or BLACK Ink.

<p>1) In computing, what does DLL stand for:</p> <p>Data Linked List Dynamic Link Library Domain Level Limit</p> <p>Answer <u>Data linked List</u></p>	<p>5) In computing, what does ATM stand for:</p> <p>At The Machine Asynchronous Transfer Mode Actual Time Modulator</p> <p>Answer <u>Asynchronous Transfer Mode</u></p>
<p>2) In computing, what does GUI stand for:</p> <p>Generally Used Information Graphical User Interface Geographical User Internet</p> <p>Answer <u>Graphical User Interface</u></p>	<p>6) In computing, what does FTP stand for:</p> <p>File Transfer Protocol Fixed Text Post Forced Termination Prenaise</p> <p>Answer <u>File Transfer Protocol</u></p>
<p>3) In computing, what does BNC stand for:</p> <p>Basic Network Connection Blind Natural Copy British Naval Connector</p> <p>Answer <u>Basic Network Connection</u></p>	<p>7) In computing, what does QOS stand for:</p> <p>Query Operating System Quality Of Service Quick Online Search</p> <p>Answer <u>Query Operating System</u></p>
<p>4) In computing, what does AWE stand for:</p> <p>Advanced Wave Effect Access With Extension Alternative Wide Emulation</p> <p>Answer <u>Alternative Wide Emulation</u></p>	<p>8) In computing, what does JDK stand for:</p> <p>Java Development Kit Jackson Ducl Kernel Joint Device Knowledge</p> <p>Answer <u>Jackson Ducl kernel</u></p>

Thank you completing this test in advanced.
This will be used in a research project headed by Jonathan Allan.
To contact him for any reason e-mail him on ja@doc.ntu.ac.uk
At no time will this be used as an assessment of your abilities.



Data linked List

Graphical User Interface

Basic Network Connection

Alternative Wide Emulation

Asynchronous Transfer Mode

File Transfer Protocol


Query Operating System

Jackson Ducl kernel

Figure A-1: Completed test script for Student 40 and their original responses written in blue ink. The binarised responses that were passed for recognition can also be seen below the completed script

xl-536

Please could you write your answer in the spaces provided and in BLUE or BLACK ink.

<p>1) In computing, what does DLL stand for:</p> <p>Data Linked List Dynamic Link Library Domain Level Limit</p> <p>Answer <u>Dynamic Link Library</u></p> <p>2) In computing, what does GUI stand for:</p> <p>Generally Used Information Graphical User Interface Geographical User Internet</p> <p>Answer <u>Graphical User Interface.</u></p> <p>3) In computing, what does BNC stand for:</p> <p>Basic Network Connection Blind Natural Copy British Naval Connector</p> <p>Answer <u>Basic Network Connection.</u></p> <p>4) In computing, what does AWE stand for:</p> <p>Advanced Wave Effect Access With Extension Alternative Wide Emulation</p> <p>Answer <u>Advanced wave EFFECT</u></p>	<p>5) In computing, what does ATM stand for:</p> <p>At The Machine Asynchronous Transfer Mode Actual Time Modulator</p> <p>Answer <u>Asynchronous Transfer Mode.</u></p> <p>6) In computing, what does FTP stand for:</p> <p>File Transfer Protocol Fixed Text Post Forced Termination Premise</p> <p>Answer <u>File Transfer protocol</u></p> <p>7) In computing, what does QOS stand for:</p> <p>Query Operating System Quality Of Service Quick Online Search</p> <p>Answer <u>Query operating System.</u></p> <p>8) In computing, what does JDK stand for:</p> <p>Java Development Kit Jackson Deul Kernel Joint Device Knowledge</p> <p>Answer <u>Java Development kit</u></p>
<p>Thank you completing this test in advanced. This will be used in a research project headed by Jonathan Allan. To contact him for any reason e-mail him on ja@doc.ntu.ac.uk At no time will this be used as an assessment of your abilities.</p> 	

Dynamic Link Library

Graphical User Interface.

Basic Network Connection.

Advanced wave EFFECT

Asynchronous Transfer Mode.

File Transfer protocol

Query operating System.


Java Development kit

Figure A-2: Completed test script for Student 36 and their original responses written in black ink. The binarised responses that were passed for recognition can also be seen below the completed script

Please could you write your answer in the spaces provided and in BLUE or BLACK ink.

<p>1) In computing, what does DLL stand for:</p> <p>Data Linked List Dynamic Link Library Domain Level Limit</p> <p>Answer</p> <p><u>DOMAIN LEVEL LIMIT</u></p>	<p>5) In computing, what does ATM stand for:</p> <p>At The Machine Asynchronous Transfer Mode Actual Time Modulator</p> <p>Answer</p> <p><u>ASYNCHRONOUS TRANSFER MODE</u></p>
<p>2) In computing, what does GUI stand for:</p> <p>Generally Used Information Graphical User Interface Geographical User Internet</p> <p>Answer</p> <p><u>GRAPHICAL USER INTERFACE</u></p>	<p>6) In computing, what does FTP stand for:</p> <p>File Transfer Protocol Fixed Text Post Forced Termination Premise</p> <p>Answer</p> <p><u>FILE TRANSFER PROTOCOL</u></p>
<p>3) In computing, what does BNC stand for:</p> <p>Basic Network Connection Blind Natural Copy British Naval Connector</p> <p>Answer</p> <p><u>BLIND NATURAL COPY</u></p>	<p>7) In computing, what does QOS stand for:</p> <p>Query Operating System Quality Of Service Quick Online Search</p> <p>Answer</p> <p><u>QUERY OPERATING SYSTEM</u></p>
<p>4) In computing, what does AWE stand for:</p> <p>Advanced Wave Effect Access With Extension Alternative Wide Emulation</p> <p>Answer</p> <p><u>ADVANCED WAVE EFFECT</u></p>	<p>8) In computing, what does JDK stand for:</p> <p>Java Development Kit Jackson Dual Kernel Joint Device Knowledge</p> <p>Answer</p> <p><u>JAVA DEVELOPMENT KIT</u></p>

Thank you completing this test in advanced.
This will be used in a research project headed by Jonathan Allan.
To contact him for any reason e-mail him on ja@doc.ntu.ac.uk
At no time will this be used as an assessment of your abilities.



DOMAIN LEVEL LIMIT
GRAPHICAL USER INTERFACE
BLIND NATURAL COPY
ADVANCED WAVE EFFECT
ASYNCHRONOUS TRANSFER MODE
FILE TRANSFER PROTOCOL
QUERY OPERATING SYSTEM
JAVA DEVELOPMENT KIT

Figure A-3: Completed test script for Student 24 and their original responses written in green ink. The binarised responses that were passed for recognition can also be seen below the completed script

**APPENDIX B – A SAMPLE OF CHILDREN’S HANDWRITTEN
SINGLE WORD RESPONSES TO NFER–NELSON’S ‘PROGRESS IN
ENGLISH 6 EXAM, EXERCISE 2 PART A’ USED IN CHAPTER 4**

brdthor	next	happy	birthday	stopped
brdth	brkweel	stoppod	njrrr	meat
brother	next	happy	birthday	stopped
brothor	next	happy	binnday	stopped
brothor	next	happy	binnday	stopped
broxer	next	happy	birthday	stappd
birday	brother	stoppd	happy	next
brother	next	happy	birthday	stopped

Figure B-1: Samples of children's handwritten single word responses to NFER-Nelson's 'Progress in English 6 exam, Exercise 2 part A' used in chapter 4

**APPENDIX C – MODEL KEYWORD ANSWERS FOR NFER–NELSON'S
'PROGRESS IN ENGLISH 10, EXERCISE 5 THE TUNNEL' USED IN CHAPTER
5**

**APPENDIX D - A SAMPLE OF CHILDREN'S HANDWRITTEN SENTENCE
RESPONSES TO NFER-NELSON'S 'PROGRESS IN ENGLISH 10 EXAM
PAPER, EXERCISE 5 THE TUNNEL' USED IN CHAPTER 5**

The animal is a dragon -

The sentence made the train sound like a lion.

Paragraph 5, line 2 the animal is a dragon.

The dragon is the animal who came out of the tunnel,

It makes the train sound like a lion

The sentence is making sound like a lion

The noise what the train makes is same as a lion.

The animal sounds like a dog

The animal is a dragon

The sentence makes the sound like a dragon.

The sentence makes the train sound like a lion

The animal is a lion

The train sounded like a dragon

Figure D-1: A sample of the binarised children's handwritten sentence response images to question 1 from NFER-Nelson's 'Progress in English 10 exam paper, Exercise 5 The Tunnel'

The engines were snorting and puffing

The smoke was snorted and puffed out

The thing that was snorted and puffed out of the steam train was smoke.

engine snorted and puffed out by the steam.

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

The engines were snorting and puffing

Smoke and steam was snorted and puffed out.

Smoke was puffed out.

The engines were snorting and puffing

Smoke was snorted and puffed out of the steam engine.

The steam-engine snorted + puffed

Steam is what was snorted and puffed out of the train.

The train was snorting and puffed out steam

Figure D-2: A Sample of the binarised children's handwritten sentence response images to question 2 from NFER-Nelson's 'Progress in English 10 exam paper, Exercise 5 The Tunnel'

When it is dark

They might have looked like serpents because their long black and rusty.

The lines might have looked like serpent because the lines are straight as well.

It was looking from an angle.

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

The railway lines looked like serpents because they are long and black.

The lines look like described like serpents because they looked like snakes.

Because he was looking from an angle.

Because they curve and in smoke it looks like the lines do move.

They looked like serpents because they were black.

The railway lines looked like serpents because they were straight.

Figure D-3: The binarised children's handwritten sentence response images to question 3 from NFER-Nelson's 'Progress in English 10 exam paper, Exercise 5 The Tunnel'

When it is roaring

The vibration of the wheels on the tracks

The train might have made a clinking and screeching sound.

The train made a rumbling sound because it was moving fast on the lines.

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

The train sounded like thunder because it was rattling on the railway lines.

They made the sound because of the wheels.

When it was chugging

Again the engines made the sound like distant thunder

It made its sound like distant thunder because

The train was screeching like distance thunder.

Figure D-4: The binarised children's handwritten sentence response images to question 4 from NFER-Nelson's 'Progress in English 10 exam paper, Exercise 5 The Tunnel'

because it didn't move

It was lazy because it floated just above the trees, doing ~~anything~~

The smoke was described as being lazy because it would hover around after the train had gone

being lazy because it was not doing anything just laying in the air

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

The smoke was described as being lazy because it just stayed there traveling from side to side

The smoke was being lazy by not going away.

Because it was all black

because it was just hovering and moving slowly.

The smoke was described lazy because it does ~~doesn't~~ do anything.

The smoke was just drifting along.

Figure D-5: The binarised children's handwritten sentence response images to question 5 from NFER-Nelson's 'Progress in English 10 exam paper, Exercise 5 The Tunnel'

APPENDIX E – PAPERS PUBLISHED FROM THE WORK CARRIED OUT DURING THE AUTHOR’S PHD

The author has currently published four conference papers in various international proceedings:

1. Automated Assessment: It’s Assessment Jim But Not As We Know it [3]
2. Automated Assessment: How Confident Are We? [4]
3. Confident Assessment of Children’s Handwritten Responses [5]
4. Automated Assessment of Children’s Handwritten Sentence Responses [6]

Automated Assessment: It's Assessment Jim But Not As We Know It.

Jonathan Allan, Tony Allen, Nasser Sherkat, Peter Halstead
IRIS, Department of Computing, The Nottingham Trent University
Burton Street, Nottingham, NG1 4BU, U.K.
Tel: (+44)0115-848-2150 Fax: (+44)0115-848-6518
Email: {ja, tj, ns, ph}@doc.ntu.ac.uk.

Abstract

An extensive literature survey on automated assessment and handwriting recognition has shown that no work has been done in addressing the area of assessment of handwritten exam scripts. This paper therefore introduces the novel concept of applying Image Extraction and Cursive Script Recognition (CSR) techniques to the area of automated assessment. We demonstrate the potential for using a holistic CSR engine as the input process for a system capable of automatically scoring handwritten responses to multi-choice questions. This innovative system utilises the constrained nature of simple multiple choice questions to enhance the recognition rate of the handwritten response. 50 writers were chosen to answer eight multiple choice questions and results show that the system yields an average 83% CSR word accuracy, which enables the system to score over 54% of all response with 99% Confidence.

1. Introduction

At the forefront of the assessment of on-line data is the commercial group behind QuestionMark [1]. This software has the ability to generate tests that accept on-line yes/no, single word, multi-word and multi-line responses. Its recognition accuracy is very good when dealing with yes/no or single word answers but, it does have a limited assessment capability when scoring single line and multi-line responses. Research has been done in applying NLP (Natural Language Processing) techniques to the user inputs in order to reduce the ambiguity found within these responses [2-3] and this helps improve the scoring confidence for multi-line responses. However, if the users are incapable of relating their thoughts efficiently via a keyboard or the cost of class testing done on a computer versus the traditional pen & paper is too

high, then a move from paper based assessment to on-line testing is impractical.

The automated assessment of handwritten responses is therefore the obvious progression. The constrained nature of assessment structure can be used to improve recognition so that the confident scoring of scripts is achieved.

2. Automated Assessment

CSR 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. 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.

It has been shown that with the use of a constrained lexicon, handwriting recognition can be improved [4]. This can be exploited in many areas and therefore it is not always necessary to work with the notion of recognising totally free unconstrained text. For instance, the recognition of addresses is one area where the concept of unconstrained text would be superfluous. Off line recognition methods such as postal address recognition, signature verification [5] and Optical Mark Recognition (OMR) are also other areas where the benefits of constraining the scope of recognition can be shown to enhance recognition.

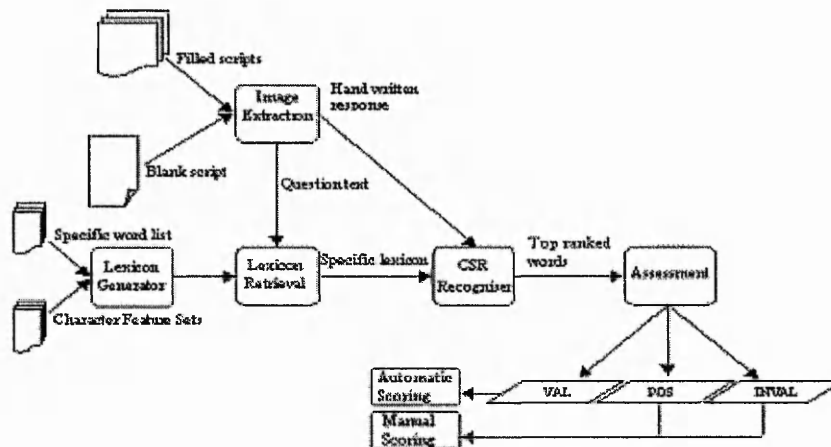


Fig 3: Handwriting Recognition System

OMR was first used to score multiple choice scripts in 1963 [6] and is still used today to mark exams all over the world. However the use of this technology is limited and some of the pupils that are tested may not be able to efficiently translate their response to the question by marking a separate piece of paper. Many young children are therefore still required to supply a hand-written answer when taking a formal exam.

written there. Use of this positional based contextual knowledge will therefore enable limited lexicons to be selected prior to recognition which, in turn, will improve the accuracy of the recognition process and ultimately the assessment process.

3. Handwriting Recognition System

Fig 3 shows the proposed system. It incorporates our existing Image Extraction system and current CSR technology as well as novel assessment technologies to automatically score handwritten responses. These areas work independently, as shown in Fig 3, and are described in sections 3.1 to 3.3 respectively.

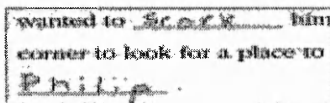


Fig 1: 1st example of an exam responses



Fig 2: 2nd example of an exam responses

Figs 1 and 2 show typical examples of the assessment formats used to assess the literacy of 7-8 year old children in the UK. As can be seen, the examinee is often expected to write their answer in a variety of locations on the script that are not predefined by a box or line following a field prompt (as in the case of conventional forms). However uniformity within exam scripts is there to be found. They possess similar features, i.e. exam scripts supply fields that demand a responses in certain areas on the script. However pronounced that field is, it is still a field in a form and will contain contextually linked data by virtue of the associated question. This can clearly be seen in fig 1 and fig 2. The handwritten responses that can be seen do not have a direct prompt (in this case a question directly in front of them) and the fields themselves are somewhat ambiguous. However the location of the handwritten response within the script does imply something about the possible responses that could be

3.1 Image Extraction

The use of colour within our exam script was felt to be important. Research shows that data extraction from forms that have a coloured structure is more accurate than that of the more traditional black and white form [8]. Thus the script structure and text used in this work was printed in red ink and the students were given instructions to enter their responses using only blue or black ink. The actual processing was done using a method devised by Wing *et al*. The scripts underwent two stages before they were ready for recognition, Quantisation and data extraction. The Quantisation process reduces the number of colours within the script whilst the data extraction process uses the colour information to directly extract the hand written responses from the image.

3.2 Cursive Script Recognition

CSR was then carried out using our existing HVBC (Holes, Vertical Bars and Cups) recogniser [9] with 3 lexicons each comprising of 9 words. These words consisted of the 3 different words that could be written in a given position in the responses plus their case alternatives, i.e. DYNAMIC, Dynamic and dynamic. The holistic recogniser works by recognising the shape of the word from features extracted from the whole word image. Each character is defined in terms of three features (Holes, Vertical Bars and Cups). Evans *et al* [9] describes this process in detail for the lower-case characters. However, the use of a lower case only lexicon was found

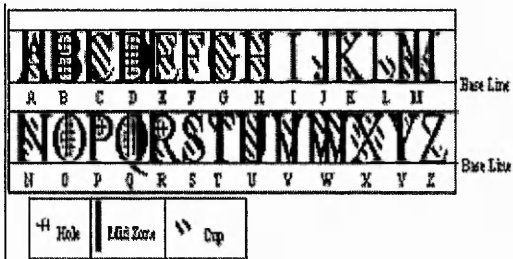


Fig 5: Example of upper-case zoning



Fig 4: The upper-case test set

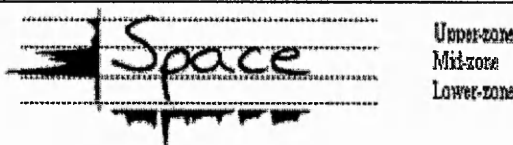


Fig 6: Example of mixed case zoning

not to be sufficient to recognise the range of responses found. Therefore it was deemed necessary to add an upper-case only and mixed case character sets to the generation stage. Fig 4 shows the upper case only character set used in this work. It follows a similar format to that used by Evans *et al*. Figures 5 & 6 show examples of the effects upper case writing has on the zoning procedure used by the HVBC recogniser.

Once the character sets had been defined the three 9 word lexicons were generated. These lexicons are specific to each question and each lexicon is unique to each word position in the response. The output images from the image extraction process were manually marked up and passed to the recogniser. An alternative word list was produced for each word image using a manually chosen lexicon.

3.3 Assessment

50 first year computing students were given a script which was made up of eight multi-choice questions. Each question had three responses, one correct and two distracters. Each response has three words and the first question can be seen in fig 7.

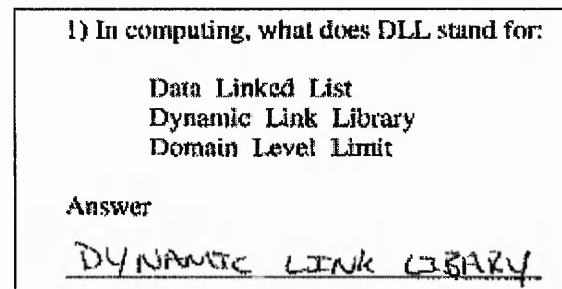


Fig 7: Question 1 of 8 from writer 10

Before processing by the HVBC recogniser was attempted, the number of words comprising each response was manually determined. If the sample contained no response or a numbered response (see fig 8) it was classified as INVALID before being passed to the recogniser. In principle these INVALID responses could be automatically pre-classified by the system as it can be programmed to reject answers that are not made up of three independently marked-up word responses.

7) In computing, what does QOS stand for:

Query Operating System
Quality Of Service
Quick Online Search

Answer

1

Fig 8: an example of an INVALID question due to a numbered response.

Since manual marking is employed, word isolation accuracy is 100% and therefore the rejection rates of the INVALID responses due to 'numbered' and 'no response' cases is also 100% accurate.

Once the written response has been passed through the recogniser, the recognised response is formed from the top ranked words in each word position and evaluated against the target response in order to classify it into one of three classes, VALID, POSSIBLE and INVALID. This is done by looking at the number of bridges within each of the recognised responses.

A valid bridge can be formed when two of the words in the recognised response correspond to a pair of words in one of the three target responses. For instance, using the actual response in Fig-7 here are examples of how each class could be generated and determined:

A pair of valid bridges between the words in the recognised response leads to a VALID classification. E.g.

Recognised Response: **Dynamic Link Library**

This is an example of a GOOD VALID classification as all words in the written response have been correctly recognised. However errors can be introduced into the recognised response when the written response has been incorrectly recognised and two valid bridges are still formed. A BAD VALID classification can be formed when the recognised response is:

Domain Level Limit

Or

Data Linked List

A single valid bridge in the recognised response leads to a POSSIBLE classification. E.g.

Recognised Response: **Domain Link Library**

Or

Dynamic Linked Library

Or

Dynamic Link List

Again these examples show the GOOD POSSIBLE recognised word responses. Errors can also be introduced that result in the classification process producing a BAD POSSIBLE class in the same way as in the creation of the BAD VALID class. Examples of this would be:

Domain Level Library

Or

Data Linked Library

In these examples the words in the first two positions have been recognised incorrectly, but have also formed a valid bridge creating a BAD POSSIBLE response.

No valid bridges present in the recognised response leads to an INVALID classification. In this instance 'Link'

is the correctly recognised word but since there are no bridges formed there must be an error in the recognition stage and the response is therefore rejected. For example:

Recognised Response: **Domain Link Level**

4 Experimental Results

50 writers were tested, producing 359 complete responses (1077 word images). Fig 9 shows the results of the response classification. It shows that the system classified 54% of all responses as VALID and did this with a scoring confidence greater than 99%. The POSSIBLE and INVALID class responses were rejected at this time as they could be assumed to be passed for manual marking.

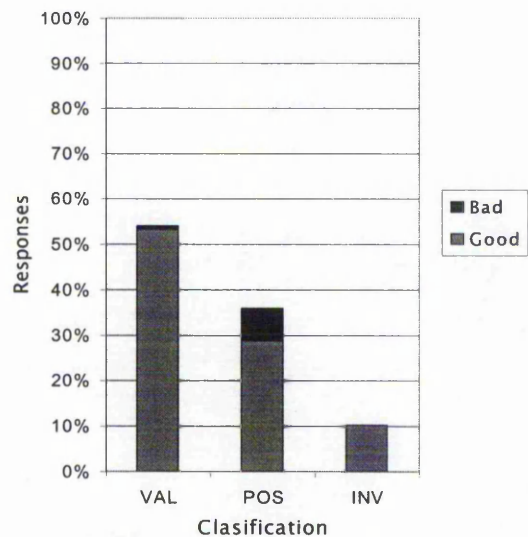


Fig 9: Response Classification

Fig 10 illustrates the individual word accuracy of the recogniser for the experiment. The overall top rank recognition accuracy using a lexicon of 9 words for each word response was 83%. This rises to 99% when the top three ranked word responses are considered.

As expected the recognition rate is high due to the constrained nature of the test set, however two issues have arisen which, if solved would improve recognition accuracy. Firstly the performance of the recogniser on upper case words and secondly errors which have not been detected at the mark up stage.

HVBC performance relies heavily upon vertical bars that ascend and descend out of the mid zone. In upper case words these vertical bars are not present thus causing

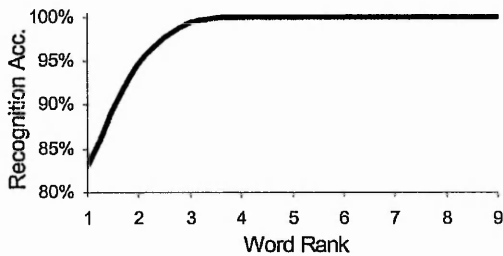


Fig 10: Recognition Rate

more confusion. Upper case writers formed 16% of the data set and when excluded the top case recognition rate increases to 89%.

Writers, have in some cases, opted to abbreviate some of their responses thus giving a word that is not expected and therefore not in the lexicon (Fig 11 shows an example of an abbreviated response). These writer errors, unlike the numbered responses, could not be detected by an automated mark up procedure because three independent word images can still be formed. Manual checks have shown that this sort of error amounts

8) In computing, what does JDK stand for:

**Java Development Kit
Jackson Ducl Kernel
Joint Device Knowledge**

Answer

Java Development Kit

Fig 11: An Example of an Abbreviated Response

to 1% of the total word responses.

5 Conclusions

This paper aims to open the debate in the area of automated assessment of handwritten scripts by identifying the research issues involved in using CSR as the input process for an automatic assessment system. The results of this initial investigation show that automated assessment using CSR is far from being as accurate as that of human recognition and judgement. However the assessment algorithm introduced in this paper is seen to be capable of adequately coping with the errors introduced by CSR. Constraining the response of the user is shown to help reduce CSR errors and enable assessment. This is indicated by the overall performance of correctly scoring 54% of all responses with a scoring confidence of 99%. Future work will focus on improving the accuracy of recognition and assessment (through constraining the scope) and will investigate the trade off

between recognition rate and constraints in order to establish the boundaries of applying CSR to automated assessment.

Acknowledgements

We would like to thank NFER-Nelson for their support over the past twelve months.

References

- [1] QuestionMark™ www.questionmark.com
- [2] Allott, N. A Natural Language Processing Framework for Automated Assessment. (1999) *PhD thesis, The Nottingham Trent University*
- [3] Lou, B. Foxely, E. STAMS – A Simple Text Assessment Marking System. (1994) *The AISB Conference of Computational Linguistics for Speech and Handwriting recognition.*
- [4] Cai, J. Liu, Z. Off-line unconstrained Hand Written Word Recognition (2000). *International Journal of Pattern Recognition and Artificial Intelligence, Vol 14, No3: 259-280.*
- [5] Plamondon, R. Srihari, S. On-line and Off-line Handwriting Recognition: A Comprehensive Survey. (2000) *PAMI vol. 22 no. 1: 63-84*
- [6] Woods, J. A comparative Review of Key Recognition Technologies, (1995). *Recognition Technology for Data Entry: A Guide and Directory, Chapt 2.B: 14-18.*
- [7] Robert M. Haralick., Document Image Understanding: Geometric and Logical Layout, (1994). *CVPR94: 385-390.*
- [8] Wing, W. Sherkat, N. Allen, T.J. Form Extraction by Colour Information, (2000) *DAS: 109 – 120*
- [9] 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*
- [10] Sherkat, N. Allen, T. J. *Whole Word Recognition in Facsimile Images, (1999). ICDAR 5th International: 547 – 550.*

Automated Assessment: How Confident Are We?

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

Abstract

This paper highlights the research issues associated with the automated assessment of handwritten scripts and introduces the notion of Scoring Confidence[1]. Using this concept, in a 3 word response environment, we prove that it is theoretically possible to achieve a Scoring Confidence greater than 98% using recognition rates as low as 81% to produce actual response yields of 50%. These results are verified by experiment.

1 Introduction

Cursive Script Recognition (CSR) has many inherent difficulties that range from coping with a wide variety of handwriting styles to the complexity of recognising multi-word combinations that cause ambiguity. High accuracy, as measured by recognition rates, is the aim of all research within the field of handwriting recognition and it has been well documented that rates of 87% and greater can be achieved [2-8,10-13]. However to gain such rates, the experiments have to be constrained. The areas that are constrained or pre-classified are one or more of the following: lexicon[3,5-6,10-12]; writers[2,4,7]; and even handwriting styles[13]. Unfortunately 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.

Automated assessment is one application of recognition that can use the constraints of the task to achieve very high recognition rates. Assessment is by its very nature constrained; only a limited number of responses should be expected to a given question. The main restriction that arises from this type of application is that it needs to be extremely accurate. The price of a recognition error is much higher than any savings that can be gained from automatically scoring a correctly recognised response. As has

been said before, existing recognition systems can correctly recognise only 87% of all responses [10]. Thus, without any post-recognition error detection, 13% of all responses will be recognised incorrectly. However, if these invalid responses can somehow be detected and filtered out for manual processing before automatic assessment is applied then the assessment accuracy of the system can be increased. Therefore automatic systems should be able to refuse to assess a response when the probability of making a mistake is too high, thus only passing on a response when it is known that the recognition system is highly confident of the response being correct.

2 Theoretical Confidence Method

6) In computing, what does FTP stand for:

- File Transfer Protocol
- Fixed Text Post
- Forced Termination Premise

Answer

FILE TRANSFER PROTOCOL

Fig 1. Example of the question and response style, 3x3,3,3

At The Nottingham Trent University first year computing students are asked to fill in an information sheet that gives an idea of how much experience and prior knowledge of computers and computing they already possess. This then helps the tutors assess who may need further help in the future. An example of the questions that are used can be seen in figure 1. It can be seen that this question is a multi-choice question with 3 possible responses that are created from 3 different words, i.e. this is a 3x3,3,3 response style question.

Since all the possible responses are known prior to recognition, it is possible to identify each word location and thus reduce the size of the lexicon used for the recognition of each word by making it word placement specific. For example, the question shown in figure 1, has three responses: 'File Transfer Protocol', 'Fixed Text Post' and 'Forced Termination Premise'. In the first word position it is only possible to have either: 'File', 'Fixed', 'Forced' or a writer error. Therefore the lexicon for this word position will be made up of the three previously stated words plus their case alternatives, making a total of 9 words per word position.

One other advantage that can be gained from knowing what responses are expected is that the three answers are unique. The syntactic structure of the 3 recognised words can thus be used to determine whether or not the three separately recognised words can form one of the answers i.e. a VALID response. Equally INVALID and POSSIBLE responses can be determined by looking at the number of bridges within each of the recognised responses[1]. Appendix A shows the complete break down of the classes and their GOOD & BAD designation in this 3x3,3,3 style of response.

Since the classes have been defined and the recognition rate is known it is possible to theoretically determine the accuracy and the Scoring Confidence of the system. If the average word recognition rate for the system across all word positions is R then the probability for correctly recognising a VALID response is defined by equation (1).

$$\% \text{ of GOOD VALID} = R^3 \quad - (1)$$

The probability for incorrectly recognising a word is 1-R. Therefore for the system to miss-recognise the written response as another viable response (i.e. written response: "File Transfer Protocol" and recognised response is "Forced Termination Premise") the probability will be (2).

$$\% \text{ of BAD VALID} = 0.25(1-R)^3 \quad - (2)$$

The factor of one quarter results from the $3/4:1/4$ split between the BAD POSSIBLE and BAD VALID classification when all three words are miss-recognised.

Since the system is unable to differentiate between a GOOD VALID and BAD VALID response these combine to produce the Theoretical Valid Response Yield (TVRY), (3).

$$\text{Theoretical Valid Response Yield} = (R^3 + 0.25(1-R)^3) \quad - (3)$$

The Scoring Confidence, which is the amount of confidence in the assessment procedure, is then defined as the proportion of the VALID responses, passed for automatic assessment, which are GOOD VALID responses. Thus the Scoring confidence for the VALID class is defined as the probability of producing a GOOD VALID responses over the Theoretical Valid Response Yield, (4).

$$\text{Scoring Confidence} = \frac{R^3}{(R^3 + 0.25(1-R)^3)} \quad - (4)$$

From these equations a graph can be produced to show the Scoring Confidence against the corresponding TVRY for all possible recognition rates, R.

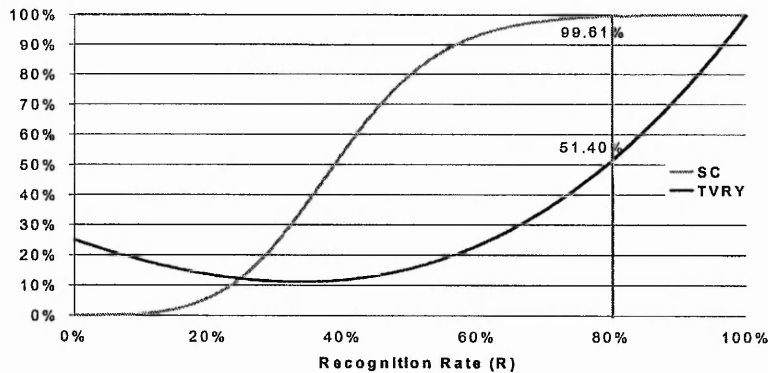


Fig 2. A graph to show the Scoring Confidence in relation to the TVRY for the 3x3,3,3 style response.

From figure 2, it can be seen that to achieve the industry standards of 50% response yield with an error rate less than 0.5% [5] a theoretical recognition rate of around 80% is required for this question.

The major significance of this method is its the portability. For instance, using the same principles, we can prove that the scoring confidence for a 3x3,2,3 response style question is given by (5)

$$\text{Scoring Confidence} = \frac{R^3}{(R^3 + 0.5(1-R)^3)} \quad - (5)$$

2) In computing, what does GUI stand for:

Generally Used Information
Graphical User Interface
Geographical User Internet

Answer

Graphical User Interface

Fig 4. Example of the question and response style, 3x3,2,3

An example of a 3x3,2,3 response style seen in figure 4. Again it is a multi-choice question, with 3 possible responses consisting of 3 different words in the first position, 2 different words in the second and 3 different words in the third. The overall classification break down for this response style is given in Appendix B.

The equation for calculating the probability of recognising a GOOD VALID response is the same as the previous example (R^3) as there are still three word positions in both examples. However the equation for calculating the number of BAD VALID responses is slightly different, $BAD\ VALID = 0.5(1-R)^3$. This is due to the fact that the probability of producing a BAD VALID response as opposed to an INVALID response has increased because of the increase in ambiguity this question style imposes i.e. less bridges can be formed and therefore there is a decrease in the possibility of confidently recognising a response.

As in the previous example a graph can be produced from equation (5). This is shown in figure 4. Here it can be seen that to achieve a Scoring Confidence of 99.5% a Recognition Rate of 87% is required. This increase is also attributed to the increase in the question style ambiguity. However with an increased Recognition Rate there will also be an increase in the Theoretical Valid Response Yields.

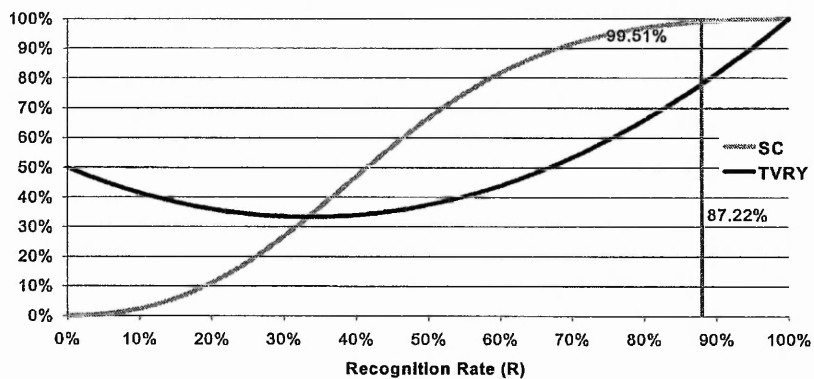


Fig 4. A graph to show the Scoring Confidence in relation to the TVRY for the 3x3,2,3 style response.

3 Results

In order to verify our theoretical prediction, automated assessment was performed on 8 question responses from 50 first year computing students. Before processing with our

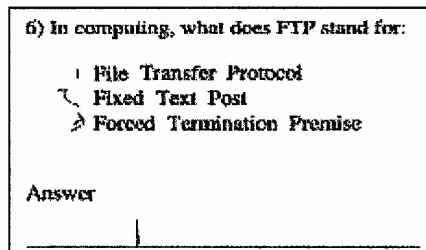


Fig 5. an example of an INVALID question due to a numbered response.

existing holistic recogniser [14], the handwritten responses were manually extracted from the scripts. During this process the number of words comprising each response was manually determined. If the sample contained no response or a response consisting of less than three words (see figure 5) then it was classified as INVALID. In principle these INVALID responses could be automatically pre-classified by the system as it can be programmed to reject answers that are not made up of three independently marked-up word responses. Note: Since manual

marking is employed, word segmentation accuracy is 100% and therefore the rejection rates of the INVALID class responses (i.e. not three words) is also 100% accurate.

Once the written responses had been passed through the recogniser, the recognised responses were formed from the top ranked words in each word position and evaluated against the target response in order to classify it into one of the three classes, VALID, POSSIBLE and INVALID.

3x3,3,3 Response Style Results

Seven 3x3,3,3 response style questions were given to the 50 students producing 313 complete responses (939 word images). The system classified 50% of all responses as VALID and did this with a scoring confidence greater than 98%. The POSSIBLE and INVALID class responses were rejected at this time as they could be assumed to be passed for manual marking. This compares well against the theoretical results (see table 1), where the theoretical results were calculated as a Scoring Confidence of 98.7% with a TVRY of 53% when the system achieves a recognition rate of 81%.

	Theory	Actual
Rec. Rate	81%	81%
Scoring Confidence	98.7%	98.3%
Response Yield	53.1%	50.3%

Table 1: Comparison table for the response style 3x3,3,3.

3x3,2,3 Response Style Results

One 3x3,2,3 response style question was given to the 50 students. The average recognition rate for this experiment was 93% and a comparison table can be drawn up between the expected theoretical results and the actual results, see table 2.

	Theory	Actual
Rec. Rate	93%	93%
Scoring Confidence	99.9%	100%
Response Yield	82%	78%

Table 2: Comparison table for the response style 3x3,2,3

With a recognition rate of 93% the theoretical probability of producing a BAD VALID is only 0.03%, therefore the likelihood of recognising one is very small. In fact, a Scoring Confidence of 100% was achieved because the data set was not large enough for the BAD VALID error to actually be quantified.

As expected the recognition rate is high in each case due to the constrained nature of the data set, however an issue has arisen. A decrease can be seen in the actual response yield when compared to the theoretical response yield. The actual response yield is seen to decrease because of other errors introduced by the writers giving an INVALID response (written response was 'Dynamic Linked Library' when the expected responses was 'Data Link Library') and in some cases writers opted to abbreviate their responses thus giving a word that is not expected and therefore not in the lexicon ('Java Dev Kit' instead of 'Java Development Kit'). These writer errors, unlike the numbered responses, could not be detected by an automated mark up procedure because three independent word images can still be formed. Manual checks have shown that this sort of error amounts to 2% of the total word responses and in the majority of case the recognised response formed is either a BAD POSSIBLE or INVALID, thus decreasing the actual response yield.

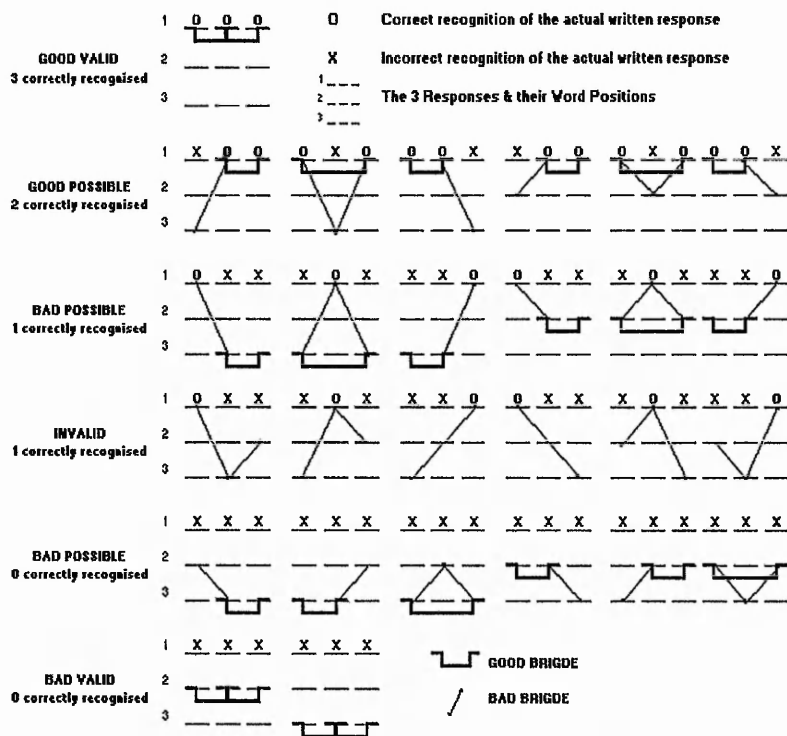
4 Conclusion

This paper aims to highlight the need for reliable confidence measures to be used in the automated assessment of cursively written scripts. It also opens the debate in the area of automated assessment of handwritten scripts by identifying the research issues involved in using CSR as the input process for an automatic assessment system. The results of this initial investigation show that automated assessment using CSR is far from being as accurate as that of human recognition and judgement. However, with the use of the theoretical Scoring Confidence introduced in this paper the system is proven to be capable of adequately coping with the errors introduced by CSR.

References

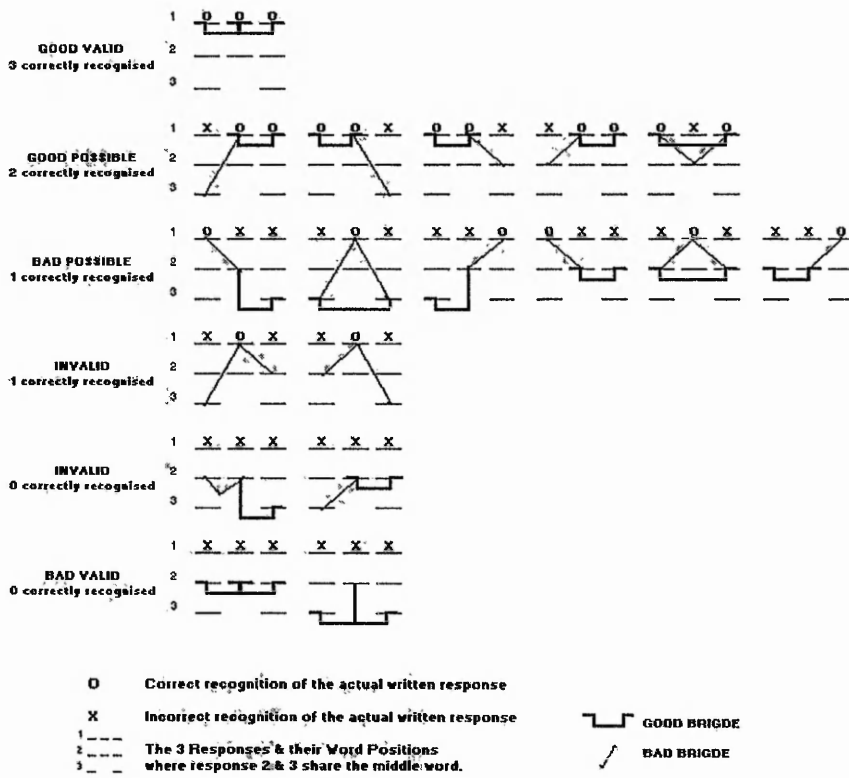
- [1] Allan, J. Allen, T. Sherkat, N. **Automated Assessment: It's Assessment Jim But Not As We Know It.** Proceedings of the 6th International Conference on Document Analysis and Recognition, Sept 10-13, 2001 pp
- [2] Cai, J. Liu, Z. **Off-line unconstrained Hand Written Word Recognition (2000).** *Int. Journal of Pattern Recognition and Artificial Intelligence, Vol 14, No3: 259-280*
- [3] Plamondon, R. Srihari, S. **On-line and Off-line Handwriting Recognition: A Comprehensive Survey. (2000) PAMI vol. 22 no. 1: 63-84**
- [4] Blumenstein, M. and B. Verma, 1997, **A Segmentation Algorithm used in Conjunction with Artificial Neural Networks for the Recognition of Real-World Postal Addresses,** *Journal of Intelligent Systems, 8, 1-13.*
- [5] Srihari S.N, **High-performance reading machines,** *Vision Interface '90,* Halifax, Nova Scotia, Canada, May 1990, 150-155.
- [6] Srihari S. N. Y. C. Shin, V. Ramanaprasad, and D. S. Lee, **A System to Read Names and Addresses on Tax Forms,** *Proc. of IEEE, 84(7),* July 1996, in press.
- [7] Suen CY. Lui, K. Strathy, N. **Sorting and Recognizing Cheques and Financial Documents.** *Document Analysis Systems.* pp173-187 1998.
- [8] Di Lecce, V. DiMauro, G. Guerriero, A. Impedovo, S. Pirlo, G. Salzo, A. **A New Hybrid Approach For Legal Amount Recognition.** Proceedings of the 7th Int. Workshop on Frontiers in Handwriting Recognition, September 11-13 2000, Amsterdam, pp 199-208.
- [9] Woods, J. **A comparative Review of Key Recognition Technologies, (1995).** *Recognition Technology for Data Entry: A Guide and Directory, Chapt 2.B: 14-18.*
- [10] Senior, A. Robinson, A. **An Off-line Cursive Handwriting System.** *IEEE PAMI, Vol. 20, no 3 (1998):309 – 321.*
- [11] Garris, M. **Unconstrained Handprint Recognition using a limited Lexicon.** *Proc. SPIE Vol.218 No. 1 (1994): 36 – 46.*
- [12] Madhvanath, S. krpasundar, V. Govindaraju, V. **Syntactic methodology of pruning large lexicons in cursive script recognition.** *Pattern Recognition Vol.34 (2001) 37 - 46*
- [13] Ebadian Dehkordi, M. Sherkat, N. Allen, T. **Case Classification of Off-line Hand-written Words Prior To Recognition.** *IAPR (2000)*
- [14] Evans, R G. Sherkat, N. Whitrow, R J. **Holistic Recognition of Static handwriting Using Structural Features (DIPM'99),** *IEE Colloquium 99/041, Pages 121-124 25th March 1999, London*

Appendix A



The complete breakdown of the 3x3,3,3 style response classifications

Appendix B



The complete breakdown of the 3x3,2,3 style response classifications

Confident Assessment of Children's Handwritten Responses.

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

Abstract

This paper introduces a novel approach for the automatic assessment of children's responses to standardised English exam questions. The constrained nature of the question and answer medium is exploited to produce an automatic assessment mechanism that is both highly accurate and produces a reasonable level of response yield. It is shown that the novel approach can achieve 100% scoring accuracy on 44% of all responses compared to a traditional lexical approach that has an error rate of 41%. When a thresholding method, similar to that used in the novel approach is applied, the traditional approach can achieve an accuracy of 100% but with a response yield of only 5%. The approach introduced in this paper is thus shown to have a significant advantage over the traditional lexical based assessment.

1. Introduction

This paper has two aims. This first is to continue the research into defining the problems associated with the automated assessment of cursively written scripts and the second is to test whether or not it is appropriate to assess children's single word handwritten responses. In such an assignment a high scoring confidence level has to be maintained even though there is little or no contextual knowledge to help improve the underlying recognition rate. In addition, in this work the legibility of the handwriting is also in question, as the children tested displayed quite a wide range of handwriting skill. A novel approach is therefore reported which does not rely upon traditional lexical recognition but instead utilises the focus of the question and answer medium to simply check if the correct word is present or not. This novel approach is discussed in more detail in section 2.2.

Our previous work has already shown that highly accurate assessment of cursively written responses is possible if the constrained nature of the responses is taken into account [1]. 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 response no improvement can be gained from such contextual knowledge.

For instance, figure 1 shows an example of a single word response style exercise used to assess children aged between 5 & 6. This exercise forms part of the "Progress in English 6" written exam, produced by the National Foundation for Educational Research & Nelson publishing company (NFER-Nelson). Many of the questions and exercises are multiple choice requiring only a single word response from a given list. In this exercise the list can be seen at the bottom of the exam page.

Exercise 2 – Part A

Choose a word from the boxes to complete each sentence.
The first one has been done for you.

The shops are shut.

Jo has one sister and one brother.

Our house is next to a road.

They were happy to meet their friends.

His birthday is in September.

The car stopped at the red light.

happy next stopped
shut birthday brother

Figure 1: Question page from Progress in English 6 published by NFER-Nelson

In order to carry out any automatic assessment of such hand written responses, the method must be able to confidently assess each response or reject it for manual marking. The ability to reliably detect and throw out errors whilst maintaining a high overall response yield is known as the Scoring Confidence [1].

The actual student responses shown in figure 1 also show that the structure of children's handwriting is notably different to that of adult handwriting. Although there are a multitude of individual styles of adult handwriting, the basic word structure is present more often than not, i.e. ascenders ascend, descenders descend and the shapes of the words are formed. In a child's handwriting these characteristics may or may not be present. The more adept the child is at writing the more structured or 'grown up' the handwriting becomes. Contrarily the more the

child struggles, the more illegible the handwriting. For some children, the acquisition of written expression skills is a difficult and enduring problem [2].

brother
 next
 happy
 stopped
 birthday

bndt
 brrrrrr
 stoped
 rrrr
 me = t

Figure 2a: Example of child 4's responses to Q1-5.

Figure 2b: Example of child 2's responses to Q1-5.

At the level of education tested in this data set there is large diversity in the children's handwriting skill. This is shown to great effect in the two samples shown in figures 2a & 2b. It can be seen that child 4 understands the exercise and is able to copy and write a legible response from the list of alternatives at the bottom of the page. Characters are formed correctly and spaced neatly and evenly. Child 2 on the other hand shows only a limited comprehension of the exercise and fails to translate his/her responses in a legible manner. It can also be seen that child 4's writing, though legible, is of a level where certain characters are miss-positioned in relation to the line and its neighbours. This characteristic is common in children's handwriting and can be attributed to the way the children have been taught to write [3]. Children are first taught how to form single characters and then taught how to combine them so that they are in proportion and in line to form a word. This developmental approach to handwriting can be seen in the formation of the characters 'r' & 'p' in figure 2a, where the height of the letter 'r' is equal to that of the neighbouring 'b' in the word 'brother' and where the letters 'p' in 'happy' and 'stopped' are placed above the line.

Both of these style characteristics would cause problems for our original HVBC recogniser [4] as it relies heavily upon the presence of features such as ascenders and descenders. In cases such as those in figures 2a & 2b, the zoning process may miss-zone the words 'brother', 'happy' & 'stopped'. This would then prevent the system from detecting possible ascenders or descenders in the image pattern leading to eventual miss-recognition. Consequently, for this work, it was found necessary to modify the feature weights in the HVBC recogniser such that it relies less upon the presence of ascender & descender features and more upon other features such as mid-zone bars, holes and cups.

2. Experimentation

2.1 Traditional Lexical Based Approach

In this paper two approaches for the automatic assessment of the response format seen in figure 1 are investigated. In the first a traditional recognition method is used, where a single input pattern is compared to a list of alternative response templates held within a lexicon. This is shown diagrammatically in figure 3.

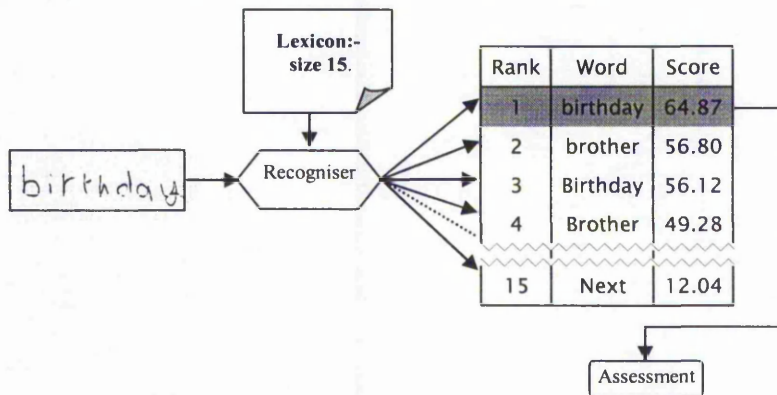


Figure 3: The traditional recognition system

In this type of assessment the 5 possible responses (see figure 1) form part of a 15 word template lexicon along with their case alternatives, i.e. lower case – birthday, next, brother, happy, stopped; mixed case – Birthday, Next, Brother, Happy, Stopped; upper case – BIRTHDAY, NEXT, BROTHER, HAPPY, STOPPED. After recognition, a single list of alternatives is produced for each response with the alternative words being ranked in order of how close the input pattern features match the template word.

In the traditional approach the top word from the list of alternatives for each question is then passed on for automatic assessment. Two assessment techniques have been adopted for the traditional approach so that a comparison can be made between it and the novel approach. The first uses the raw results from the recognition process where the top words for each response are passed for scoring straight away. The resultant assessment accuracy will be then the raw recognition accuracy of the system. The second method uses a threshold to filter the correctly recognised and incorrectly recognised responses in an effort to minimise the errors produced by the recognition process.

Scoring is the process of marking and classifying the response as either CORRECT or INCORRECT. Table 1 shows the different possible outcomes from the recognition process.

Input	Automated		Classification
	Recognition	Scoring	
Correct	Correct	CORRECT	GOOD
	IncorrectA	INCORRECT	BAD
	IncorrectB	INCORRECT	BAD
	IncorrectC	INCORRECT	BAD
	IncorrectD	INCORRECT	BAD
IncorrectA	Correct	CORRECT	BAD
	IncorrectA	INCORRECT	GOOD
	IncorrectB	INCORRECT	BAD
	IncorrectC	INCORRECT	BAD
	IncorrectD	INCORRECT	BAD

Table 1: The possible classification outcomes of the system

It can be seen that if a correct handwritten response is inputted in this system, then there is the possibility that the recogniser may either recognise it correctly or may miss-recognise it as one of the four incorrect answers. If the recogniser recognises a correct response as correct then the response will be scored as CORRECT with a GOOD classification. However if the recogniser mistakes the response for one of the wrong answers then the system will score the response as being INCORRECT. This is manually classified as BAD because it is an unaccountable error within the automatic system that could not be detected without human intervention. In a similar manner, an incorrect handwritten response may be correctly recognised and scored as INCORRECT (classified GOOD) or incorrectly recognised as the correct response or another incorrect response (classified BAD) and incur the same unaccountable scoring error.

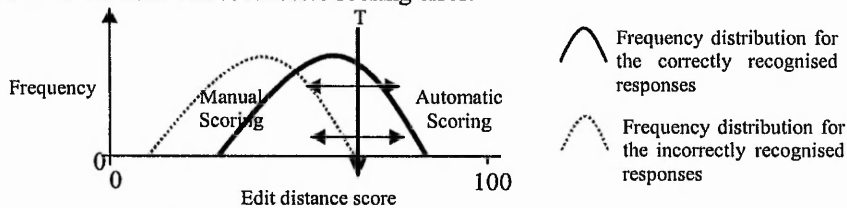


Figure 4: A stylised example of a distribution graph showing the frequency distribution of the correctly and incorrectly recognised response scores

Applying a threshold to the associated edit distance scores [4] of the top ranked words is one way in which this error can be overcome automatically. This threshold can be calculated from a training set by plotting the frequency distribution graphs of the correctly and incorrectly recognised responses. The threshold (T) is then taken

from the point at which there are no further incorrectly recognised responses. An example of this can be seen in figure 4.

Top ranked responses with a score higher than T will be automatically scored and responses with score lower than T will be automatically passed for manual scoring. Only a high score threshold may be implemented as results gained from a low score threshold would only identify the definitely incorrectly recognised responses but will not indicate whether the actual response is CORRECT or INCORRECT. Since any word which scored below the high threshold would be passed for manual scoring the low score threshold would be redundant.

2.2 Novel Word Verification Based Approach

The second approach is more novel, in that it exploits the nature of the question and answer medium by only comparing the input pattern to the template of the correct answer for that specific question, see figure 5.

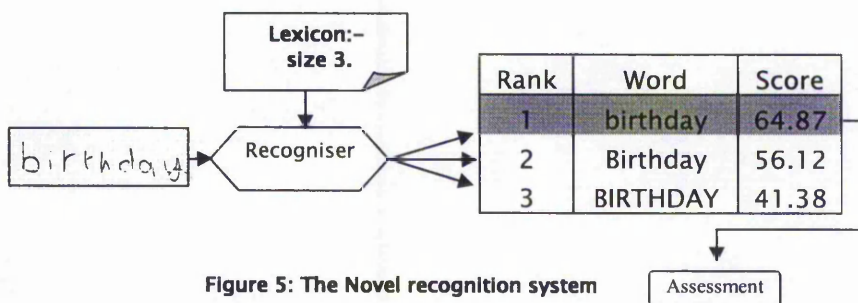


Figure 5: The Novel recognition system

For example, in the question "His _____ is in September" the correct answer would be 'birthday' and therefore the lexicon used for this question would only consist of the feature templates for the words: birthday; Birthday; & BIRTHDAY. Of course by neglecting the other four target responses this approach will always generate the correct answer as the recognised response. However, the recognised responses can then be thresholded on the basis of the top ranked word score. A high score implies that the recogniser has achieved a close match between the input pattern and one of the correct word target templates. The system can thus confidently score the response as CORRECT. However, if the word has a low score, this means that the recognised word is either a wrong answer or is illegible and it can therefore be automatically scored as INCORRECT.

Figure 6 illustrates this with three example responses. It can be seen that the written response birthday achieves a high score that peaks above the upper threshold t_2 , thus scoring it automatically as correct. The response "Stopped" shares little features with the feature templates of 'birthday', 'Birthday' & 'BIRTHDAY' and therefore has a low score allowing it to be automatically scored as incorrect. The response 'brother' on the other hand is a closer match to the 'birthday' template and attains a score that is higher than t_1 but lower than t_2 . This therefore should not be assessed by the system but must be set aside automatically for manual assessment.

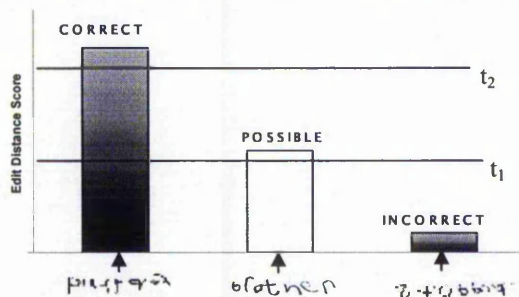


Figure 6: three examples of written response that have been assessed using the thresholds for the question "His ____ is in September"

3. Results

A total of 29 children completed the 5 questions (shown in figure 1) as a part of the Progress in English 6 exam. This gave a total of 145 handwritten responses. A professional human assessment¹ scored 65% of the responses as correct and 35% as incorrect. This was achieved with 100% accuracy.

3.1 Traditional Lexical Based Approach

In figure 7, an example of the actual written response and the recognition results from the traditional approach for Child 8 can be seen.

Rank	BROTHER	NEXT	HAPPY	BIRTHDAY	STOPPED
1	next 67.3	next 74.0	Next 66.5	Brother 54.6	brother 60.2
2	HAPPY 64.6	HAPPY 65.8	STOPPED 53.6	Birthday 53.7	next 58.5
3	happy 64.5	NEXT 64.0	HAPPY 53.3	Stopped 49.1	STOPPED 57.5

¹ The exam papers were pre-scored by an external examiner employed by NFER-Nelson before being used in this work.

Figure 7: Example of the traditional recognition results for child 8

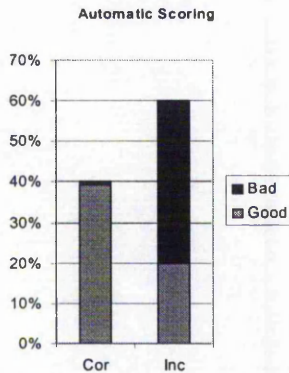


Figure 8: Graphs to show the Automatic Scoring in the Traditional Approach.

This system achieved an overall recognition rate of 59% (shown in figure 8), which would obviously imply an error rate of 41% if the top ranked words were simply passed for scoring. The actual correct and incorrect responses as determined in the manual scoring is also shown in figure 8, where it can be compared to the automatic scoring. Although the automatic process achieved 99% accuracy for scoring the correct responses it only achieved 33% accuracy for the incorrect responses. The main reason for this is the inability of the recogniser to cope with poorly written responses such as those seen in figure 2b.

Applying the threshold to the traditional approach firstly involved creating training and test sets. 75% of the data was randomly removed and used for the training set and 25% was set aside as the unseen test set. The frequency distribution graphs of the correctly and incorrectly recognised response word scores were produced from the training set and a threshold (T) of 81 was determined (see figure 9). This threshold was then applied to the test set. From figure 10, it can be seen that only 5% of the responses could be passed for automatic assessment. However a scoring accuracy of 100% was achieved with this 5% response yield.

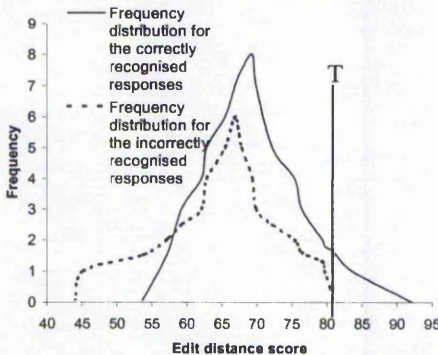


Figure 9: Graph to show the smoothed frequency distribution of the correctly and incorrectly recognised response scores and the value of threshold T.

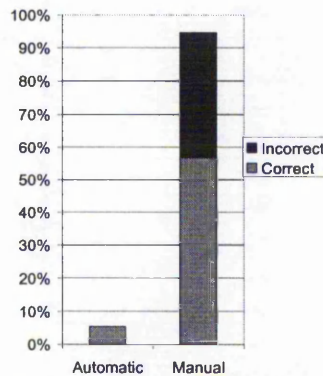


Figure 10: Graph to show the response yields after the threshold has been applied.

3.2 Novel Word Verification Based Approach

The results from the novel approach for Child 8 are shown in figure 11. It can be seen that in each list only the correct answer is present. Again only the top ranked responses are passed for assessment.

Rank	BROTHER	NEXT	HAPPY	BIRTHDAY	STOPPED
1	brother 52.7	next 74.0	HAPPY 53.3	Birthday 53.7	STOPPED 57.5
2	BROTHER 51.4	NEXT 64.0	happy 39.7	BIRTHDAY 47.9	stopped 55.0
3	Brother 47.9	Next 47.7	Happy 28.2	birthday 45.6	Stopped 54.2

Figure 11: Example of the novel recognition results for child 8

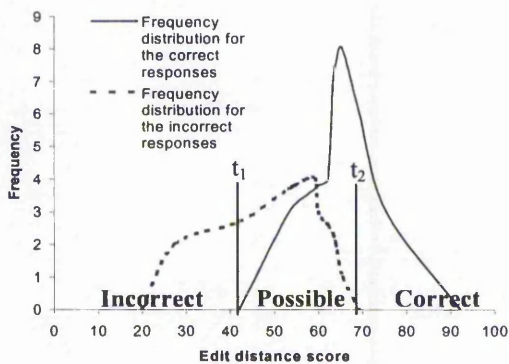


Figure 12: Graph to show the smoothed frequency distribution of the correct and incorrect response scores and the values of thresholds t_1 and t_2

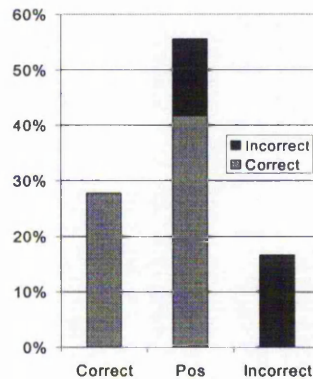


Figure 13: Graph to show the response yields after the thresholds have been applied.

Training and test sets for the novel approach were created in the same proportions as in the traditional approach and the frequency distribution graphs of the correct and incorrect responses were plotted (see figure 12). From the graphs t_1 and t_2 were determined as 41 and 69 respectively and then used to analyse the test set. Figure 13 shows that the novel approach scored 28% of responses as correct and 16% as incorrect without error. This gives an overall response yield of 44% with an accuracy of 100%.

4. Conclusion and Future Work

This paper has introduced a novel approach for the automated assessment of handwritten single word responses. The novel approach performed very well when compared to a traditional lexical based approach. Yields of 44% were gained with 100% Scoring confidence using static thresholds on real world data. This compares favourably to the traditional approach where the low recognition rate of 59% is compounded by the fact that a constrained lexicon of only 15 words was used. The development of confidence methods is one avenue of further work that could be investigated. These methods could be trained to classify the outputs from the recogniser into legible and illegible classes in order to increase response yields whilst not compromising accuracy [5,6]. For instance a neural network could be used to dynamically determine writer specific thresholds from the individual exam papers. Other avenues of investigation that have yet to be explored involve using style dependent classifiers [7] to detect a recogniser specific illegible writer, thereby allowing the system to reject the responses prior to recognition.

Acknowledgement

I would like to thank Mike Peppiatt of NFER-Nelson who spent a lot of time and effort to supply the data for these experiments.

References

- [1] 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.
- [2] Luttinger, L. Gertner, M. **Learning Disorder: Written Expression.** eMedicine Journal, April 30 2001, Volume 2, Number 4.
- [3] Payne, G. Isaacs, L. **Human Motor Development; A life span approach,** 4th Ed. Chap. 12 (1999) Mountain View California: Mayfield Publishing Company.
- [4] Evans, R G. Sherkat, N. Whitrow, R J. **Holistic Recognition of Static handwriting Using Structural Features.** Document Image Processing and Multimedia (DIPM'99), IEE Colloquium 99/041, Pages 121-124 25th March 1999, London.
- [5] Gori, M. Scarselli, F. **Are Multilayer Perceptrons Adequate for Pattern recognition and Verification?** Published in IEEE Transactions on Pattern Analysis and machine Intelligence, Vol 20, No 11, November 1998.
- [6] Vermeulen, P. Barnard, E. Youghong, Y. Fanty, M. Cole, R. **A Comparison Of HMM and Neural Network Approaches To Real World Telephone Speech Applications.** IEEE international Conference on Neural Networks and Signal Processing, 1995. Pp 796 – 799.
- [7] Ebadian Dehkordi, M. Sherkat, N. Allen, T. **Case Classification of Off-line Hand-written Words Prior To Recognition.** IAPR (2000).

Automated Assessment of Children's Handwritten Sentence Responses

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

Abstract

This paper compares two approaches of recognising handwritten sentence style answers, which are then assessed. The first, is a conventional lexicon approach to handwriting recognition. An approach that generalises the problem, thus is able to recognise all possible responses and as a result it will give a recognised response every time. The second method employs the specific word assessment technique to evaluate each word in the written response against only a model answer. If the technique is not confident about a word that it has recognised then it will not give a response. In both approaches there is the option of rejecting a response and thus passing it for manual assessment. The use of a questions history is also exploited to give a more robust assessment and is used to assess the recognised responses from both of the two approaches and is compared to a baseline results were the history has not been used. Results show that the Specific Word Assessment Technique with History performs best with an overall assessment accuracy of 100%, however the high accuracy has been achieved at the expense of the total amount of responses assessed, 33.2% were 66.8% of all responses the approach was unconfident in marking thus requiring human intervention. The Conventional Lexical Approach with History managed an accuracy of only 46.3%, with a yield of 66.9% of all responses being automatically assessed.

Keywords: Automated Assessment, Handwriting Recognition, Confidence.

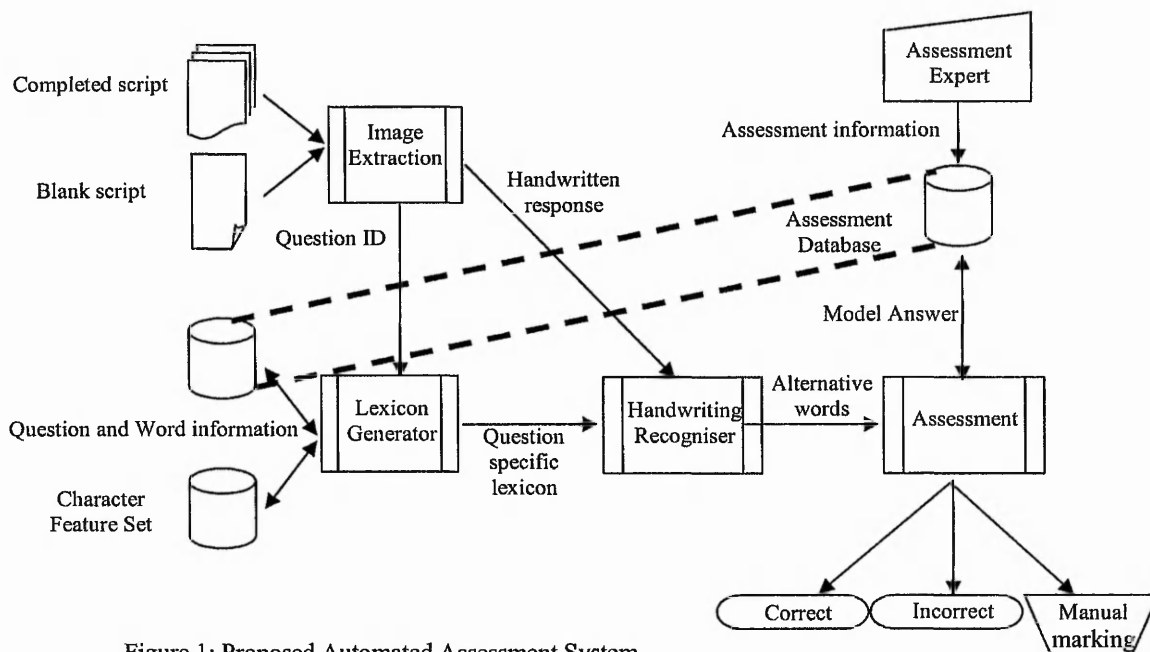


Figure 1: Proposed Automated Assessment System

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.

Image Extraction is considered a mature research field and is out of the scope of this paper. 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 them independently, the holistic approach is to recognise 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 were 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.

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



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. The use of a questions history is also investigated to show that the model answers alone are insufficient and that additional 'real world' knowledge is necessary to mark all the answers accurately.

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 in 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 best match in each case is then used to build a recognised response that is used in the assessment stage, "once vibration eyes The where on five made".

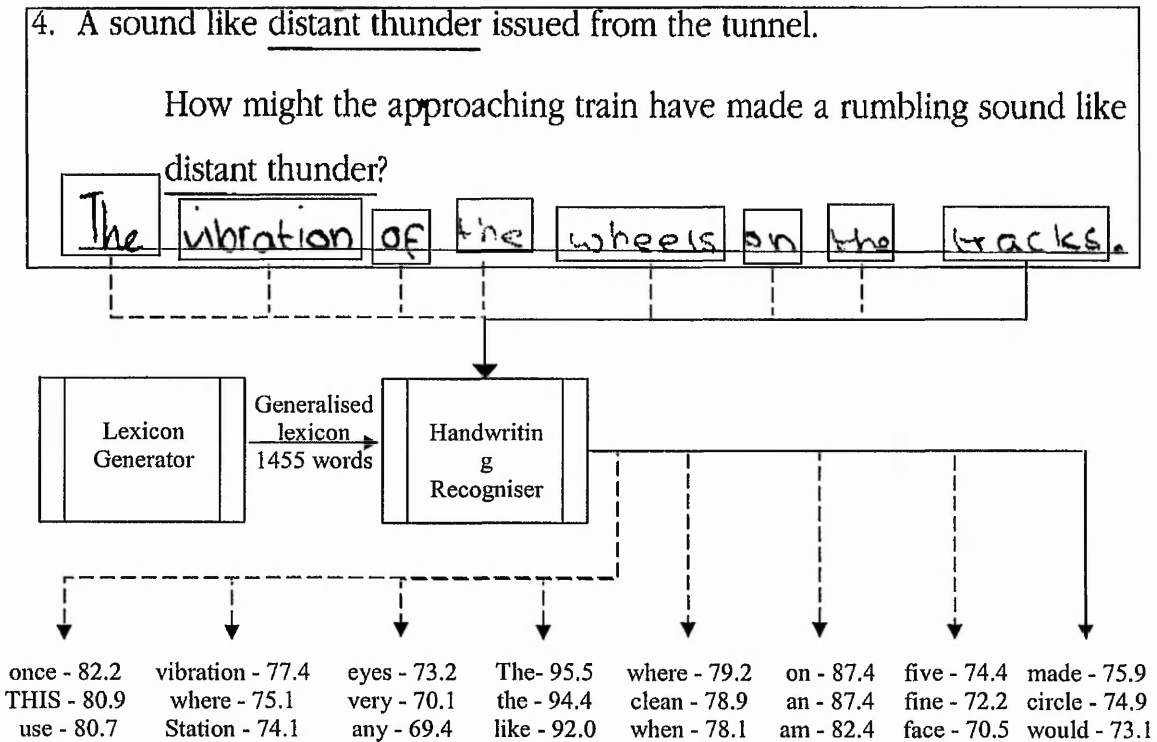


Figure 3: An example of the Conventional Lexical Approach being employed to recognising a handwritten sentence

Specific Word Assessment Technique

SWAT exploits the nature of the question and answer medium by only comparing the input pattern to the template of the correct answer 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 top ranked recognised words a keyword (KEY), possible (POS) or not a keyword (NKY) (see figure 4).

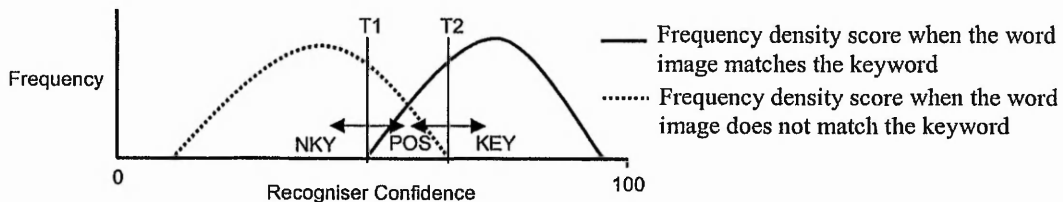


Figure 4: A stylised frequency density graph to obtained the two confidence thresholds for SWAT

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. 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 t_2) implies that the recogniser has achieved a close match between the word image and 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 t_1), this means that the recognised word either is a wrong 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 5 shows an example of how SWAT can be applied to build a recognised responses for the written response: "The vibration of the wheels on the tracks" using only the model answer as the lexicon.

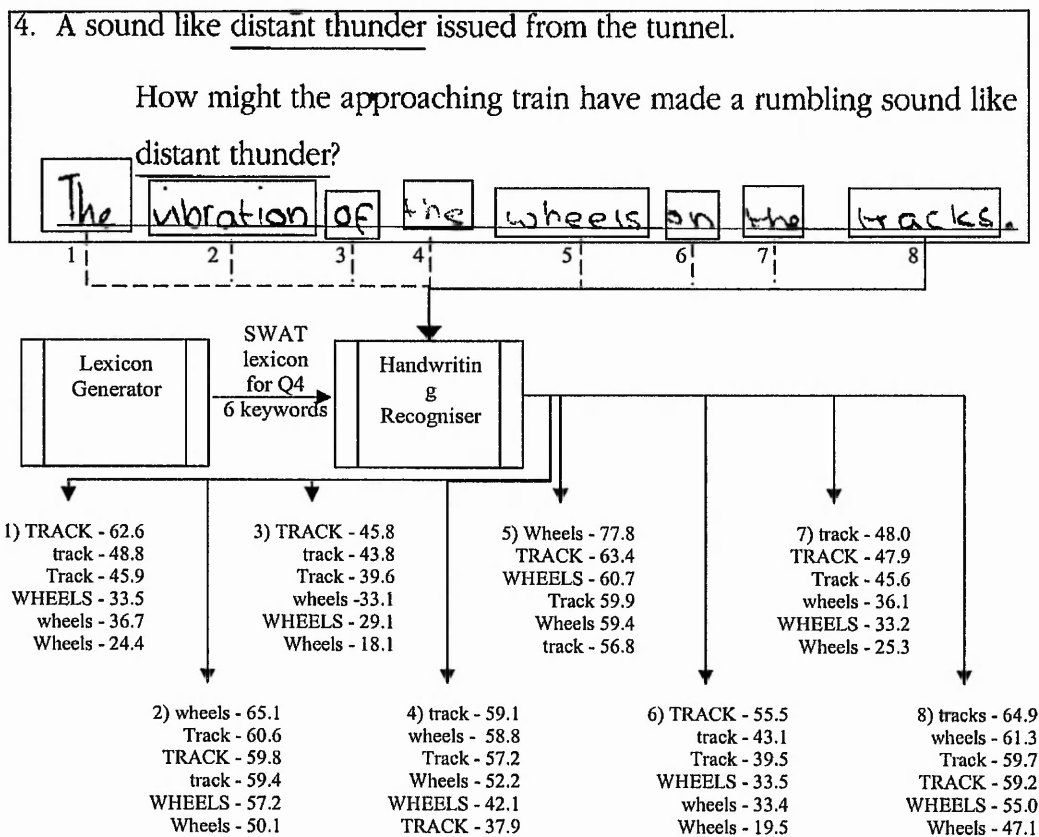


Figure 5: An example of the Specific Word Assessment Technique being employed to recognising a handwritten sentence

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 would be 'POS POS NKY NKY **KEY** NKY NKY POS' using the thresholds. 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'.

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 model answers are given by NFER-Nelson. 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 information has been found relating the recognised response to the model answer then the response can be marked as incorrect. To make an initial comparison in respect to the inclusion of knowledge into the experiments a History set is created from previous correct, and frequent incorrect answers. This set is used to augment the model answers to form new assessment criteria, as shown in Appendix Aii. Using this criteria the written example in figures 4 & 5 would be automatically marked as incorrect by CLA and passed for manual assessment by SWAT as a partial answer was found.

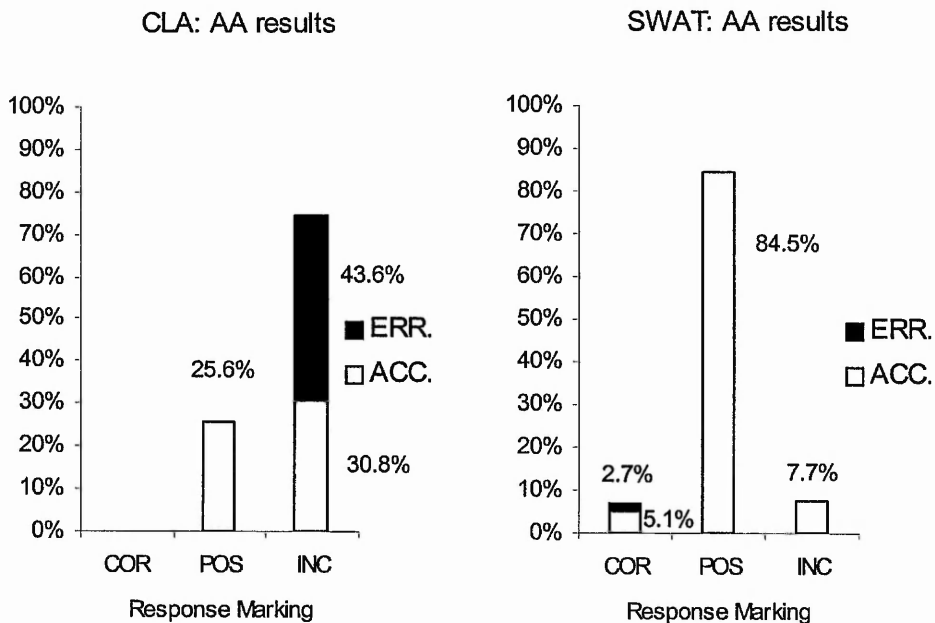
Results

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% with all words being recognised (both use same lexicon and only the assessment criteria changes). This compares to SWAT with History's (SWATH's) keyword recognition rate of 97.1%. However, SWATH was confident on only 63.9% of the words, leaving the remaining 36% unusable in the assessment of the responses. 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 all 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 which are actually correct answers that have been miss-recognised and erroneously assessed. This is error rate of the response automatically assessed using CLA and the poor performance was contributed to the low word recognition rate. In contrast SWAT automatically assessed 15.5% of all response rejecting 84.5% for manual assessment, as it was could not confidently mark them. 2.7% of the responses where incorrectly marked as correct. This was a result of two children giving an incorrect response 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 high assessment accuracy, however the number of responses was low as a result of the assessment criteria being too limited. This shows that human assessors must 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

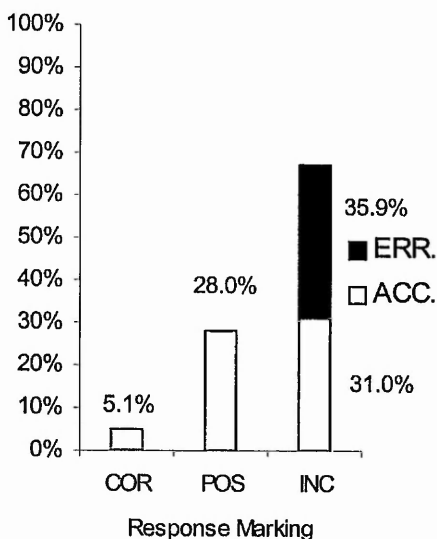
As in the previous experiments the graphs for the automatic assessment of the responses using the two approaches can be shown, see figures 8 & 9. With the inclusion of new assessment criteria the accuracy of both methods has increased. However the of number responses automatically assessed by CLAH has decreased while it has increased using SWATH. This is a result of CLAH recognising more partial model answers as there are more keywords and also the assessment criteria is better improved therefore SWATH can confidently assess more responses. When the history data is added to the assessment criteria, the accuracy of the system increases to 100%. This was due to the approach being able to assess the responses 'steam engine' in question 2 because without history this response is marked as correct. The addition of the history made up for the lack of 'common sense' therefore making it possible to automatically score more responses.

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

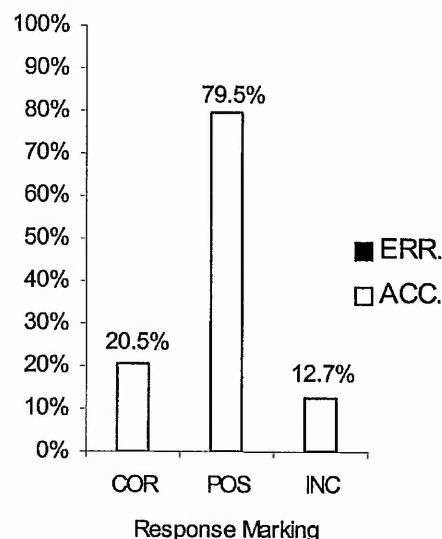
	Responses Automatically Assessed (%)	Assessment Accuracy (%)
CLA	74.4	41.4
CLAH	72.0	46.3
SWAT	15.5	82.8
SWATH	33.2	100

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

CLAH: AA results



SWATH: AA results



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

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 as 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 however has a very high keyword recognition 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%). However in this case the increase in accuracy does not decrease the response yield but instead with the added criteria the approach can identify more responses and therefore automatically assess them. SWATH assessed 17.7% more responses than SWAT, automatically marking 33% of the responses with an accuracy of 100%. A large scale trial of SWATH is required to determine if these results can be sustained and therefore be a viable solution to ease the burden of marking handwritten sentence 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 TRAIN 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