Automated assessment of non-objective textual submissions.

CHRISTIE, J.R.

2003

Automated Assessment

of

Non-Objective Textual Submissions

by

James Rennie Christie

A thesis submitted in partial fulfilment of the requirements of The Robert Gordon University, Aberdeen for the degree of Doctor of Philosophy

February 2003

James Rennie Christie

Doctor of Philosophy

Automated Assessment of Non-Objective Textual Submissions

Abstract

The author sought to explore the methodlogy of automated marking of non-objective textual submissions. In so doing he tasked himself to discover how far automated assessment of text could be applied in the range of submissions from the single word to a multi-sentence, multi-paragraph piece of text.

In the course of his research the author had to determine the extent and severity of any problems associated with manual marking of such submissions. In fact, the thesis informs the reader of the myriad problems associated with, and generated by, the subjectivity inherent in the manual marking process and the author indicates to what extent automated marking can remove or reduce these problems.

In parallel the author had to determine the existence, if any, of automated assessment of such type of submissions. The literature survey did show that some work had been done in this subject area, but that most of the work dates from the 1960's for style marking and only recently has some work on content marking been published.

In the event the author devised algorithms and implemented them in a software system, called SEAR (**S**chema **E**xtract **A**ssess and **R**eport), that would process word-processed submissions and award marks. Lack of suitable marked data sets prevented the full development of the style algorithm. However the author was able to demonstrate that content marking is possible, within a range of submissions.

The proposed style algorithm is based on a novel idea of determining a set of metrics that could be applied to all textual submissions. The data structure that was developed for the marking of content is unique.

In passing the author compiled a set of potential criteria for use in the evaluation of the methodology of automated marking of textual submissions. These criteria were applied to his software system.

Acknowledgements

The author wishes to acknowledge the following individuals for their support of his research project.

Supervisory Team
For the unstinting support and encouragement throughout the time of the research project:

Director of Studies: Emeritus Professor H I Ellington

Second Supervisors: Ms S Earl and Dr R Butler

Statistics
For help and advice on statistical matters: Mr A Wilson

Library
In particular Inter-Library Loans and Subject Librarians

Second Markers
Ms C Norton, Lecturer at Aberdeen College
Dr M Sharif, Lecturer at Napier University

Family
For unswerving support, encouragement and patience throughout the long hours.

Table of Contents

List of Tables

List of Figure(s)

Chapter 1: Introduction

1.1 Definition of an Essay

This chapter acts as a starting point to this research project by laying down some basic definitions of essay, markers and so on in preparation for the later chapters.

In the thesis resultant from this research project the term "essay" needs to be defined so that the subject of the algorithms for marking style and content is clearly understood.

The author accepts that every reader will have his or her own concept of what constitutes an essay. In the course of the author's research the author has come to regard the complex definition, attributed to Stalnaker, as the 'best'. The author means that in his personal view Stalnaker's definition is 'best' in terms of outlining the expected composition of an essay and 'best' by the holistic manner in which the many aspects of an essay are encompassed succinctly, while simultaneously introducing the problems of marking an essay. These problematic aspects are the ranges of essay size possible together with the flexibility of content and the flexibility of style. All these aspects are found in Stalnaker's definition that is partially quoted below:

> *"(essay question) is defined as a test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject.*
>
> *The most significant features of the essay question are the freedom of response allowed the examinee and the fact that not only can no single answer be listed as correct and complete, and given to clerks to check, ...*
>
> *... but even an expert cannot usually classify a response as categorically right or wrong.*
>
> *Rather, there are different degrees of quality or merit which can be recognised."*
>
> Attributed to Stalnaker, JM by Coffman WE, in Thorndike, Educational Measurement, p. 271, 1971

Before the reader carries on with this thesis, the author wishes to point out that many references may be considered rather long in the tooth. This research area is not popular and thus, particularly for automated essay assessment, there is not an avalanche of modern references to cite.

1.2 Types of essay

Essays are used in teaching and learning and in assessment where they are widely accepted as being a traditional method of assessment. Essays are quicker to set and easier to set than possibly most of the other forms of assessment, although not all subjects at all times and levels must be, or even should be, assessed by essays. The author accepts that the setting of better quality, or better performing, essay based assessments is a craft that requires some skills of the essay setter. The problem with using essay-based assessments is the amount of time often of significant duration, which it takes to mark the essay, together with the effort required to provide feedback to the essayist. As the current trend is towards an increasing number of students in further and higher education this marking problem is set to increase.

For teaching and learning the use of essays provides staff with a mechanism in which to base evaluative / formative / diagnostic and summative feedback, and formative feed-forward, on their teaching practice which consequently affirms other operational information.

In this research project the author only considers the use of essays in assessment, where the nature of assessment may be diagnostic, formative or summative. The purpose of these three forms of assessment is generally the same, but the impact is vastly different.

In diagnostic assessment the essayist is examined at the start of a programme of study to find what course or courses of action should be taken by the essayist, or to establish the suitability of options for the essayist.

Formative / evaluative assessment is used for shaping or modifying the essayist's performance during the course of study. In this mode of assessment the aim is to find out what the essayist has done right and indeed done wrong, and use this information as a basis for future activity of the essayist. This "finding out" may be given to the essayist by the examiner, and / or learned directly by the essayist. There is a trend nowadays to use peer assessment, where a tutorial group of essayists mark each other's work against a marking scheme. Self-assessment can also provide formative assessment for the essayist.

However the stakes change dramatically when summative assessment is considered. On the basis of the result of summative assessment the essayist may progress to the next stage in a programme of study, or graduate, or obtain employment, or otherwise advance.

In this research project the author deliberately makes no distinction between the three modes of assessment, as the methodology of marking does not alter in the light of the mode of assessment applied. Only the impact of the assessment on the essayist changes with the mode of assessment.

The author acknowledges that other practitioners may have different ideas or concepts on type(s) of essay and on the variant(s) thereof that could be attributed to different types of essay. An exploration of the typical range of types, and the purpose of each of these types is given by Cockburn and Ross (1978).

According to Cockburn there are two main types of essay: coursework and examination. Following the ideas of Cockburn a "coursework essay" is an open-ended, open-book assessment that is expected to be properly structured with proper references, whilst accompanied by some parameters for completion and other points of guidance. Such essays may be used in group seminars or in individual seminars. Such essays may be re-submitted by the essayist after receiving feedback information. On the other hand an "examination essay" is usually unseen before the examination and is developed by the essayist under controlled conditions such as time and place, using approved materials, where structure and referencing may not be expected by the examiner to be perfect. There is, generally, no chance for the essayist to re-submit and no feedback other than the mark the essayist is awarded for his or her efforts.

For each type of essay that Cockburn and Ross identify there are several variants, as detailed below:

Essay Types	**Variants**	**Identifying Characteristic**
Coursework	Short essays / Essay outlines / Seminar notes	~ 500 words
	Standard essays / Seminar paper	~ 1,000 – 2,000 words
	Extended essays / Extended time essays	~ 2,500 – 5,000 words
	Dissertation / Thesis	~ 5,000 – 15,000 words
Examination	Short Answers	~ 10 – 20 answers in 3 hours
	Traditional	~ 3 – 4 answers in 3 hours
	Special	~ One essay in 3 hours

Table 1.2 Types of Essay

The Essay Definition given in section 1.1 above is equally applicable to each of the above listed essay type and variant combinations.

Coursework: Short essays or essay outlines or seminar notes

Example: "What are the eight principles of the 1984 Data Protection Act?"

This type of essay is used to determine how well the essayist has mastery of facts, understanding of principles and concepts, but it tests the ability of the essayist to selectively organise materials into cohesive text. This is often used in formative assessments and used in tutorial group work. When used in the initial formative assessments, the essay outlines are used to ensure the essayist has created an essay with acceptable structure and with acceptable content. This is done to prevent the essayist having to make a large investment in time and effort in creating a full essay only to find that it is not acceptable.

Coursework: Standard essays or seminar paper

Example: "Apply the eight principles of the 1984 Data Protection Act to the case-study."

This is used to elicit how well the essayist weaves disparate formats of information into a balanced argument using analysis and evaluation techniques.

Coursework: Extended essays or extended time essays

These are similar to the previous type, but the essayist will be working on a more complex problematic area. The essayist will be expected to conduct further reading or further background work in the process of preparing the essay.

Coursework: Dissertation or Thesis

This is a major piece of work in which the essayist has to show a properly structured, complete, holistic in-depth study in a particular subject area.

Examination: Short Answers

Example: "Write short notes on four of the following topics".

This is much used and possibly much loved in examinations to ensure that the coverage of the syllabus is complete. "Short answers" type essays are favoured by those essayists who find essay writing a difficulty in an examination situation. Often the expected answers are in the format of bullet points or the ubiquitous "see page X paragraph Y of the textbook".

Examination: Traditional

Example: "Answer any four questions in 2 hours."

This is the usual format for university level examinations. Generally the rubric is any 4 from 6 or any 5 from 8. Examination times are usually set to last 2 hours, or 2.5 or 3 hours. In practice there will be a mixture of monolith questions (worth an award of 20 or 25 marks for the single answer), and multi-part questions (for example " ... (10, 10)" or " ... (10, 5, 5)"). Multi-part questions often include the "Short Answer" essay question type mentioned above.

Examination: Special

Example: "Answer only 1 question in 3 hours."

The author has not set nor has encountered such an examination. However in fields such as law, health education and business there is extensive use of this type of essay assessment based on a case study that has been issued well in advance of the actual assessment. Essayists undergoing this type of assessment are expected to do background reading, research and so on and thus come prepared to the assessment. Such an examination would constitute a very high stakes assessment that could only favour those among the population who can easily produce essays on a particular topic.

In the late 1990's Catterall (1998) produced a detailed study guide as to what a university essay is. This four-page guide seeks to explain to the newest group of university-level essayists what their essay markers are seeking in essays in general. Catterall does not appear to fully explain the myriad of essay types possible. Caterall's study guide details what to and how to, include or use the various factors to manipulate essay content and style in order to maximise marks. This guide finishes by encouraging students, as essayists, to seek further help if they, the students, feel that their essay skills are weak. There are several such guides available to prospective essayists, for example Donley (1978).

The author is acutely aware that the term "essay" covers a large range of potential submissions. In the course of this author's research the author has deliberately chosen not to work on a specific essay type or variant. This allows the author's research to be unfettered by any artificial or arbitrary limitations caused by the selection or choice of a particular essay type or variant combination(s) to work with. The size of text the author used to develop this research ranged from the trivial "The cat sat on the mat." to a text of some 800 words on the history of The Robert Gordon University, Aberdeen. In Chapter 4 this author describes both the developmental and test essay sets in detail.

1.3 Definition of an Essay Set

The author defines an "essay set" as being a collection of essays, produced by individual essayists such as the students in a specific cohort, in response to a particular assessment instrument. This instrument of assessment has been set by an examiner who is normally a member of staff, usually designated the first marker, at the educational organisation.

1.4.1 Definition of First Marker

This is a member of academic staff who is an expert in a particular subject area. The first marker is usually the person who is normally responsible for the teaching of the students who are in this case the essayists, and for the setting of the instrument of assessment. In the context of this thesis it would be the essay topic and all the associated parameters of size, submission date, format and so on which would be part of the instrument of assessment. The setting of the instrument of assessment should include the marking schema for style and more especially for content. Finally the first marker would be responsible for the marking of the submissions which in the context of this thesis would be essays.

1.4.2 Definition of Second Marker

The second marker is another member of academic staff who is deemed expert in the same subject area and who is furthermore an expert academic practitioner in the marking of submissions. The second marker's duties are to audit and where applicable validate, according to the protocols of the educational organisation, the work and the marks awarded by the first marker.

1.5 Assessment Life Cycle

British Standards Institute Standard 7988

It is a certainty that every reader of this thesis will have his or her own ideas about what constitutes an assessment life cycle. As a compromise benchmark the author cites the one contained in the draft standard BSI 7988 "*A code of practice for the use of information technology in the delivery of assessments*". This draft was issued by the British Standards Institute on the 10th October 2001 (BSI Standard 7988:2001) as being a typical assessment life cycle. This particular assessment life cycle is adopted for the BSI Standard as it covers or emphasises, implicitly, the mapping of computer assisted assessment in general to various stages of this particular assessment life cycle.

As an aside, the reader should be made aware that the scope of the BSI 7988 Standard concerns the delivery of assessments to examinees and the recording and scoring of their responses, where the delivery, the recording and scoring is mediated by any computer mediated system.

However the reader should note that this standard specifically includes a disclaimer to the application of the standard to text in general and to essays in particular. It should be further noted that this standard does not apply to the content of the assessment instruments.

The author's software that underpins this thesis is called "SEAR", where SEAR is an acronym representing the four stages through which an essay passes in the marking process: Schema, Extract, Assess and finally Report; these same four stages constitute the model of manual essay marking that the author currently uses. The column headed "Covered by PhD software SEAR" maps the author's research onto this assessment life cycle and incidentally also maps onto BSI 7988. In a second disclaimer the BSI 7988 standard specifically does not cover human marking in the production of marks.

The assessment life cycle typified by the draft BSI Standard 7988 is outlined in the table below:

ID	Assessment Life Cycle Stage	Covered by BS 7988 Standard	Covered by the author's PhD software SEAR
A	Identification of need to assess		
B	Design of outcomes / assessment methodology		
C	Preparation and Calibration		Y ~ Schema
D	Pre-registration		
E	Distribution *	Y	
F	Authentication *	Y	
G	Delivery *	Y	Y
H	Response return *	Y	Y
I	Scoring, result determination +/- Feedback *	Y	Y ~ Extract & Assess
J	Data return *	Y	Y ~ Report
K	Analysis		
L	Appeals		
M	Certification		

Table 1.5: Assessment Life Cycle

Each of the six items specific to automated assessment and shown as asterisked items in Table 1.5 above and are each explored in the following paragraphs.

Distribution

This item is concerned with the distribution of software used for assessment. Here software includes both the assessment engine and the actual instruments of assessment. The assessment engine and assessment instrument software are expected to comply with guidelines on inter-operability as being developed by the IMS organisation, in order to maximise flexibility in the distribution of assessment software. The IMS standard seeks to maximise the integration of assessment software with other software. At the end of this section there is a brief commentary on the work of the IMS.

Authentication (including Identification)

This is concerned with proving that the examinee is a bona fide candidate and is also concerned with the prevention of academic misconduct such as impersonation. Generally, in examination halls or in examination rooms, invigilators can check the examinee's face against an identification card and these cards can be checked against examinee lists. Unfortunately with coursework there is a rising tide of plagiarism and the question of candidate authentication is extremely topical. The problem of authentication is magnified when remote or off-site assessment is being conducted.

Delivery

For the actual assessment the examinee, the required equipment such as computer, furniture and screens etc and the specialist software comprising the assessment engine and assessment questions have to rendezvous at the same time and place. The delivery of the assessment engine does not necessarily need to be secure, but the delivery of assessment questions has to be as secure as possible without compromising access by examinee and invigilator.

Response return

The examinee's responses have to be collected, and that collection has to be as secure as for the delivery of the questions. The responses have to been attributed to the examinee.

Scoring, result determination and / or feedback

Having been collected the examinee's responses are then marked against the answers previously supplied by the examiner. The result of each question and the overall result have to be determined. For diagnostic and formative forms of assessment providing feedback to the examinee is a vital component of assessment. Feedback may occur at different times such as on-the-fly as each question is answered, at the end of the session or at a later stage from for example a tutor.

Data return

The examinee has to be informed of the result, and also the result has to be communicated to the examiners, administration and others who may be involved in the assessment process.

IMS Global Learning Consortium

There does exist an international organisation responsible for setting the standards in automated objective assessment: the IMS Global Learning Consortium (www.imsglobal.org). This organisation was set up in 1997 as a National Learning Initiative, a part of EDUCAUSE initiative.

By December 2002 the membership of the IMS includes many software suppliers, academics and other representatives across the education community and the work of the IMS is being universally adopted as best practise. The IMS develops standards or specifications for the educational environment that range from such aspects as notes through to the tracking of student progress and other administrative functions. The main objective of the IMS is to develop standards for inter-operability in general and to develop standards for items used in automated objective testing packages in particular. By the term "inter-operability" is meant that data collected in one computer-based system may be imported into another computer-based system.

However, this IMS Global Learning Consortium does not provide guidance on any standards associated with the automated marking of essays or free text responses. The author is given to understand that as a future task the IMS Global Learning Consortium intends to look at inter-operability in free text responses and essays.

1.6 Basic research questions

This chapter should have whetted the reader's interest in essay marking in general and caused them to ponder the possibilities of automated essay marking. The author suggests the following as a list of basic research questions that are relevant to this project:

- What, if any, are the problems of manually marking essays?
- What, if any, is the potential of using computers to mark essays?
- Is there any existing software that will mark essays?
- If not, could such software be developed?
- How effective could this software be in assessing essays?

This author predicts that it should prove possible to mark **some** essays by computer.

The author believes that computer marking of essays for style seems more promising than marking for content. The reason for this is that style does not require the software to attempt to make sense of the essays, merely to conduct word counts, sentence profile and other such objective measures.

Marking for content would require a deeper understanding of the use and interplay of words. Bloom's taxonomy ranges from *knowledge* to *evaluation* – so this gives rise to another question.

- How far along this taxonomy would automated content marking be possible?

A technical essay, say, on the operation of milling machines, must require different essay skills from an essay, say on the evaluation of a artist's painting.

1.7 Outline of the SEAR software

At this point of the thesis a brief outline of the **SEAR software** is warranted.

In essence SEAR marks **word-processed essays**, where [different] versions of Microsoft Word ®™ was used by the essayist.

For **marking style**, the examiner must first mark a statistically significant sample (at least twice the number of the metrics to be used) of the essays from the essay set. The sample is then analysed by the software to produce a weighted linear model, which is then used to mark all the essays in the given essay set.

For **marking content**, the examiner must supply a detailed answer schema and a model essay. The detailed answer schema is used to create content data structure, which is used to mark the model essay first, then to mark all the essay in the given essay set.

If the reader wishes to gain more information on the SEAR software system then in Chapter 3 and in Appendix K the author details his SEAR software system.

Chapter 2: Review of the current situation and literature on manual and automated marking of essays

2.1 Literature on manual and automated marking of essays

The purpose of this chapter is twofold. First, is to explore what literature there is on the marking of essays especially the problems therewith in order to establish if there is a case for considering automated essay marking. Second to explore what has already been achieved in the area of automated essay marking to create a foundation for this research project.

There is a considerable amount of information available on essays and essay writing, and this is mainly for the benefit of the essayist. Examples of this can be seen in Catterall (1998) and CLASS (2000).

Unfortunately there is very little information available to help essay markers conduct the process associated with manual marking. And there is even less information available on the process of automated essay marking.

This shortage of information leads the reader to assume that either the manual marking of essays must be very easy, indeed so easy as to be unworthy of discussion, or, that such marking is shrouded in academic mystery!

As a contradiction to this first assumption the author intends to show in the remainder of this chapter that manual marking is not easy. In fact, the author shall show that manual marking is a very difficult and onerous procedure. The shortage of information for helping essay markers may be seen as being an indicative measure of how hard essay marking is. If essay marking were as easy as writing essays, then the volumes of information available for essayists and essay markers would be more evenly apportioned.

Some of the researchers' results discussed in this chapter would lead the reader into thinking that essay marks were awarded by some mysterious process! Indeed the reader could think that the term 'magical' would be a more appropriate term to use than 'mysterious'. The inter-marker studies and the mark-remark intra-marker studies discussed later in this chapter indicate that assigning a specific mark to an essay sometimes appears to be a rather obscure procedure as opposed to the clear cut application of engineering or scientific principles.

For manual marking the case could be made for the establishment of policies or procedures aimed at the testing and indeed even re-testing of existing markers and for the recruitment of markers. The utilisation of such testing and re-testing policies and procedures facilitates the measurement of the quality, in terms of accuracy, reproducibility, and so on, of markers.

Initial measurements which are quantifiable in nature may thus be improved upon. For new markers there may be the opportunity for training in the awareness of the various problems associated with essay marking and thereby the opportunity for new markers to take the utmost precautions in the avoidance of these problems.

Shortage of information on automated marking may be caused by a variety of reasons.

In the first instance, there are very few researchers operating in this definitely unpopulated and problematic field. As a consequence of this, the source pool of information providers is very limited.

Secondly, manual essay marking is such a problematical and contentious subject area that only the most determined researchers will choose to operate in a field of study where there are no clear measurable baselines present against which to measure research.

Thirdly, obtaining "good" essay sets is not easy. Here the term "good" relates both to the quality of the manual marking and to the size of the essay sets. In this sense "good" would mean the performance of manual marking is high in statistical robustness.

Fourthly, getting essays marked by many experienced markers is an expensive process. This is expensive in monetary terms, in human resource terms and in terms of time.

There already exists an objective methodology in education. Educationalists have been exposed to an "atom-level" approach to learning and assessment, namely the research work of Skinner and his followers. Building on behavioural psychology, Skinner introduced the idea of programmed learning that dominated the progressive educational environment in the 1960's. The core of programmed learning is that learning materials are broken into small pieces of knowledge that may be (formatively) assessed at suitable points in the learning programme. In the context of this research project, Skinner's "atom-level" approach to programmed learning may supply a route into the marking of essay content.

2.2 Manual marking

In order to place the author's research in context it is necessary to establish what, if any, problems occur in the normal manual method of essay marking. Although essays have reputedly (Page claims that essays had been used by the Chinese some 2,000 BC) been used for assessment during the last 4,000 years, the marking of essays has several major problems. In spite of the many very varied problems covered in the following paragraphs of this section, assessment by essay is still widely practised and still widely accepted by staff in the role of marker and students in the role of essayist alike as a valid mode of assessment.

Virtually all the research conducted on the marking of essays and the problems of such marking, use essays written by school children in the early secondary education at onset of Piaget's formal operations stage. This presents the question as to why this particular age group is predominant. This author does not know, but presumes that these essays are of a size that is neither too large, nor too small. Essays from this age group are generally reported to be about 100 to 200 words per essay. Furthermore the thought invoking demands or cognitive demands placed on the children would not be too high. Essays produced by younger children would be too simple to provide significant data to reveal the effects being researched. Older children and young adults would be capable of producing fairly lengthy essays, and these essays should be more challenging of cognitive skills. Thus the demands of the markers used in research would be all the greater, more time consuming and, of course, more expensive were essays from this age group employed. The experimental design that could be required for research involving more complex essays might have to be more rigorous than that used for simpler essays.

The author feels that the approach used should be based on simple structures which may be built into more varied and complex structures and this indeed may have been the driving factor behind early approaches to essay marking research.

In the majority of research work read in this thesis the markers used have been a mixture of inexperienced student teachers and experienced teachers who are familiar with the teaching of early secondary school children. This author expects that in the case of experienced teachers it would have been appropriately experienced teachers who were used as the markers. Marking material related to their area of expertise is after all a basic aspect of their employment. The author is puzzled at the use of student teachers. Surely they are only learning their trade.

This author would understand the inclusion of student teachers if the research were specifically directed to establishing the effect of teacher training programmes on the evolution of marking skills.

Very often students of psychology are used as markers. Most research work cited in this thesis is reported in journals relating to psychology. The author wonders if it would be wrong to infer that marking essays does have some deep-seated psychological bias.

Style and content marking

Setting aside the differences between marking of style and marking of content, the major essay marking problems are described in the following sections. Having stated this, some of these problems are probably equally applicable to both the marking of style and the marking of content.

2.2.1 Subjective nature of marking

What it takes to make an "excellent" essay excellent is not clear. This section explores the considerable scope for subjectivity in marking essays. Sources of subjectivity are varied. Some sources are inherently related to this specific type of assessment instrument. Other sources are related to possible bias or possible discrimination associated with the markers. Yet other sources are directly attributable to the marking hierarchy of the marking process itself, first and second (and in some cases third) marking and external marking. Particular emphasis is brought to bear on two major sources of subjectivity. The first source of problems is the markers themselves. This particular source leads nicely to inter-marker and intra-marker considerations. The second problem source resides in the time-consuming nature of essay marking coupled with the provision of feedback to the essayist.

2.2.1.1 What is the effect on marking of the absence of a marking schema?

As there is generally no objective marking methodology for an essay, essay marking can only be subjective. This means that when awarding a particular mark the marker must be satisfied as to how fully the essayist has covered the marking schema. Has the essayist thoroughly explained key facts or merely mentioned appropriate key words or appropriate key phrases in all the right places? If there is no marking schema present then how is reproducibility, ability to be audited and so on to be demonstrated or ensured? Without a marking schema there is neither a basis for providing feedback to the students, nor a basis for conducting meaningful second marking. However, it is becoming de rigeur standard in Universities that detailed marking schemas are used especially for the awarding of degrees.

There is a trend or tendency or temptation nowadays to "share" the marking. This is especially true when large classes are being considered. And it is true that in the present educational environment large classes are not unusual. The term "sharing" covers different situations. Sharing may be applied to a situation where a team of markers applies a common marking schema but only mark material from their own tutorial group.

Sharing may also apply to a situation where the markers are expected to develop their own schema (private communication) for application to their own tutorial group. The team may consist of experienced staff either alone or in conjunction with inexperienced research staff or a mixture of both. In other words "shared marking" can range from "distributed marking" to "devolved marking". The implications arising from poorly orchestrated shared-marking systems are potentially very daunting. From private communications to this author from a number of Awarding Bodies, some markers have been known to be so liberal in their application of the given marking schema such that they, the markers, have reduced the value of the schema to being worthless.

2.2.1.2 What is the effect of the position of the essay in the set on the mark awarded?

While marking a large essay set, the marker may change what is worthy of a particular mark in the light of the standard of the essays marked. So the position of an essay in the essay set, or the order in which the essays are marked, may subjectively influence the actual mark awarded.

There is always the problem of the one essay that stands out from the rest. For example after a run of many very poor essays, an adequate essay may be awarded a better mark or grade than it deserves. Likewise, after a run of very good essays, a mediocre essay may get a particularly lower than deserved mark!

A study in 1975 (Hales and Tokar, 1975) showed that the supposition of the previous paragraphs is in fact correct. Hales and Tokar's study extended a theory originally conceived in 1951 by Helson. This is namely a theory of adaptation: an examiner will adapt the initial series of essays as being the norm. If this initial series of essays is of poor quality, then on encountering a better essay, the examiner will offer a better than deserved mark. The converse is also true. The numerical data produced by the Hales' 1975 study shows just how subjective this effect is.

Two target essays were graded as being 3.0 +/- 0.2 and then remarked following an initial run of 19 poor quality essays. This resulted in the first target essay mean mark being uplifted from 3.0 to 4.1 (i.e. a range of 1.1 marks), while the second target essay remained that same mean mark. Conversely when 19 good quality essays were used as the initial run the target essay mean marks were downshifted to 2.5 and 1.7 respectively. The second essay almost lost half its original mean mark.

Hales therefore did show that there is a considerable difference in the marks awarded to two target essays depending on the quality of the essays previously marked. So it may be concluded that the position of an essay in an essay set has significant effect on the mark awarded to it.

In 1984 Hughes and Keeling (Hughes and Keeling, 1984) conducted a study that both reproduced and also extended Hales' 1975 study. The extension was to provide half of the markers with three pre-marked model essays.

This study was carefully constructed. First 38 high school pupils aged 13 to 14 years wrote an short essay of about 250 to 350 words long on the topic of their hopes and aspirations for the next ten years. These 38 essays were typed verbatim and then formed into booklets to be marked by 25 experienced teachers. These teachers all had at least five years experience of marking similar essays. The marking scale used 0 to 25 marks. The mean mark awarded and the standard deviation by the 25 teachers were assigned to each of the 38 essays. Of these 38 essays, 23 were selected to be used in the next stage of this study. There was a two-fold selection process as essays were selected so as to cover as far as possible the whole of the range of the marking scale and where essays of similar mean marks were found then those essays with the smaller / smallest standard deviation were selected.

The 23 essays that were selected were bundled into three distinct sets. These were the context set of 15 essays, the model set of three essays, and the covariate set of five essays. Each set was packaged into separate booklets.

The context set consisted for five "good essays", five "poor essays" the one criterion essay, and four "filler essays". There were two versions of the context booklet produced, the "good context" and the "poor context". In the "good context" version the five good essays (in random order) were presented first then the criterion essay followed by a mixture of the poor and filler essays in random order.

For the "poor context" version the five poor essays were presented first (in random order) then the criterion essay followed by a mixture of the good and filler essays in random order.

The model set were three essays that were assigned the mean mark produced by the 25 teachers, but rounded to the nearest whole number. These essay were also labelled 'Superior', 'Average' and 'Inferior'.

The production of the context booklets combined with the model booklets created a four-fold experiment – good context +/- model set and poor context +/- model set.
For the final stage in this study Hughes employed 156 first-year college students taking a course in educational research methods to be the markers. 39 markers were randomly assigned to each of the four experimental conditions. The main finding of this study was that the availability of model essays did have an effect on the marking, but that the model essays did not reduce nor eliminate the context effect. In fact, the context effect is more significant than the effect of model essays being present.

The essays in the covariate set were randomly arranged in their booklet and they were used to determine marker performance. This was a subordinate objective of this study. All the markers used in the final stage of this study marked this essay set. There was no significant difference detected between the markers with and those without the model essays. However there was a significant difference between the students' and the teachers' marks in that the students awarded higher marks than the teachers. So it appears that either the teachers were more severe in their marking, or the students were more lenient in theirs!

Hughes finished the report on this study by suggesting that model essays may help with factual accuracy but may not help in marking style. Hughes also suggested that there may have to be an acceptance that context effects are not avoidable.

2.2.1.3 What is the effect of vocabulary, voice and spelling or grammatical errors on the mark awarded?

For a given essayist the vocabulary and voice used, the tone and the volume of both spelling and grammatical errors will certainly be candidate factors which consciously affect the marks awarded for style. But the question still remains as to what effect, if any, do these factors sub-consciously have on the awarding of marks for content. This author has found no published research for the factors of vocabulary, voice and tone in this area of marking.

A study by Marshall (1967) investigated the effect of spelling, grammar, punctuation errors and combinations thereof in essay marking. The results of this study were twofold. Firstly, the results showed that spelling errors and grammatical errors do affect the mark awarded to the essays. Spelling errors and grammar errors have the effect of lowering the mark awarded to an essay. Secondly, even when directly instructed to unambiguously ignore spelling errors and grammatical errors, examiners still awarded a lower mark to essays in which these errors occurred. And it was found that the more errors present in an essay the lower the marks awarded to that essay. Marshall concluded that the severity of effect of the errors depends on the types and number of errors. Spelling errors and grammatical errors give relatively equal effects and the effects of both of these types of errors are more significant than the rather weak effect due to the errors of punctuation.

This is not too surprising as spelling errors and grammatical errors tend to greatly impede the readability and understandability of text. Punctuation errors do not too greatly impede the reader's interpretation of the text.

2.2.1.4 What is the effect of the introduction and the conclusion of the essay on the mark awarded to that essay?

In 1993 a specific study (Townsend et al, 1993) was conducted into the effect of alternative introductions and alternative conclusions in the marking of essays. This study revealed that changing introductions has a significant effect on essay grade whereas changing conclusions showed little significance. For each essay marks were allocated for the introduction, the main body and the conclusion. In this study the marks awarded to the bulk of the essay remained unaltered while the effects of the alternative introductions and of the alternative conclusions were being considered. Marks did reflect a change according to these alternatives and thus shows the importance of each of these parts on the variability of marks and also highlights the importance of recognising just how sensitive markers are to these two components of an essay.

2.2.1.5 What is the effect of the marker's mood at the time of marking?

A study in 1989 (Townsend et al, 1989) was specifically conducted to find the significance of the effect of the marker's mood at the time of marking. In this study groups of markers had their mood "conditioned" by watching an appropriate film immediately prior to marking the essays. The study concluded that there was no support for the existence of a mood effect in the marking of essays. This, in the author's view, does not seem to be true.

However, Townsend's study contradicts its own conclusion as it repeatedly notes that the first marked essay is affected by the mood of the marker. The markers were categorised as "good mood" and "poor mood" by watching specific films before they undertook their marking duties. Markers had their "good mood" induced by watching a comedy film. Markers had their "poor mood" induced by watching a sad film. It was found that markers in a "good" mood tended to mark the first essay higher than the control group of markers, whilst in the case of "poor mood" markers the mark awarded was less. However, the study gives no clear conclusion of the effect of marker's mood. In the discussion section it states that the limited effect seen may arise from the essay set used in the study. The essays used were short, simple and not cognitively challenging. Townsend's study suggested and alluded to the possibility of greater mood effect when essays of a more challenging nature are being marked. The question of how mood would effect higher stakes essay marking is left unanswered.

2.2.1.6 What is the effect of the media on which the essay is submitted?

A study in by Peacock (1988) revealed that word-processed essays usually scored one grade higher than hand-written essays. Moreover Peacock found that the effect was more pronounced the poorer the essay. However, no discussion of the quality of the handwriting was included in this study. The reader would accept the assumption that variation of word-processing in the aesthetic sense is considerably more limited than the variation that exists in hand writing.

The study by Markham (1976) provided confirmation that handwriting quality affects marks awarded to essays. This study showed that, regardless of content, those essays with better handwriting consistently received higher marks than those essays with poorer handwriting. Obviously if the writing cannot be deciphered by the marker then no marks will be awarded. However, the quality of handwriting is not discretely scaled, nor for that matter is the markers' ability to read any 'level' of handwriting. Indeed humans are inherently variable in their abilities to read handwriting. Nevertheless there remains the suspicion that sub-consciously a marker equates better quality handwriting with better quality content and as a consequence of this a better mark is awarded.

As mentioned previously the markers' ability or skill in reading various hand-written documents is also a factor here. The author has not found any research where this skill has been investigated.

2.2.1.7 What is the effect of subject on the marking?

Referring to the summary of Starch's three subject studies (Starch and Elliott, 1912 1913a 1913b) found in Table 2.2.2.1.b below, the data presented there clearly indicates that there is a difference in the performance of marking by subject. Starch considered the three subject areas of English, Mathematics and History. The results of this study showed that the marking of English essays fared better than Mathematics (specifically Geometry) calculations and History essays. This result is counter-expectation, as Mathematics is one of the so-called numerate subjects where the answer to a problem is clearly right or clearly wrong. The probable error for English, between 4.0 and 4.8, is about half that for the other two subjects researched by Starch, where he found the probable error was 7.5 for Mathematics and 7.7 for History. A wide range of marks was awarded to the same standard paper in all three subjects studied.

So, it is concluded that subjective marking is not the province of any one particular subject.

2.2.1.8 What about the effect of bias, or discrimination, by the markers on the essayist?

There have been numerous studies into possible marker bias, and marker discrimination in essay marking. Such studies have included the effect of first name, the effect of perceived race, the effect of perceived gender, and various other perceptions. This author enlarges on several of these factors in the following text.

2.2.1.8.1 First Name Stereotyping

Both Hararp and McDavid (1973) and Erwin and Calev (1984) have conducted studies on first name stereotyping. Their results are summarised in the following table dealing with how to receive better marks than expected by choice of first name.

Effect	Girl	Boy	Researcher
Good	Alison	Steven	Erwin, 1984
Good	Adelle*, Lisa	David, Michael	Hararp, 1973
Bad	Beryl	Norman	Erwin, 1984
Bad	Karen, Bertha	Elmer, Hubert	Hararp, 1973

Table 2.2.1.8.1 First Name Stereotyping

Both these tightly orchestrated studies revealed that the first name of the essayist does have a significant effect on the mark awarded for the essay. According to Erwin, attractive first names attract better marks than do anonymous names. Unattractive first names received poorer marks than anonymous names.

However, these studies reveal more than pure effect of choice of first name. In Hararp's study first names were pre-processed to determine the attractiveness of first names by gender. This showed that Adelle, the name asterisked * in the table, was ranked as being a fairly unattractive girl's name. Further, contemporaneously with the essay marking study a woman called Adelle Davies, apparently a famous health food nutritionist, was figuring largely in the press and for 'Adelle' the essay marks awarded in the study were inflated. It is possible that positive media coverage had a positive effect on the outcome of the study as the original pre-determination of 'Adelle' was that is was not attractive. Does this show the effect of the Press? Hopefully it does not infer that to get a better essay mark an essayist should get his or her name in the Press!

Hararp's study showed that the more experienced the marker, then the more pronounced the stereotyping effect was. This is a surprising finding. Before reading Hararp's study, this author would have suggested that it is the more inexperienced markers who would have been susceptible to bias, and not the experienced markers. For experienced markers it would be expected that professionalism, coupled with years of practice would have eliminated first name stereotyping.

Erwin's study also showed that the markers showed stereotyping effect. The nature of, and the degree to which, the effect is present is comparable to Harai's findings.

2.2.1.8.2 The Race of the Author and (Reverse) Discrimination

A complex study by Fajardo (1985) showed that, even when deliberately instructed to ignore all information on race, markers respond subjectively to race information. This study used four pre-processed essays of different standards and 160 markers. The essays used varied in quality. One essay was of poor quality, two essays were of average quality and one essay was of good quality. The study showed that those essays purporting to be authored by black persons were awarded significantly higher marks than the same purporting to be authored by white persons. The study thus revealed that discrimination was a significant factor and that as well as this discriminating effect in certain circumstances some sort of "affirmative action" or "reverse discrimination" took place.

It was noted that the effect of the reported discrimination was not constant. Those pre-processed essays classed as average suffered a greater effect than those essays that the pre-processing classed as poor or good.

2.2.1.8.3 Gender of the Essayist

The effect of First name stereotyping was found to be more pronounced for male names than for female names. Both the separate studies conducted by Hararp and McDavid (1973), and, Erwin and Calev (1984) noted the gender effect. Both researchers have made attempts to explain possible factors as to the source, or indeed sources, of the gender effect. Neither gave objectively provable sources for the gender effect, but both gave a different possible source or sources.

A third study, which investigated the gender relationship between student and examiner, was conducted by Carter (1952). This study also showed a gender effect of the "essayist" coupled with the gender of the examiner. Female examiners generally awarded more marks than did male examiners and it was found that male students were awarded fewer marks than female students were.

Unlike the two studies cited in the earlier paragraphs, Carter's research did not involve English essays, but was in the subject of elementary algebra. To obtain the maximum subjective gender effect in elementary algebra it appears that one would have to be a female student and seek to be marked by a female examiner. Carter's study mentions eight other researchers' works {namely in alphabetical order: Day 1937, Douglass 1938, Edmiston 1943, Garner 1935, Lobaugh 1942, Newton 1942, Shinnerer 1944 and Swenson 1942} that also revealed pronounced gender effects in the marking process.

2.2.1.8.4 Markers' Expectations

A very unexpected side effect of first name stereotyping is that of the individual marker's own name bias. Erwin and Calev's study (1984) provided evidence of this effect. However Hararp and McDavid's study (1973) explored this significant effect further. Harap found that examiners were favourably disposed towards attractive or common first names and were not favourably disposed to uncommon or unattractive first names. The scale of favourability was identified by the mark awarded. The favourable element was shown by the award of higher marks whereas the opposite was shown by the award of poorer marks. A strange effect found was that if the marker and essayist had a first name in common there was a trend for more marks being awarded to the essay. The author speculates that this is a case of judging essayists by their names.

In the 21st century the multicultural nature of European countries may have reduced this problem to zero, but there appears to be an absence of literature to that effect.

2.2.1.8.5 Appearance of the Essayist

This appearance characteristic is an extension of the above in terms of examiner's expectation. However the study by Landy and Sigall (1974) was based on essayist's facial characteristics. Landy conducted this complex study in 1974 when there was less political correctness than there is today. Its purpose was to investigate the effect of an essayist's appearance on the marks awarded by the examiner. The study revealed a strange subjective effect on the marks awarded when the examiner is shown 'incidentally' an image of the essayist. In this study Landy used both essays alone and essays in conjunction with a photograph. Landy found that the same essay when marked by some 60 male psychology students rated those essays purporting to be produced by an "attractive female" (via a photograph) higher than those essays that were 'missing' the photograph. In turn these 'faceless' essays were marked higher than those purporting to be the work of an "unattractive female" (again via a photograph). This subjective manipulation was shown by Landy to be greater when poorer quality essays were being marked than for better quality essays.

Landy then projected this effect to other arenas in which performance is measured. For example, Landy further posed a question of whether a person's appearance has any significant effect in the case where that person performs impressively.

Landy also went on to pose the question of how significant does a person's appearance affect the appraisal of a sub-standard performance. He wondered if it were possible that the "better looking" person gets favourable treatment when producing inferior work. Landy extended this even further by hinting that "attractive people" will be given the benefit of the doubt when performance is of dubious quality and that "ugly people" might not have the luxury of such benefit offered to them.

Pushing this subjective reaction to a person's appearance yet even further, Landy posed questions on the more general situation of how other people react to one's appearance and how this impinges on their reaction to one's accomplishments.

2.2.1.8.6 Integration of discrimination effects

At the University of Indiana a professor of Educational Psychology, Clinton Chase (1986) reported on his investigation into the integration of the subjective variables of sex, race, reader expectation and the quality of the handwriting.

Chase had created a student essay of moderate quality on the topic of pollution, which was purported to be a fifth grade social studies essay. The essay contained errors and finished with a false conclusion. Two hand-written versions of the essay were produced one 'poor' (rated about 2 on the Ayres Handwriting scale) and the other 'good' (rated about 8 on the same scale).

Having created the essay Chase then prepared a packet for each of the markers. The packet contained the following items:

- One version of the essay,
- Scoring instructions on 10-point scale,
- The source text on which the essay was based, and
- A synthetic student school record.

This record contained details on the student such as academic records, photograph some biographical details and so on. In all, some 16 different packets were prepared.

Eight-three experienced teachers covering a wide variety of classroom backgrounds were used in this study. Seven markers of the markers were black while 76 were white, and eight were male while 75 were female.

In summary Chase's findings confirmed the discriminatory effects found by the other researchers.

2.2.1.9 What is the effect of the size of the essay set being marked?

The marking of a large essay set will take time and effort. The time alone is often of significant magnitude. So, which of us as a marker can guarantee that he or she would be consistent in both accuracy and diligence over an extended period of time? Often the time is fragmented, and more often than not the period of time occurs late at night, early in the morning (time categorised as non-quality) and at the weekends (time considered anti-social). This represents poor rest and recreation opportunities for the staff, with all the resulting medical and legal effects. It has been proven that most people can only concentrate for about 20 minutes at most at any one stretch of time.

Take a hypothetical example. Suppose there is a first year class of 200 students who each submit one essay and that it takes 15 minutes to mark each essay. This situation then represents some 50 hours of marking activity for the first marker. Additionally there is the time taken for second marking. For each additional layer or level or operation the time taken for marking increases.

It is quite apparent from this that the larger the essay set the more time consuming will be the marking process. Coupled with the direct stress on markers to mark large essay sets per se, there is the administrative stress on markers. Administrative stress is caused by the fact that many educational organisations expect, and lead the essayists to expect, that feedback will be provided by the markers within some time period after the submission deadline. Often this period of time is set to two to three working weeks. The complaints made to, and the enquiries to, markers who fail to achieve this feedback deadline also creates more staff stress and removes staff time for other purposes. This, surely, must have a negative impact on marking performance and a negative impact on subjectivity in marking. Markers, essayists and educational organisations find the foregoing impacts undesirable.

2.2.1.10 "Own tutee" effect

In the early 1980's Colin Byrne (1980) conducted a study of Open University inter-marker marking reliability. This study is covered in the next sub-section of this chapter in 2.2.2.11.

However Byrne detected in all three sets of inter-marker comparisons on which he reported that tutors appeared to be lenient to their own tutees and appeared to be more severe with tutees from other tutors. This was found to be particularly so in the group derived from Arts, Social Studies and Education. Therefore it would appear that within a cohort not all students are treated equally and that a degree of familiarity could affect the marks awarded.

2.2.1.11 Essay Length

A further factor in the marking of essays is that of essay length. Two studies that have been conducted into the effect of essay length on the marks awarded to essays are considered here. The first is by Tollefson and Tracey (1980) and the second is by Hall and Daglish (1982). Both studies conclude that "length matters" rather than "quality matters". Each of these studies will be considered in the text that follows.

2.2.1.11.1 Tollefson and Tracy

After a complex selection process, N Tollefson and D B Tracy (1980) selected ten essays from a pool of thirty. These ten essays were classed into two groups of five each, one 'poor' group and one 'good' group. In each group of five there were two long essays, four moderate length essays, and four short essays. The breakpoints between short-moderate and moderate-long were about 60 words and about 140 words respectively. Eighty-eight markers were involved in marking these ten essays.

After extensive statistical analysis the findings showed that, apart from one case, the longer the essay then the mark awarded was significantly better. The exception to this pattern was the particular case of good quality essays for which there was no significant difference in marks awarded between short and moderate length essays.

The good quality long essay had the highest average mark and the smallest variance in marks awarded. However, average mark for the poor quality long essay proved to be not significantly different from short and moderate length, good quality essays. It appeared to Tollefson and Tracy that there were serious questions about the validity of marking procedures used by many markers, and that marks that were awarded were mainly due to length of the essay not its content.

2.2.1.11.2 Hall and Daglish

C G W Hall and N D Daglish (1982) conducted an exploratory study on the effect of essay length on the awarding of marks. The report on their study started by criticising the work done by Tollefson and Tracy (in the sub-section immediately above) by expressing doubts on the scripts that Tollefson and Tracy used, which were from 35 words long to 145 words long, as not being "proper essays".

In this study Hall and Daglish used eight markers of varying experience (from three markers who were senior lecturers, down to two markers who were post-graduate students who had never marked before). These eight markers were tasked to mark five pairs of essays on a 25-point scale.

Great care was taken to ensure that each of the five pairs of essays were clearly of different standards of quality and that for in each pair the quality was as close as possible equivalent. The qualities chosen were A (20 out of 25), B1 (18 out of 25), B2 (16 out of 25), C (14 out of 25) and D (12 out of 25).

In each pair of essays one essay was classed in terms of word length as being 'long' and the other classed as being 'short'. By a simple inspection of the original marks versus essay length in words the study concluded that the longer the essay the higher the mark awarded.

There was only one notable exception. The exception was in the case of the pair of essays for the 'C' quality rating where the "long essay" was marked lower than the "short essay".

In their discussion Hall and Daglish felt that this study did neither statistically confirm nor deny the effect of essay length on the mark awarded. This author thinks that, circumstantially, there is some evidence in Hall and Daglish's study that essay length does have some effect on the markers and hence on the marks awarded.

Two further specialised aspects of this subjectivity problem are dealt with in the following sections. These are inter-marker problems and intra-marker problems.

2.2.2 Inter-marker problems – human versus human

The following text assumes, of course, that there is an inter-marker problem existing in the first instance. In order to gauge just how extensive this problem is, this author repeats the quotation that McDonald and Samson (1979, pp. 45 - 46) had included in their paper.

> "*Fifty candidates sat one three-hour single essay paper. The scripts were then marked out of 100 by five different examiners. When the five marks were averaged out all candidates' marks were very close, so close in fact as to make it impossible to grade the results into 'classes'. But the difference in the marks given to individual candidates by separate examiners varied by as much as 36 marks out of 100. The average difference on an individual script between the five examiners was 19.*" McDonald and Samsom (1979, p. 45 - 46)

With the aid of the numerical data in the text quoted above then, clearly, there is a problem in the reliability and consistency of markers.

Subjectivity resides in any situation where there is a human decision-maker. The previous section sets out the potential factors contributing to the subjective nature of marking.

In the light of this, the involvement of a second marker only compounds the subjectivity problem. No two markers will have exactly the same experience of marking. In fact, it is difficult to see how two markers can ever have similar experiences of marking.

This subjectivity is illustrated by the following two quotations, taken slightly out of context, from Lynda Markham's (1976) investigation on the effect of handwriting quality on the marks awarded to essays.

The first quote taken is

> "*older teachers did not give significantly different ratings to papers than did younger teachers*". Markham (1976 p. 280)

The terms older and younger are clearly defined in the study as referring to the experience of the teachers rather than to their physical age. This study grouped the teachers involved into three groups. These groups were categorised according to the teachers' experience as 0 to 4 years, 5 to 9 years and more than 9 years. The data set, investigated was generated by children in their "fifth grade" at school. In the light of this data set it may be asserted that the cognitive requirements placed on the essayists and markers might not have been too demanding.

The second quote is again taken from Markham's investigation and states

> "*The student teachers tended to rate all papers higher than did teachers.*"
> Markham (1976 p. 280)

Markham felt so strongly on this finding that the last paragraph in the results section of her study is completely devoted to this difference in the marks awarded between student teachers and experienced teachers.

The Kolb Cycle is a well-known cycle pertaining to the practise of teaching. This author now turns to this cycle to further clarify thoughts on marking. The Kolb Cycle identifies four stages, which have been labelled Concrete Experience, Reflective Thinking, Conceptualisation, and Plan New Learning.

It is possible that for an essay on the Kolb Cycle that different markers will award full marks for a 'perfect' description. But what if the essayist's description was slightly wrong? In this case what marks should be awarded? Should the mark awarded by the markers be zero marks, some marks, half marks or something else?

Situations all too commonly occur where markers are found to be consistently:

- too lenient or generous,
- too severe or hard, and
- inconsistent.

This produces two further questions:

- Who can decide which marker is awarding the "true mark"?
- How may it be proven that the "true mark" has been awarded?

During the course of his research this author has discovered the existence of reported instances where agreement between human essay markers produces an excellent correlation coefficient of 0.8. The range of values for inter-marker reliability that has been reported is shown in the following table giving the researchers' name, date of reporting their research and the results of their investigation. The lowest correlation coefficient was a poor 0.009 and this was produced from Cast's brace of matching studies (Cast, 1939 1940).

The table below serves to provide an overall view of the range of correlation for manual marking reported by various researchers as well as an indication of the correlation reported by researchers for automated marking. The range of correlation for automated marking extends from 0.927 down to 0.44, as seen in section 2.5.3 below.

Each of the researchers' work is briefly described in the following the table. Note that Starch and Elliott (1912 1913a 1913b), whose research work is described in this section, does not appear in the table because they did not use correlation in their discussion of inter-marker performance. Instead, Starch and Elliott used a methodology based on calculating the standard error of a mark. In this respect it is therefore false to include Starch and Elliott's findings in this table, but their findings are very germane to the consideration of the inter-marker problem.

Researcher	Number of Markers	Style (S) or Content (C)	Overall Correlation Max	Overall Correlation Min
Blok 1985	16	S	a) 0.670 b) 0.669	a) 0.180 b) 0.269
Cast 1939, 1940	12	S	0.779	0.009
Fajardo 1985	46+	S	0.96	0.88
Finlayson 1951	6	S	0.824	0.591
Kniveton 1996	2	S	0.72	0.19
Jacoby 1909	6	C	0.973	0.516
Nyberg 1980	6	S C	0.949 0.952	0.592 0.609
Wiseman 1949	4 4	S S	0.85 0.73	0.72 0.53

Table 2.2.2: Summary of Inter-Marker Correlation (Manual)

The results revealed in Table 2.2.2 show that a 'high-water mark' of about 0.8 for inter-marker correlation may often be achieved. Yet there are many references and evidence, some taking the form of many private communications with the author, of inter-marker correlation being quite poor, often about 0.3. Thus the range of this reported inter-marker correlation extending from 0.009 to 0.96 is very large. This is significantly large enough to cast doubt on the reliability of any human essay marking. It follows that this state of affairs creates a huge opportunity for appeals against results, grade, and degree classification.

2.2.2.1 Blok

Blok (1985) created a large scale study using 16 elementary school teachers as markers working with 105 essays on two separate marking sessions. In effect this study is both an inter-marker study and an intra-marker study. The instructions to the 16 experienced markers were to holistically mark the 105 essays as quickly as possible, on a 10-point scale where 1 represents "very poor" and 10 represents 'excellent'. On this 10-point scale only the two extreme values, that is 1 and 10, were labelled. All the other scale points were left unlabelled.

From the first of the two marking sessions the maximum inter-marker correlation was found to be 0.670, whilst the minimum was 0.180. The second marking session however produced a maximum of 0.669 and a minimum of 0.269.

This shows little change in the value of the maximum inter-marker correlation but the minimum value is almost doubled in value. It could be argued that even with a maximum correlation of about 0.67 these 16 markers did not show any excellent performance.

On deeper examination it is revealed that different markers were responsible for these maximums and minimums. In the first marking session, markers 11 and 15 showed the maximum, while markers 3 and 5 showed the minimum. In the second marking session it was markers 2 and 12, 12 and 14 respectively. So by considering the markers at the extreme ends of the correlation range in both marking sessions different markers have involved, therefore no one marker appeared to be extremely consistent. Marker 12 in the second marking session was paired in both the maximum correlation, with marker 2, and at the same time paired with the minimum correlation, with marker 14.

2.2.2.2 Cast

Cast (1939 1940) considered four different methods of marking essays namely:
The first method was the Individual in which the marker uses their own method of marking. The second method was Achievement of Aim in which a judgement is made as to how well the essayist achieves the purpose of the task. The third method considered was the General Impression in which a holistic view is taken of the essay, and the fourth method Analytic in which a fixed detailed marking schema is used.

Of these four alternative methods Cast stated that the last method, the Analytic Method, was the best. He considered this 'Best' by virtue of the statistics of range of marks awarded, the average and the standard deviation of each of the four alternates.

These statistics showed that the Analytic Method produced the lowest values in all three statistical measures with a significant gap in performance between this method and the closely grouped other three methods.

Although it performed best of the four alternative methods of marking essays, Cast did report some worrying values of correlation for the twelve markers listed in the report. The maximum value found was 0.799, which is a respectable level of correlation. The average correlation value for the twelve markers was 0.493. This average is not too acceptable. Worse was to follow as the minimum value found was 0.009, with the value of 0.102 as the next lowest. A value of 0.009 is nearly no correlation at all.

Even if one were to disregard this value as being so extreme that is not likely to be ever repeated the next lowest value of 0.102 is still not acceptable.

It is not known what threshold would be a minimum acceptable value of correlation, but surely a value of 0.102 would be below this threshold.

2.2.2.3 Farado

Much of Farado's work has been covered earlier in this chapter, excepting inter-marker correlation. In the pre-experimental stages for this research study, the correlation among 32 senior-level under-graduates, 14 graduates (who taught freshman composition) and an un-specified number of "C.E.E.B judges" (College Entrance Examination Board) was determined. Farado (1985) reported that the correlation determined ranged from 0.88 to 0.96 among the four essays that were ultimately used for the main experiment. This represents very high correlation among the various markers.

2.2.2.4 Finlayson

Finlayson and Wiseman carried out research into essay marking but both these two researchers reported on their work in the area of intra-marking. Intra-marking is discussed later on in this chapter.

Finlayson (1951) employed six markers in his original two-essay study in inter-marking in which he found that the range of correlation extended from a maximum of 0.824 to a minimum of 0.591. The level of standard deviation (in marks) of the marking ranged for essay set X from at best 2.184 to 3.475, whereas with essay set Y the range extended from 2.218 to 3.62. In terms of mean mark produced by the six markers the range is from 7.558 to 11.492 for essay set X and 7.766 to 11.132 for essay set Y. So, there is a suggestion from these descriptive statistics that inter-marker correlation would be good. In fact, the correlation ranged from a minimum of 0.591 to a maximum of 0.824, which is acceptable.

In examining the six individual performances the situation appears less acceptable. Using Finlayson's own identifiers marker "2" showed a limited spread of marks, while marker "3" had a much bigger spread.

Between marker "1" and marker "6" there is a difference of four marks in their means. That difference represents a fifth of the range of marks awarded, as the essays were to be marked out of 20. If the pass mark were to be set at 10 out of 20, then on average marker "1" would fail more essayists then marker "6". This is clearly not acceptable.

2.2.2.5 Wiseman

Turning now to Wiseman's two-part study (1949) into inter-marking, the first part of this study took place in 1943 and the second part in 1948. Wiseman identified his markers as A, B, C and D (in the 1943 part) and A, E, F and G (in the 1948 part). The marker identified as "A" was the same person in both parts of this study. In 1943 the range of correlation was found to be from 0.72 to 0.85, whilst the corresponding correlation 0.53 to 0.73 in 1948. Wiseman claimed that it was the lack of self-consistency of marker "G" that lowered the 1948 inter-marker correlation performance. According to Wiseman an inter-marker correlation of less than 0.6 is not likely to be generated if all the markers employed are in fact "self-consistent" markers. By the term "self-consistent" Wiseman expected that the mark-remark correlation of an individual marker would be better than 0.7, and that any value less than 0.7 would be unacceptable.

2.2.2.6 Kniveton

Kniveton (1996) reported on a research study in which a correlational analysis was conducted between essays and multiple-choice assessments. Here the current author concentrates on the essay part of this research study.

From his background preparation to the study Kniveton knew that marking of essays was not reliable. Therefore he tasked two independent markers to mark an essay initially. The following decisions and actions were then taken. If the marks for any essay were different by no more than 10%, then the average mark was used. Where the difference was greater than 10% then a third independent marker was used. The essay mark then used was the average of the closer of the two marks. This resulted in a third marker being used in 11% of all the essays used in this study.

The reasons for the use of a trigger value of 10% difference between the first two markers in order to evoke the third marker is not clearly given in the research report. Regardless of the reasons this policy resulted in about one essay in every ten necessitating the input from the third marker. This author assumes that setting the trigger to 10% was a compromise. If the trigger were set to, say 5%, then probably there would have been a very much higher demand for the services of the third marker. Setting the trigger to, say 15% then the amount of essays with widely differing marks could lead to problems in the statistical inferences that Kniveton hoped to make.

The essayists had to write on a choice of one from two topics that had been made known to them some weeks beforehand. The topics were in the same subject area as the multiple-choice questions also used in Kniveton's study. Under examination conditions each of the essayists was allowed 55 minutes to write the essay. These essays were marked according to an honours degree classification scheme. This scheme is where a score of 70% and over was rated "1st" class, a score in the range of 60-69% was rated "2.1" second class division 1, a score in the range of 50 to 59% rated as "2.2" second class division 2 and a score in the range of 40 to 49% rated "3rd" third class. Under this scheme any mark under 40% was rated as "fail". In Kniveton's reference there is a fulsome description of what would represent each of these grades.

This study involved some 1,000 undergraduate students of both genders studying at different years in 23 different Psychology courses hosted by three different UK universities. Classes ranged in size from 14 to 99 students.

In the study Kniveton limited his report to 720 students from 16 different courses. The maximum correlation found was 0.72 with a minimum of 0.19, which is quite a large difference. A non-weighted average of 0.40 was also discovered. However, Kniveton further examined the essay marks and found more alarming data. Across the 16 courses reported, the range of standard deviation extended from a maximum of 14.84 to a minimum of 3.19 marks representing a difference of 12 marks, with an average of 9.6 marks! The range in marks awarded by course had a maximum of 65 marks to a minimum of 11 marks, with an average of about 39 marks!

Further, the course exhibiting the maximum standard deviation had the maximum range of marks, while the course that exhibited the minimum standard deviation also had the smallest range in marks. Unfortunately Kniveton did not appear to follow up on the reasons for such a wide range of essay marking performance across these 16 courses within this study.

2.2.2.7 Jacoby

In 1909 Professor of Astronomy H Jacoby (1910) sought to ascertain the precision with which essays are marked. Jacoby sent 11 samples of his own students' work to five other astronomy professors. These samples represented a fifth of the class. During this study no marker was permitted to see the marks awarded by the other markers. Jacoby had the 11 essays accompanied by instructions of how to mark the essay to enable standardisation of marks.

The instructions were to ignore spelling errors, level of neatness and so on and thus to ensure that marks would only be assigned on the basis of astronomical skill of the essayists. A code letter identified each marker. The pass mark was set by Jacoby at 6 out of 10. Unfortunately one marker appeared to set the pass mark to 5. As an aside the author wonders just how often markers fail to follow marking instructions.

This is the only research work found by the author in which the raw data is so clearly given. A copy of Jacoby's original data is in the Table 2.2.2.7.a below, which also includes the variety of the various results obtained from each marker. This is followed by Table 2.2.2.7.b, which shows the inter-marker correlation revealed, by this study.

Essay	Maker A	Marker B	Marker C	Marker D	Marker E	Marker F
1	9	9.0	8.5	7.2	9	7.3
2	7	6.6	7.0	5.9	6	6.5
3	9	9.0	8.8	7.2	8	8.0
4	10	9.4	9.9	8.0	10	9.2
5	7	6.2	6.7	5.8	7	5.9
6	10	9.8	9.6	7.6	10	9.5
7	6	5.8	6.3	4.6	7	5.4
8	9	9.3	9.7	8.0	9	8.8
9	8	5.7	9.0	6.7	10	8.7
10	10	8.5	9.1	6.2	9	9.0
11	9	9.0	9.5	6.1	8	9.0

Table 2.2.2.7.a: Jacoby's inter-marker raw data

Marker	A	B	C	D	E	F
A	-	0.876	0.911	0.776	0.762	0.898
B	0.876	-	0.784	0.751	0.516	0.707
C	0.911	0.784	-	0.815	0.824	0.973
D	0.776	0.751	0.815	-	0.726	0.726
E	0.762	0.516	0.824	0.726	-	0.809
F	0.898	0.707	0.973	0.726	0.809	-

Table 2.2.2.7.b: Jacoby's inter-marker correlation

Of the six markers, three passed all the essayists, two markers each failed two students and the remaining marker failed three students. This is a sad reflection on the six markers. Worse follows when the four essayist failures are examined in further detail. One essay was awarded a fail three times: by markers B, D and F. Another essayist was failed twice by markers: D and F. However, the remaining two failures noted were awarded in each case by a single marker: by marker D and marker B respectively. So there is no consistency in the failure mode other than the suggestion that marker B (and possibly marker F) appears to mark harder than the other markers.

On considering the passes awarded there is no consistency. Some essayists were awarded full marks by some markers, and yet other markers awarded those same pieces of work bare pass marks.

This study appears to neatly encapsulate the problems in manual marking. The author begs the question as to which of the markers is actually producing the accurate marks.

2.2.2.8 Nyberg

VR Nyberg and AM Nyberg (1980) undertook a study that specifically investigated the reliability of essay scoring. Six experienced teachers of high school English were given marking models and, importantly, instructions to practise using these marking models. Each marker graded each essay twice, once for content and once for style.

Not surprisingly, excellent inter-marker correlations were obtained. Although the results obtained were excellent correlations they were not perfect correlations as there was still a large spread in performance of the markers. Nyberg, in the last sentence of the report, gives some chilling advice in that at least six essays should be used in order to make reasonably sound decisions for essayists! This is sound advice indeed!

2.2.2.9 Starch

In the early 1910's Starch and Elliott (Starch and Elliott, 1912 1913a 1913b), produced a study on the reliability of marking of three different subjects. The subjects they considered were English, Mathematics (specifically geometry) and History. The table below highlights some of the important statistics produced by this three-subject study.

Subject	**Number Of Markers**	**Paper**	**Maximum Mark Awarded %**	**Minimum Mark Awarded %**	**Median %**	**Probable Error As %**
English 2 papers A, B						
Principal Teachers	142	A	97	60	88.2	4.0
		B	98	50	80.2	4.8
Student Teachers	86	A	99	74	92.4	
		B	74	65	84.5	
Educational Measurers	98	A	99	63	86.7	
		B	63	64	80.5	
Small Schools	39	A	97	60	89.5	
		B	93	50	82.0	
Large Schools	42	A	97	50	86.8	
		B	93	61	80.3	
Mathematics 1 paper	138		93	28	70.0	7.5
History 1 paper	114		92	43	70.8	7.7

Table 2.2.2.9: Summary of Starch's three subject study

Each of these three studies covered four papers for the four subjects, and marking was performed by a considerable number of principal subject teachers. Marking schema and clear instructions accompanied each of the papers being marked. These papers consisted of one paper each in both Mathematics and History and two papers in English.

Starch and Elliott revealed several findings with these studies and the main points are given below.

In the first instance they discovered that surprisingly the marking of English showed a probable error in the order of half that of Mathematics (Geometry) and in the order of half that of History. This is contrary to the commonly held assumption that the marking of Mathematics should offer very high performance. It is commonly held that a low probable error, even possibly zero, should be exhibited by mathematical submissions.

It must be noted that the data produced by Starch and Elliott were adjusted so that a virtual pass mark of 70% was maintained across all three of the subjects, and this was used for all the markers recorded in these studies.

Secondly Starch and Elliott noted that the marking of English by principal teachers in small schools appeared to be significantly higher than their counterparts doing the marking in large schools. This author surmises that perhaps in small schools teachers are more acquainted with their students, while large schools are somewhat more impersonal; and this could possibly lead to the awarding of higher marks. But it could also be perhaps that large schools deploy a policy of stricter marking to preserve their academic rigour and thus maintain or increase their academic standing.

Thirdly Starch and Elliott discovered that student teachers marked more leniently than both principal teachers and educational measurement experts. This may be due to student teachers being more willing to give the "benefit of doubt" to the paper, while experienced teachers are more grudging in offering "benefit of doubt". Of course it could be that inexperienced teachers may have an in-built reluctance to mark strictly.

Fourthly Starch and Elliott revealed that the range of marks awarded was wide, and this was regardless of which subject was examined. Again the English marking appeared to be better than that of the other two subjects researched. According to the marks, each of the same four standard papers (two English, one Mathematics and one History) could be marked as either:

- Being near perfect (mark of 99%), or
- A reasonable pass (marks in the range of 80-90%), or
- A very bad fail, for example in
    - English marks about 50–60%,
    - Mathematics a mark of 28% and
    - History a mark of 43%.

Starch and Elliott's fifth point was concerned with the quality of the paper. In the English study Paper A was pre-determined to be of good quality, while the Paper B was predetermined to be of lesser quality. Results from this study showed that Paper B was on average marked 8 percentage points lower than paper A. This is fine so far. But it must be noted that of the 142 examiners in this study, 23 examiners marked Paper B 15 or more percentage points lower than Paper A. Further, another 19 examiners marked Paper B higher than Paper A. Perhaps by being of a poorer quality, Paper B gave more opportunity for a wider range of marks to be awarded?

A final point was raised by Starch and Elliott in the study of Mathematics. They analysed one of the questions in more detail. This particular question, was Question 10 in the paper. They found that it produced a probable error of around 8.8%, which is a larger probable error than the overall probable error when the Mathematics paper was analysed holistically.

There are other research studies into the inter-marker reliability that help to further explore aspects of the inter-marker problem. Each will be outlined in the following sub-sections.

2.2.2.10 Brown (Brown, Rust and Gibbs, 1994)

In the publication by Brown Rust and Gibbs, there is a summary of an exercise that had been conducted by one of the three authors, Graham Gibbs. This study was undertaken to investigate the marking of two specially constructed essays using a scale of 0 to 10 with the pass mark set to 4. In this study one poor essay and one good essay were marked by several hundred teachers drawn from many subject disciplines. The poor essay, "Essay 1", received a mark in the range of 3 to 7, with an average mark of about 5. The good essay, "Essay 2", received a mark in the range of 5 to 10 with an average of 7. From these results it appears that nothing unexpected was detected.

However some disquieting facts did emerge. Firstly the range of marks awarded per essay was larger than the difference between the essays. Secondly Essay 1 was awarded the equivalent of a first class honours while Essay 2 barely passed, and thirdly Essay 1 was sometimes marked higher than Essay 2. At this point of the current thesis these facts such not be too surprising for the reader.

Although lacking a full statistical analysis, this one inter-marker study clearly shows how difficult manual essay marking is.

2.2.2.11 Byrne

In this study Byrne (1980) investigated the reliability of Open University tutors in their marking of tutor marked assignments, commonly referred to as TMAs. For each Open University course there may be as many as 8 to 12 TMAs per course module. TMAs are intended to provide two pieces of information for the student. Firstly the student receives a mark representing the student's performance, and secondly, the TMA provides a basis for student and tutor dialogue that encompasses feedback and so on.

Assignments were taken from sixteen different second-level courses spanning all six Open University's faculties. Byrne involved 48 tutors to participate in the study. In groups of three, experienced tutors marked the assignments. Each tutor marked four of their own students' work and four from each of the other two tutors in the group. Thus each tutor marked 12 assignments, where the expectation that each assignment would take between 30 to 50 minutes each to mark representing a minimum investment of six hours per tutor. All the marks were on the Open University standard marking scale ranging from 0 to 10.

Byrne reported three sets of results from the 16 possible in his study. The first set was taken from Arts, Social Sciences and Education, the second set taken from Physical Sciences and Technology and the third set was Mathematics. Table 2.2.2.11 below, summarises Byrne's findings, where the number in brackets indicates frequency of occurrence:

	Set 1	Set 1	Set 2	Set 2	Set 3	Set 3
	Mean	Range	Mean	Range	Mean	Range
Marker 1	5.8	4.0	5.3	3.7	7.9	3.6
Marker 2	6.8	4.3	5.6	7.0	7.4	4.6
Marker 3	5.8	7.3	6.2	5.4	7.8	4.1
Range across markers	1.0	3.3	0.9	3.3	0.5	1.0
Maximum discrepancy	3.3 (3)	-	2.3	-	1.5	-
Minimum discrepancy	0.7	-	0.9 (2)	-	0.2	-

Table 2.2.2.11: Summary of Byrne's inter-marker study

This table shows that there is a spread of marks in each of the three sets. In eleven assignments out of the 36 covered by the above table the discrepancy between the three markers was at least 2 marks (out of 10). The study uncovered a difference in marker performance in the various subjects (sets).

The more numerate the subject area, the better the marker performance became. As well as considering the range of marks Byrne considered the ranking of the twelve assignments in each set.

Byrne found that here too was a spread of ranking, in the sense that an assignment awarded the top mark by one marker was not necessarily awarded the top mark by any of the other two markers. The same effect was seen in the case of the bottom mark awarded and was also in evidence with the marks awarded in between these two extremes.

Nevertheless, Byrne examined his results further. From a more searching examination, he found that markers were more lenient with their own tutees and more severe with the tutees from the other two markers. This effect, the so-called "own tutee" / "own tutor" effect, was the more predominant in set 1.

This is a similar effect to that of the "marker's expectation" effect discussed earlier in this chapter (2.2.1.10), but this time it is operating on a sub-group of the cohort being marked. It must be noted that, although similar, the "own tutee" effect is not exactly the same effect as that of "marker's expectation".

2.2.2.12 Coffman (Coffman, 1966; Coffman and Kurfman, 1968)

In 1966 Coffman reported on a study in which he was one of three researchers, the others being Godshalk and Swineford, who conducted a study in 1966 into the relative reliability of multiple-choice marking compared to essay marking. This study involved 646 essayists writing short essays on each of five different topics. Each essay was then marked by five different human markers. This study also included a secondary objective which was to determine the validity of using a single essay.

In essence the marks awarded for each of the five essay topics were compared with each other, and with the multiple-choice marking. These marks were then compared with the results of multiple-choice marking.

In this author's view the most significant part of this 1966 study is Coffman's last sentence, which is quoted below.

> *"In order to obtain validity coefficients of comparable magnitude using only essay tests, it would be necessary to assign at least two topics to each student and have each read by five different readers or to assign three topics and have each read by three different readers."*
>
> Coffman (1966 p. 156)

It should be very clear to the reader that Coffman's suggestion would prove to be too expensive to implement, especially in today's academic environment.

Coffman and Kurfman (1968), conducted a further study, the purpose of which was to compare two different methods of marking essays, namely the holistic method and the analytical method. Both methods were based on a 15-point scoring scale being applied by four different markers on 120 essays. The essays were organised into 15 batches where each batch contained 15 essays. The marking was spread over two days.

The holistic method for marking was to assign a single mark to the essay based on the overall, or holistic, impression of the essay. The analytical method for marking required the marker to assign marks in functional areas, then sum these functional marks to give the final essay mark. The three functional areas were Generalisation and Interpretation; Factual Content; Presentation and Mechanics.

The findings of this second study by Coffman's is that the differences in marks due to the two methods of marking are not significant when viewed in the light of the differences between the markers, and in the light of the differences in marking performance of each of the five markers in the two days of this study. In conclusion Coffman suggests that to reduce the errors in essay marking then the final mark awarded to any essay should be based on the marks produced by several different markers.

2.2.2.13 Eysenck (1939)

Perhaps the answer to Coffman's question on what is the sufficient number of markers necessary to produce a single essay mark which is reliable (and perhaps the answer to why Page employed five human markers in his research) lies in this 1939 study by Eysenck. In this study Eysenck extended a previous study (by Gordon (1924) on the discrimination of various weights) into the area of aesthetics.

Eysenck sought to find out if the validity of judgements increases with the number of judges used in areas where some affective value is to be assigned. As an example of an area in which affective values are made, Eysenck mentions ranking essays.

Using twelve uncoloured pictures Esyenck obtained 900 rankings. Seven hundred of these rankings were used to create the "standard rankings", while the remaining 200 rankings were used as the experimental group. Table 2.2.2.13 below shows how the average inter-ranker correlation increases with the number of pooled rankings. The shaded column in the table contains values Eysenck quotes from the work of Gordon.

Number of pooled rankings	Eysenck	Gordon
1	0.47	0.41
5	0.77	0.68
10	0.86	0.79
20	0.94	0.86
50	0.98	0.94
200	1.00	-

Table 2.2.2.13: Eysenck's correlation of pooled rankings

From this table it appears that five pooled rankings are barely adequate in terms of acceptability, but that 50 pooled rankings is close to perfection. The question still remains on whether it is acceptable to use five rankings.

By changing 'pictures' to 'essays' and 'rankings' to 'marks' then Eysenck's findings could conceivably be applied to essay marking. But the problem arises as to which academic organisation has the resources to employ five markers for every high-stakes essay yet alone 200.

2.2.2.14 Newstead (Newstead and Dennis, 1994)

Newstead and Dennis, realising that marking essays is not too reliable, decided to investigate the marking of six third year Psychology essays produced during a three-hour long examination with a rubric of a choice of three essays from seven. Their investigation made use of fourteen external examiners and seventeen internal examiners.

The external examiners were drawn from a list of known external examiners maintained by the Association of Heads of Psychology Departments. Ten of these examiners were from "old universities" and the remaining four examiners were from former polytechnics. One examiner was female. Newstead and Dennis believed that all four parts of the United Kingdom were represented.

The seventeen internal examiners were drawn from the University of Plymouth's Department of Psychology, but the examiners' areas of expertise were not in the specific subject area of six essays used in the study.

These six essays were chosen so as to span the normal percentage scale. This scale was constructed more or less, to the standard university honours classification where:

- 0 – 39: Fail,
- 40 – 49: Third class honours,
- 50 – 59: Lower Second class honours or 2.2,
- 60 – 69: Upper Second class honours or 2.1, and
- 70 – 100: First class honours

All 31 markers in this study were given the same marking instructions and photocopies of the six essays together with the actual question "Is there a language module in the mind?". The marking instructions were to mark each of the six essays using the above scale and to award marks according to the five aspects of the:

(a) Quality of argument,
(b) Extent, accuracy and relevance of knowledge displayed,
(c) Level of understanding,
(d) Insight, originality and critical evaluation,
(e) Relevance to, and success in answering, the question asked.

Table 2.2.2.14 below shows the results obtained from all the 31 markers, highlighted to differentiate marker type.

	External Examiners				Internal Examiners			
Essay	**Mean**	**Min**	**Max**	**SD**	**Mean**	**Min**	**Max**	**SD**
A	52.6	46	59	3.3	53.9	42	60	5.1
B	55.7	45	62	5.3	53.1	37	62	6.1
C	65.9	55	75	5.5	64.5	50	78	6.7
D	63.0	52	76	6.7	61.4	56	67	4.1
E	61.7	52	80	6.6	60.8	47	68	5.4
F	66.5	50	85	9.2	64.8	55	75	5.8

Key		
	Min	Minimum mark awarded
	Max	Maximum mark awarded
	SD	Standard Deviation

Table 2.2.2.14: Newstead and Dennis' Psychology results

This set of results is worrying, as the standard deviation and the range (maximum mark less minimum mark) of marks is both large for each the six essays in general.

In the case of essay F the marks awarded are particularly disturbing, as both the external examiners and the internal examiners rated this essay as a clear 'First' at best and hovering around a 'Third' or '2.2' at worst! The table shows that the reliability of markers awarded by external examiners is as reliability as that of internal examiners.

Newstead also analysed the five aspects of marks mentioned above so that it could be determined if the spread of marks was due to any particular aspect, and so that it could be determined if any aspects were uniformly marked. In other words did any aspect exhibit a low range of marks with small standard deviation. The result of this showed that no aspect was found to have been marked uniformly. All five aspects were highly inter-correlated for both the external examiners and the internal examiners.

In the discussion section of this study, Newstead and Dennis proposed a raft of ideas on how to improve reliability of marking. These ideas ranged from training the markers (both internal and external) to having nationally set examinations assessed by a team of markers. Another idea proposed was to use more objective assessments in place of essays.

2.2.3 Intra-marker problems

The inherent nature of essay marking, being subjective, gives rise to a potential situation where the marker may award different marks for the same essay at different times (Wiseman 1949, Finlayson 1951, Blok 1985). This problem arises from both the general experience of the marker and from the particular experience of the marker in marking that type and level of essay. In general the less experienced the marker the more likely the marks awarded will fluctuate with time. In fact, consistency in marking will not remain at all constant.

2.2.3.1 Wiseman

In Wiseman's research, he used two groups of markers, one marker being common to both groups of markers. Wiseman's research findings back up Finlayson's research work which is described in the next sub-section.

Year	Marker	Correlation
1943	A	0.76
	B	0.72
	C	0.85
	D	0.79
1948	A	0.86
	E	0.83
	F	0.70
	G	0.60

Table 2.2.3.1: Mark-Remark Correlation by Wiseman

The study undertaken by Wiseman in 1943 produced an average intra-marker correlation of 0.78, while his study of 1948 produced an inter-marker correlation of 0.75. This reveals two items worthy of further consideration.

Firstly, there is a superficial implication that on average all groups of markers will end up with a similarly high average intra-marker correlation. Hesitantly, it must be noted that this intra-marker correlation depends on the selection, the quality and the experience of these markers themselves. This hesitancy can be explained by the actual spread of correlation. In the 1943 study the spread of correlation was 0.13, while in the 1948 study the correlation spread was 0.26 – twice the spread of that obtained in 1943.

Secondly, the results show that Marker A's correlation improved from 0.76 to 0.86 – which represents a whole 0.1 in five years. But the question remains unanswered as to how and why this correlation improved. With the lack of further evidence it does not necessarily mean that there was a genuine improvement in Marker A's performance. It could be that the 1948 markers were a poorer group of markers. Indeed, the improvement could reflect both a genuine improvement in Marker A's performance and that the second group of markers were of poorer quality. There is insufficient information to answer this. Another question arises from this: which, if any, of these markers are correct? High correlation, as in the case of Marker A, does imply high consistency or high precision or high reproducibility. But high correlation, in itself, does not imply high accuracy.

Wiseman set two correlation breakpoints for the threshold of acceptability. For intra-marking a correlation of 0.7 was set as the minimum acceptable value. Any smaller value would strongly indicate that for the particular marker was lacking in sufficient self-consistency. The second breakpoint set was for inter-marker correlation at a value of 0.6. For inter-marker correlation it is not likely that, for a set of self-consistent markers, a value of less than 0.6 is expected.

Both Finlayson's and Wiseman's studies give little, or no, information on the time span separating the original marking from the remarking. Neither researcher provided any information on what these markers were doing between periods of marking. It is possible that the markers were actively marking essays or were marking similar essays. But this is not necessarily the case and the question arises as to whether or not these markers were gaining suitable experience in the interim period.

The results of the studies of both of these researchers show a mark-remark correlation of about 0.76. This statement ignores the exceptionally high correlation result in the Finlayson study.

2.2.3.2 Finlayson

Finlayson (1951) conducted a mark-remark study during which four of his original six markers remarked their previous essays. In order to avoid the effects of previous markers in this study the original marks awarded were not made available to the current markers. Table 2.2.3.2 below shows the results obtained by Finlayson from four markers using two essay sets (X and Y) for the mark-remark Rxx and Ryy.

The table also shows, for comparison (Rxy), all six of the markers inter-essay set correlation:

Marker	Rxx	Ryy	Rxy
1	-	-	0.724
2	0.731	0.766	0.650
3	-	-	0.673
4	0.678	0.636	0.601
5	0.966	0.959	0.798
6	0.887	0.856	0.697
Average	0.816	0.804	0.690

Table 2.2.3.2: Mark-Remark Correlation by Finlayson

Finlayson claimed that these mark-remark results were similar to two previous studies conducted by Wiseman (1949). These two studies were undertaken in 1943 and 1948 some two years before Finlayson. The data obtained from these studies showed that the intra-marker subjectivity was considerably less of a problem than was the inter-marker subjectivity. However, one or two sets of data are not sufficient evidence to deduce that all mark-remark exercises will produce the same, possibly acceptable, results.

Closer examination of Finlayson's findings shows that Marker 5 appears to be the most consistent marker of this set as it is Marker 5 who has the highest correlation in each of the three cases considered. Of course, high correlation does not imply that Marker 5, or indeed any other marker, produces a "true" mark. As a corollary, if high correlation implied high quality marking then Marker 4 is clearly the poorest marker in this set. Plainly this implication is not necessarily true and it is certainly not directly provable.

2.2.3.3 Blok

As part of his 1985 study using 16 markers and 105 essays, Blok also undertook a study in mark-remark correlation. The best mark-remark correlation of 0.815 was produced by marker 16 with the worst correlation of 0.403 produced by marker 3. Both these correlation values are better than the inter-marker correlations produced on the two sessions of marking, but about 0.1 in value.

The second marking session, namely the remark part, of Blok 's study was conducted three months after the first marking session using fresh copies of the 105 essays. The rationale behind what was the choice for a three-month time interval between the two marking sessions is not clear, as Blok did not give any clues to this.

This author wonders what the effect(s) of repeating the remarking at other time intervals, say 6 months and again at 12 months would be.

Using the Kolb cycle example again, the marker may not be able to award exactly the same mark to the same submission regardless of when the re-marking takes place.

2.2.4 Time-consuming features of marking

If a feedback comment such as "in parts your essay there was a little ...", then the marker is under an obligation to guide the essayist to the specific parts of the essay which exhibit the "..." . Hence the marker has to explain the reasoning behind such a comment and thus indicate what improvements have to be made. This creates an opportunity for the essayist and the marker to engage in a discussion of what was written and a discussion of the comments made thereon.

Clearly this feedback is very subjective in nature and slow to administer to the individual essayist. This problem is magnified when dealing with a cohort of essayists and is particularly exacerbated as the size of the cohort increases.

2.2.4.1 The marking process

Ignoring the trivial essay that consists of a few words, most essays are comprised of many words organised into sentences and hence into paragraphs. Thus to mark an essay requires the marker to scan the essay in order to establish the relationship between that essay and the marking scheme. This scanning may be undertaken several times for each essay considered. This repetitive pick'n'mix approach is neither fast nor efficient.

For example, consider the marking of one essayist's attempt to describe Kolb's cyclical model of experiential learning. To do this, even the simplest essay could require a few hundred words with or without any accompanying diagrams. Kolb uses an ordered four stage cyclic process but a valid essay answer may not necessarily start with the same stage as Kolb or start at the same stage as the marker's marking schema. Should, or must, the starting points be different?

Further, any description is subjective so any errors may not be seen clearly. Any subsequent review by a single marker or several different markers, or during a feedback session, may reveal further errors in the description. The problem then arises as to the effect that this discovery has on the mark already awarded.

It leads to the dilemma of whether or not the mark should be adjusted. Therefore the feedback to the 'essayist' is slow, possibly fragmented, and potentially recursive.
Often there is no marking schema provided by the marker. In these situations there is no common basis for marking, and no formally agreed method of allocating marks. And in these situations there exists the probability that very little opportunity exists of retrospectively producing one for second or external markers. Nowadays the occurrence of this problem is reducing, as the awarding of degrees is becoming more predicated on detailed answer schemas.

Often marking schemas are expressed in open descriptor terms. Using open descriptor terms the essay components are usually split into sizeable parts such as, for example: Style 20%, Introduction & Conclusion 20%, Content 55%, and References 5%, where the percentages shown represent the allocation of marks. A level of mark granularity being couched in such vague high level terms does not provide clear measures that would aid the marking process. Lack of clarity is evidenced by questions such as what warrants full marks for style.

Similarly there is no indication as to what must be present in the essay for the awarding of marks for references. This could be interpreted as the use of a specific number of references. On the other hand, it could be interpreted as the citation of specific references. A further interpretation might hang on how well the essayist integrated the references cited into the essay. The author suggests that the last option is probably the option most markers would be most comfortable with. Yet there may be assessments where the essayist is expected to find and use specific references. For example some legal and medical contexts particularly lend themselves to using specific references.

Alternative schema formats may similarly be expressed in vague terms. These vague terms do not aid the first marking process, let alone the second marking process and any requirement for re-marking.

2.2.4.2 Provision of feedback to markers and essayists

Giving feedback to the essayist is important in terms of letting the essayist know the mark or grade awarded and why that particular mark or grade was assigned. The marker should indicate opportunities and strategies for the essayist to improve his or her performance.

Due to the nature of essay marking being subjective then it is highly possible that the marker and the essayist end up discussing the awarding of marks and the feedback proffered as well as the content. In its extreme the latter case may end up with the essayist explaining the content and the marker adjusting the marks awarded in the light of this explanation.

Such a discussion will take time for each essayist. In an environment where a one-to-one feedback protocol has been adopted, feedback is not simultaneous and essayists are potentially forced to waste their time to varying degrees waiting their turn to receive their feedback. In the worst case scenario an essayist may not receive feedback quickly enough to effect immediate improvement in his or her standard of essay and thus may be severely disadvantaged. This scenario is not guaranteed to be even-handed due to the fact that some essayists may receive feedback before the completion date of their next essay, while other essayists in the same cohort may not have received their feedback.

Another aspect of the time-consuming problem with respect to providing feedback is that the larger the essay set the greater the time delay in providing feedback to the essayist. Should an essayist be required to submit several essays both for the same subject or for other subjects in rapid succession then it is possible that the essayist will have submitted another essay, or essays, before receiving feedback from the first. That is, the feedback from one essay will not have been incorporated by the essayist in order to enhance the next essay submitted. Thus the essayist may have negative reinforcement of erroneous concepts, style, tone, voice and so on. This is an unacceptable situation as habitual or entrenched errors of the essayist are very difficult to eradicate.

Markers will default to their own chosen style for lecturing, but should not for marking essays. If markers did use their own marking style then essayists are potentially exposed to contradictory feedback. For example, one marker may suggest that the essayist uses short sentences and long paragraphs, whilst another marker may suggest the use of longer sentences. The essayist is thus left on the horns of a dilemma. What is the essayist supposed to do? The essayist may feel that it is pertinent to target a particular style for each marker.

This leaves the essayist with two questions to be answered. Firstly, is there a correct essay style, and if so, what is it? Secondly, what essay style should be used when submitting work to a new marker?

How does the essayist develop his or her own style, if he or she simply follows a marker's advice. Each marker may also have in their turn 'followed' or have adopted a compromise of their own marker's advice, and so this situation is a recursive one extending back in time.

An important point to ponder is the one on how markers glean their own feedback. On what basis do markers objectively learn what went right and what went wrong in the essay set they have just completed marking? This leads to a further point: how do markers enhance their performance for the next time they conduct marking?

There have been at least two attempts (McDonald and Samson, 1979; Borja and Spader, 1985) to make the marking of essays both more efficient in the sense of quickening the marking process and more effective in the sense of trying to make feedback more objective in nature. Both attempts were similar in approach in that both advocated the use of structured objective feedback when the essay is returned to the essayist.

There has been at least one attempt that combines a feed forward and feedback strategy (Donley, 1978). This was the use of a 26-point checklist.

2.2.4.2.1 McDonald and Samson

McDonald and Sansom (1979), who investigated the use of assignment attachments. These attachments or slips were given to the essayist when their essay is returned. The attachment consisted of a doubled-sided single sheet of paper.

On one side of the attachment were six parts each with a number of four-point itemised scales. There was one part for structure (2), argument (3), style (4), presentation (2), sources (2) and mechanics (4). The number in brackets indicates how many items were in each part, and a total of 17 items were used. A few examples should serve to illustrate the wording of the scales. From the structure part a scale was "*Essay relevant to topic ... Essay has little relevance*", and from the style part a scale was "*Fluent piece of writing ... Clumsily written*". Both these examples show that the wording at the extremes of each scale has to be exactly opposite in meaning for the scales to operate properly.

On the second side of the assignment sheet the upper half of the paper was devoted

- to an explanation of the system,
- a warning to the effect that no correspondence should be made between the location of the ticks on the itemised scale, and,
- the overall mark assigned to the essay and a guide to the overall mark assigned to the essay.

The lower half of the side was left blank to allow the marker to pen any explanation and or any comments felt meaningful.

Some twenty to thirty markers in two Australian universities have used this approach to objective feedback. The subjects covered by these markers ranged from anthropology, history and education through environmental sciences to veterinary studies. A few hundred essayists received assignment attachments to their essays (Billing, 1979).

In a similar vein to Borja and Spader whose work is described in the next sub-section, McDonald and Samson advised caution in the use of highly structured feedback. McDonald and Samson also made the originally blank assignment attachment available to the essayist so that the essayist might use it in the composing process.

By developing structured, written, objective feedback then a number of constructive points can be made. Firstly inter-essayist comparisons are easy to make, without much prompting to do so. Both markers (professionally) and essayists (avidly) will make these inter-essayist comparisons. Secondly essayists become more informed about the manner of how their essays were marked and thirdly markers have their detailed feedback open to wide scrutiny. Markers may find this latter point a little disconcerting!

2.2.4.2.2 Borja and Spader

Borja and Spader (1985), developed a list of 15 feedback codes that could be written in the margins of the essay. The self-same codes had already been made known to the essayists. The list of codes and their meaning is outlined in Table 2.2.4.2.2 below.

Code	Meaning
AMP	Answer missed point of the question
AWK	Awkward phrasing hides your meaning
BIB	Brief answer / inadequately developed / lacks breadth
CAI	Critical analysis and/or evaluation inadequate
CAM	Critical analysis and/or evaluation missing
CON	Contradicts other statement(s)
ERI	Erroneous interpretation of concept / argument / theory
ERR	Erroneous reasoning or inference / lack of coherence with previous statement(s)
FAP	Failed to answer part of the question
NFD	Needs further development
ICC	Inadequate comprehension of concept / argument / theory
IRS	Irrelevant statement(s)
UNC	Unclear statement(s)
UNS	Unsupported statement(s) / a personal opinion with no plausible reasons or evidence given
VAS	Vague statement(s)

Table 2.2.4.2.2: Borja and Spader's Feedback Codes

A number of points may easily be raised from this table.

The first point is that all the codes are for fault correction and there is no code for praise. Secondly, these codes are in such a format as to be easily learned and understood by markers and essayists alike as the codes are mnemonic in nature. Thirdly, these codes are located in the essay directly where they apply, so the essayist receives the feedback precisely where it applies. There is nothing to stop any marker in the future from using a similar list or a larger list of codes in marking essays.

This author would recommend to markers considering this approach to feedback that the code list should include codes for praise, the recognition of good work and encouragement to the essayist. Codes used for the provision of summary or holistic feedback could also be easily devised. Borja and Spader, themselves, highlighted these points as a matter for future work. Importantly Borja and Spader also cautioned that the use of codes to provide the essayist with feedback has to be properly deployed to maximise the benefits all round.

2.2.4.2.3 Donley

Donley (1978) published an article on marking advanced essays. In this article Donley listed a 26-point checklist to be used by the essayist to ensure that their essays are of the highest quality possible.

However, this checklist can be used by the marker to indicate in situ both positive and negative comments for the benefit of the essayist. By marking the position in the essay and using the code letters from the checklist combined with some indication of good or mediocre or poor performance then the essayist should gain objective feedback and some indication on the reasoning behind the actual mark awarded.

The actual wording of the checklist given by Donley in this article may have to be changed to better convey feedback to the essayist.

2.3 Opportunity for change (source of the research problem)

All the subjective effects of manual marking detailed above make a very strong case for the deployment of automated marking systems in general, and for considering automated essay marking in particular, should it ever be proved that automated essay marking be possible.

Accepting that there are substantial multifaceted problems in marking essays for style and content, what can be done about it? The pursuit of a manual marking methodology offers little potential for improvement apart from establishing a need for staff development.

However, there is now wide scale deployment of word processing facilities amongst markers and essayists alike. This is coupled with the ever increasing requirement that essays be submitted as word-processed documents. These two aspects provide an opportunity for harnessing the power of modern computing technology for marking.

The usual reasons for replacing manual procedures with computer based procedures are that computer based procedures offer faster throughput, are more reproducible, can operate un-attended, do not suffer fatigue and subsequent inconsistencies, are systematic and error free. It must be noted that this author harbours grave doubts about the "error free" part of these claims!

Essays produced by word processing facilities are already computer-mediated files.

Computer mediated files are easy to transmit, and to store, and are easy to examine by a computer. Modern computer systems that encompass both hardware and software together with very powerful computer programming languages offer the technical window for at least examining the feasibility of automated marking of essays.

2.4 Automated marking

In the case that automated essay marking were proven and accepted, it is necessary to contemplate the potential benefits which would arise for both the marker and the essayist. There are several perceived benefits.

Firstly, automated marking would allow for the provision of faster marking.

Secondly this marking would be systematic in operation as well as being repeatable and reproducible.

Thirdly the nature of the marking would be objective with subjective aspects reduced to a minimum or possibly absent.

Fourthly were automated marking to be deployed then such a methodology would generate feedback quickly and be able to stand up to both close scrutiny and auditing.

Fifthly with the increasing incidence of plagiarism it would provide a basis for the detection of this form and possibly other forms of academic misconduct.

2.4.1 Specific benefits of automated marking

Thus, from the essayist's perspective the existence of fast marking with its subsequent more timely feedback would facilitate better essay submission and allow the essayist to more easily hone his or her essay writing skills. It would immediately provide an objective basis for the mark awarded as well as providing an objective basis for feedback. For a content-based essay there is the offer of the potential bonus of feedback on style.

Thus, from the marker's perspective the provision of fast marking and timely, objective feedback facilitates better education immediately. There would be focused interaction between the marker and essayist over the feedback given. The time saved from marking could be spent providing more essayist focused feedback, or free up time for the marker to conduct other, often academic related, duties. In addition, cohort analysis would be facilitated.

Plagiarism, and other types of academic misconduct, would be easy identified. Finally any request for re-marking would be much more easily accommodated.

2.4.2 Possible criteria for the acceptance of automated marking

There are two questions that hover over the acceptance of automated marking. These questions are:

- Who has to accept automated essay marking?
- What would make automated essay marking acceptable?

The key stakeholders as perceived by the author are some professional bodies, employers, external markers, academic staff as well as even administrative support staff, and, of course, students, all of whom would all be interested in the deployment of automated essay marking. Each of these stakeholder categories would provide input pertinent to the acceptance of automated marking in general let alone automated essay marking in particular. Any radical change and not especially a radical change in the assessment methodology is, correctly, not to be taken lightly. No one wants to be the first to try anything new, especially if the stakes are high as would be so in the case of degree results.

The most obvious criterion for acceptance would be that automated essay marking has to be "at least as good as manual marking". This naturally requires the use of statistical proofs concerned with the comparison of automated and manual marking. However, as seen earlier in this chapter, manual essay marking should not engender high levels of confidence in the reader. Thus, there is a potentially low threshold for automated marking to exceed. Replacing one poor methodology by another is not progress.

A more rigorous set of criteria would be required, and that set has to be multi-facetted. Randy M Kaplan (1992) suggested a set of six criteria for short textual submissions. Kaplan et al (1998) later suggested a different set of four criteria. The first set was employed in Kaplan's paper of automated content marking, while the second set was employed in his paper of automated style marking. This author's refection on these ten criteria is that both sets in combination are potentially applicable to automated essay marking for style and for content.

The author's interpretation of Kaplan's ten criteria is outlined below. It must be noted that this set, although it may not be complete or correct, will at least serve notice of the range and nature of what factors may have to be included in the criteria for the acceptance of automated essay marking.

2.4.2.1 Ease of creating a scoring schema

One implication of automated essay marking is that the examiner has to create a computer mediated marking schema prior to any marking taking place. This, in itself, has inherent problems. One such problem occurs if the creation of such a schema proves to be difficult. If such a state prevails then systematic errors could be generated and staff would be reluctant to initiate the use of automated marking. Another consideration here is that it could also be difficult to make changes to the schema and hence the provision of feedback could be mired with complexity and uncertainty. The existence of all these factors raise the breakeven point of cost-effectiveness when it comes to the adoption of an automated methodology.

It goes without saying that the schema must permit a high level of detail to be contained within it. A marking schema using "see set book page x" is totally useless. But, there is a trade off between detail and time to develop the schema: time used to create a schema increases directly in line with the level of detail to be included in that schema.

It must be borne in mind that the more straightforward, easy or simple a schema is to create, then the more likely it will be to attract uptake by markers. Once essayists realise that simple straightforward schemas exist and that the markers are basing marks and feedback of this schema then essayists themselves will welcome uptake as well. In this context, making changes to the marking schema would be easy. In order to appraise benefits the marker has to balance work needed to create the automated marking schema before the submission deadline, with work that is needed to manually mark essays after this deadline. The lower this breakpoint then the lower the point at which deployment of automated marking becomes cost-effective.

As with all schemas used for manual marking so with automated marking schemas: the examiner not only has to specify what constitutes a correct answer but also has to specify how marks are to be assigned to this answer.

However, in all automated marking systems it is necessary that all alternative answers be included in the marking schema in an explicit manner. There is no room for human interpretation here. The computer must be "told" explicitly what is required in all possible forms.

For example, consider the ordering of the planets from innermost to outermost. Most people will give the sequence as "*Mercury, Venus, Earth, Mars, Jupiter, Saturn, Uranus, Neptune*" and then "*Pluto*". This is usually strictly correct, except for those years, (for example in 1989 and other years, of course), when due to its strange orbit Pluto shifts inside Neptune's orbit. So, in 1989 the strictly correct answer ends with "... *Uranus, Pluto"* and then "*Neptune*". Were a computer to mark an essay on the ordering of the planets this variation would have to be built in.

The voice used by the essayist is also significant. Some essayists will favour the active voice while others will favour the passive voice. Thus in terms of the voice used alternative expressions will have to be catered for. For example the active voice version "*The cat sat on the mat."* is equivalent to the passive voice "*The mat was sat on by the cat.".*

2.4.2.2 Ability to score on various mark regimes

As in all marking regimes an automated marking regime must cater for an answer being awarded full marks or part marks. The marking schema must offer the capability of a range of marks, not limited to just "1" representing success or "0" representing fail. What an answer is worth is fully dependent on the schema in use. Indeed, an answer in one schema may attract a different mark when used within the context of another schema. Offering part-marks for an incomplete answer could be accepted as a fairer marking regime than one in which an all or nothing marking protocol is applied. In essence, the latter is an approach using finer granularity than the former and consequentially creates a further problem. Indeed it must be accepted that allowing part-marks opens up the question of how exactly these part-marks are to be awarded.

2.4.2.3 Ease of identification of non-scoring elements

In an acceptable situation there should be some indication as to how much of the submission mapped onto the schema and how much of the essay is irrelevant. It is thus important that there should be some indication of what marks were awarded where. So the process of marking should indicate the mapping of marks. This has a threefold purpose. Firstly, it provides some form of quality assurance. Secondly, it provides accountability. And thirdly, it facilitates the provision of feedback.

A submission that has a low match vis-à-vis the marking schema could mean one of four options.

In the first instance it could mean that the essay is intrinsically very poor.

Secondly the essay could be a bad faith submission in which the student has deliberately submitted rubbish. This is not a very likely situation, but it is nevertheless possible and should not be discounted.

Thirdly the essay is written by either someone with limited or ill mastery of the language in which it is submitted, or by someone with a special learning difficulty. In this case a human marker using knowledge of the essayist's background may make, or will make, passive allowances or active allowances for the poor language skills.

Fourthly, and this option is probably the most sensitive aspect of all, is that the essay submitted is superior to that which was expected. This latter situation, although problematic, is nevertheless rewarding for any marker to find.

A brace of examples should serve to indicate what a superior answer is, and the danger therein for the awarding of marks.

Example 1. Suppose an examiner asked "*How far out does the planet Earth orbit the Sun?*" with the expectation of "*about 150M km*" as the correct answer. However, the marker may get the answer in terms of "*a perihelion of 147M km and an aphelion of 152M km*"! Clearly this would be a superior answer – probably from a budding astronomer. This answer is "more correct" than the "correct answer", yet does not strictly match the set marking schema! This is a highly problematic situation for the computer.

Example 2. This example shows the inherent danger of the examinee giving a superior answer and is not limited to just the written word. This author nearly failed an oral examination because he gave a superior answer.

Several years ago the author was undertaking the Bronze Medallion in Life Saving. When asked for all the possible different types of burning that would injure a human being, the author gave an answer along the lines of: high thermal flux, friction, chemical – acid, alkali, organic, radiation – Beta, Gamma, micro-wave, UV / sunburn, radio frequency, and so on. Because the examiner's answer schema had only 'heat', the result would have been a fail because that is the one word the author did not use in his oral answer. The result 'Fail' would simply have arisen because the answer given did not match exactly the "required answer".

Only the intervention of the instructor ensured a pass for the author for this question. This raises the question of what result would the author have received if the answer had been expected in the form an essay, a scenario where there was no possible intervention by an instructor. In this situation a "fail" would not be a true reflection of the essayist's performance.

2.4.2.4 Ease of modification should scoring errors occur

In the normal process of designing and creating an instrument of assessment, the examiner should develop the associated marking schema concurrently. However on reflection, sometimes in the light of comments from peers or second markers, sometimes in the light of comments from one (or more) external examiner and sometimes during the process of actual marking, the examiner may wish to alter the marking schema. This is a natural and necessary process.

With this in mind it is thus important that automated marking procedures offer mechanisms which allow ease of alteration to the original marking schema. It is conceivable that the required change or changes be achieved either by some sort of automatic adaptive procedures or by manual procedures. Changes to the original marking schema may necessitate remarking essays already marked. If this is the case it is as important that any remarking that is thus required must be easily achieved.

2.4.2.5 Consistent, reproducible scoring

It is important that every time an answer or a part answer is encountered across the set of submissions the same mark be awarded without exception. If the marking, or remarking, is repeated then the new mark must be reproduced again without exception. It follows that a mark should remain unchanged if that part of the schema was unaltered for the re-marking.

2.4.2.6 Acceptability of scores or results to human markers

Here lies an echo of the simple criterion mentioned in section 2.4.2, namely "at least as good as manual marking". This is a self-evident and expected criterion. The author would suggest that for automated marking to be acceptable then it must model the human marking methodology and produce similar results.

2.4.2.7 Defensibility

On one plane the mark awarded to each and every essay must be justifiable. Ideally there should be a facility of showing where and how each mark is generated both at the holistic final mark and also at the part and sub-part level.

On a higher plane the actual algorithm has to be defended as an educationally acceptable method of assessment. If the algorithm is in itself not acceptable then it becomes a totally worthless methodology regardless of how accurate and reliable it proves to be.

2.4.2.8 Accuracy

Obviously automated marking must be accurate. If such marking is not accurate then it is worthless and discredited. However, what is meant by accuracy, or more demandingly what level of accuracy achieves a minimum threshold of acceptability? In manual marking the norm of acceptability begins with the assumption that the single human marker is always being accurate. The norm continues with the second marking and the external examining roles working to prove that the first marker was indeed accurate. By setting automated marking as the first marker then the norm still applies. A second marker sets out to prove that the first marking is correct. Here the second marker will always be a human marker.

This author does not, and never will, accept the situation where both the first and second markers are automated markers, leading to marking software package A and marking software package B determining any essay marks in collaboration.

2.4.2.9 Coachability

Whatever the marking algorithm is it should not be coachable. That is the essayist should not be able to write an essay that highly maps onto the marking algorithm but does not map onto the context of the essay topic. For example, suppose that most of the marks came from the number of words in the essays and the percentage of words whose length is greater than five characters. By knowing this, an essayist could deliberately construct an essay that is both excessively wordy, that is long in length, and where short words are replaced by longer words found in dictionaries and Roget's thesaurus.

2.4.2.10 Cost

However spectacular automated essay marking may become it will not become operational effective if it is not cost effective. Costs in this particular paragraph refer to both time costs and money costs. Should the organisational set-up costs, time and money, of an automated marking system be too high then there will very be little uptake by educational organisations. If operational costs and time costs, are too high then there will be very little uptake by academics. For the academic the operational costs would cover the time to set up the marking schema, the time to mark the essays, the time to obtain the marks with corresponding feedback and the performance of the actual marking software.

Further, the human examiner must retain, and exercise, the right and obligation to alter the automated mark or marks in the light of any review process of the marks. Examiners are only too aware that the superior submission may be present, and also that "special circumstances" apply to many of the students being assessed and that these should be factored into the resultant mark.

Rewording the simple criterion stated in 2.4.2 above to read "automated marking must be at least indistinguishable from human marking" conveys a minimal level of acceptability to human examiners.

Suppose that the resources were available to deploy, say, five highly competent human examiners to mark the same set of submissions. If the marks produced by the automated marking system are indistinguishable from the expert human examiners, then what rational reason is there for refusing to accept the automated marks?

Given how automated marking almost completely avoids the problems of subjective marking, and does so in the absence of further problem creation, and given that automated marking offers a superior performance then what are the obstacles to its deployment?

2.5 Historical information on automated essay marking

There are two major approaches to automated essay marking: one for marking content and one for marking style. Each approach has one or more clearly identifiable 'founding source' researcher associated with it.

Page (1966), latterly of Duke University, founded the original approach for automated essay style marking whilst he was at the University of Connecticut.

The original approaches for automated essay content marking are based on natural language processing. Landauer and his co-workers, of the University of Colorado at Boulder, and Kaplan, of Educational Testing Service based at Princeton New Jersey are the two founders of marking content by natural language processing. However it must be noted that the specific approaches adopted by these two latter researchers are different. In addition Allott, and the author, have separately adopted a data structure approach to mark content.

Professors Page, Landauer and Kaplan have each had derivative approaches generated by other researchers.

This author is not so narrowly focused on automated essay marking as to ignore other research work that could have a significant impact on the author's own research work. In this author's opinion the most typical instances of this lateral type of approach to his research project are mentioned in the following paragraphs about Iker and Harway (a team working on the analysis of medical interviews), and Don Foster and David Holmes (both working on the attribution of authorship). Finally in consideration of essay traits there is the research work of Dollard who researched tension in documents and Hiller et al who conducted researched on opinionation, vagueness and specificity-distinctions. Each of the researchers mentioned in this paragraph are further explored in the following paragraphs.

2.5.0.1 Iker and Harway (Harway and Iker, 1964; Iker and Harway, 1965 1967)

This section discusses work was of Howard P Iker and Norman I Harway. This research pair worked at the University of Rochester in the early to mid 1960's where their particular interest was in recognition and analysis of content of unstructured interviews with their own psychiatric patients. By using their own assumption that words that have a common meaning have a high correlation with each other. Where there is a low commonality of meaning then there will be a low level of correlation between the words.

Iker and Harway used a software package called "WORDS". In the fullness of time the ten or so programs that made up the software system WORDS transformed into the software package called "The General Inquirer", which is a package that has been much used over a few decades.

Iker and Harway used certain pre-processing procedures conducted manually in order to improve their software package's performance.

The following is a list of the pre-processing improvement procedures, which they undertook. These procedures are not listed in any particular order. The procedures were to:

- Replace contractions, for example "don't" is expanded into "do not",
- Replace all pronouns by their referents,
- Delete all punctuation,
- Delete certain small words, for example 'the', 'a', 'an', 'were' and so on,
- Delete certain parts of speech,
  such as articles, prepositions, conjunctions and so on,
- Stem all verbs,
  for example 'asks', 'asked', 'asking' would be replaced by the root 'ask',
- 'Thesaurus' away word alternatives,
  that is replacement of similar words by a single word.

It must be clearly noted that all of the above pre-processing procedures were manual. These procedures were implemented at the same time as the interview transcripts were being hand coded onto punch cards. It must be noted that this study dealt with transcripts of interviews and not essays. The transcripts were often passed more than once through this coding – re-coding procedure before the actual computer based analysis was performed.

The manual procedures used in this study may, today be automated and Iker and Harway's procedures may still be relevant in the areas of automated content marking of essays.

2.5.0.2 Don Foster (Crain, 1998)

Foster, with his work in forensic linguistics, was thrust onto the public stage as the result of two events.

The first event was his rapid and accurate identification of the author of the novel *Primary Colors* as being one Joe Klein. In spite of Klein's denial over a period of six months he did eventually admit to the authorship of this novel, proving that Foster was in fact correct.

The second event was in the prosecution of Theodore Kaczynski as the Unabomber. Originally the defence tried to employ Foster's skills to refute the FBI's own textual analysis. However, after reviewing the evidence Foster joined the prosecution.

As a result of these two events, and other activties, Foster now uses his skills to undertake other legal activities.

2.5.0.3 David I Holmes (1998)

Holmes, together with his frequent co-worker Richard S Forsyth, has investigated how provenance or attribution may be assigned by the use of style in a scientific method known as stylometry.

Holmes, Forsyth and many other similar researchers, hold the belief that each author is not conscious of their own particular style. They have found that each author's style has its own quantifiable features or metrics that are distinctive and largely un-alterable. An author could seek to disguise their authorship of a piece of text, but not without conscious major deliberate efforts on the part of the author.

2.5.0.4 Stylometry and Attribution

The current author asks the question as to what sort of metrics may be used in stylometry and attribution. The answer this author puts forward is that basically the same as those suggested by Page, and others, for the automated marking of essay style. These metrics include:

- Word length profile,
- Sentence length profile, in both linear and logarithmic scales,
- Frequencies of specific word types,
- Frequencies of specific words, pairs of words, and so on.

So, there already appears to exist a considerable overlap between the metrics used for automated marking of essay style and the metrics used for the automated attribution of authorship.

But there are other, seemingly more specialist, metrics used in stylometry. These include:

- Hapax Legomena that is words that appear only once per essay,
- Dis Legomena that is words that appear only twice per essay,
- Distribution of the last five syllables of each sentence,
- Classification of syllables into either being 'long' or being 'short,
- The probability of an author's new text containing new words,
  based on Fisher's statistical prediction of finding a new butterfly species,
- The use of certain "marker words",
- The frequency of short words, that is words of two and three letters long,
- The frequency of vowel words, that is words starting with a vowel,

In this author's opinion there is no clear case, or reason, for any of these metrics to be excluded from the pool of potential metrics to be considered for the automated marking of essay style. This is taken from the stance of not being an expert in linguistics let alone an expert in computational linguistics. The bigger the pool of metrics becomes then surely there may be a basis for generating a better performance in the marking of essay style.

In chapter 6 the author has indicated that the detection of plagiarism, and other forms of academic misconduct, may be a possible route for furtherance of the author's research and software. The work of both Foster and Holmes, and no doubt many other researchers would appear to be a creditable starting point in the development of anti-plagiarism procedures.

2.5.0.5 Essay traits

The traits of tension and opinionation, vagueness and specificity-distinctions are determined using the level(s) of usage of certain key words or key phrases in the text under examination. In order for this to be achieved there has to be an underpinning lexicon of the appropriate words and phrases.

2.5.0.5.1 Tension trait

Dollard and Mowrer (1947) developed a new measure of tension called the Discomfort-Relief Quotient (DRQ). The formula for the DRQ is:

DRQ = (Number of Discomfort Words) / (Number of Discomfort and Relief Words)

Dollard and Mowrer were not concerned with the marking of essays so there are no performance figures for any marking. The DRQ was applied to written transcripts of psychotherapy interviews recorded by medical staff with their patients. Strangely Dollard and Mowrer carried out an examination of the reliability of those whose task it was to determine the DRQ.

In fact they plotted various performance curves, just as if essay markers were being examined. Dollard and Mowrer considered three different scoring units. These three scoring units are by word, by clause or thought unit and by sentence. The pattern of DRQ curves produced as a result of using these three different scoring units is very similar in each case.

In addition to these curves, Dollard and Mowrer used the Horn statistic to numerically determine how closely the performance of the DRQ scorers matched, where the smaller the value of the Horn statistic then the greater was the level of coincidence or matching.

It appears that a Dr Daniel Horn, of Harvard University, developed this statistic specifically for this study as it is based on the score obtained on any page of the case notes. The Horn statistic may have a value in determining inter-marker reliability and equally the Horn statistic may be useful in detecting plagiarism.

2.5.0.5.2 Opinionation, vagueness and specificity-distinctions

Hiller of the Night Vision Laboratory, and his co-workers Marcotte and Martin of the University of Connecticut (1969), used the essays that Page had already processed. In fact Page sponsored this particular research work and examined some 256 essays for these three traits of opinionation, vagueness and specificity-distinctions. Hiller et al used lexicons of size 130, 60 and 90 for opinionation, vagueness and specificity-distinctions respectively. The number in brackets is the size of the underpinning lexicon. The results obtained for each trait were correlated with the essay grades already awarded. All the correlations were significant at $p < 0.01$.

Some examples of lexicon entries for each of these three traits are:

Opinionation – "I feel", "in my opinion", 'all', "beyond a doubt" and "height of absurdity".

Vagueness – 'probably', 'many', 'sometimes', 'generally' and 'usually'.

Specificity-distinctions – 'analyse', 'distinction', 'exception', "but here", and "for example".

This author has, as yet, been unable to find the exact contents of each of these three lexicons but would like to do so.

2.5.1 Approach for marking style

The Table of Metrics for Style Marking contained in Appendix A details eight researchers' work, and year of the publication of their work, together with the specific metrics they chose to deploy in style marking. These researchers are listed alphabetically and are Bishop, Christie {this author}, Gajar, Johnson, Larkey, Page, Slotnick and Whalen. The original work by Page deployed a fixed set of 29 metrics (Page, 1966). However by 1996 Page had increased the number of metrics used to the best 50 metrics from a pool of over 290 (Page, 1996). By 1995 Page claimed that he was making use of some 4,000 metrics (DeLoughry, 1995). This is the highest reported number of metrics used.

However as the next volume of metrics used by Page is a pool of 290 this figure of 4,000 is subsequently ignored. The smallest number of metrics deployed is by Johnston who used only six (Johnston, 1996). The average number of metrics used by researchers in this area of research is 22 metrics. Only Page using a dynamic selection of metrics taken from a pool of metrics, while all the other researchers appear to use a fixed set of metrics.

2.5.1.1 Page's approach using Surface Metrics

(Page 1966a 1966b 1968a 1968b 1994 1995a 1995b 1995c 1996a 1996b 1997a 1997b)

Ellis Batten Page is the founder of research into the concept of automated essay marking and the first in the literature to develop appropriate software. The idea probably came sometime during the five years Page was a high school and college English teacher. When Page became a professor of education at the University of Connecticut he started work on realising his idea of essay marking by computer.

Page through the use of suitable statistics showed that it is possible for an essay to be marked reliably by human markers. By this it is meant the production of an inter-marker correlation of 0.8. However he stated that it requires four or so markers to achieve such an acceptable marking performance. The use of four markers for each and every essay is prohibitively too expensive in both time and money.

Page's research was first published in the mid-1960s. In it he developed an approach based on "proxes" and "trins". According to Page a "trin" is an intrinsic factor or variable of interest to the marker that is present in the essay under examination. These intrinsic factors are characteristics such as "aptness of word choice", or "complexity of sentences". A "prox" is a measurable quantity that is a variable of interest in approximating to the intrinsic factor sought. The larger the number of the proxs that are used the more accurately the trins are modelled. It is only Page who espouses the idea of proxs and trins. Of all the researchers who have followed Page's metric approach, very few has used these two terms.

One example of this relationship between 'proxes' and 'trins' is shown when one considers the factor of vocabulary aptness. Using an appropriately recognised word list then a computer can be made to generate the proportion of words of the essay that are found on the word list. A second example is in developing a "prox" for the "trin" of sentence complexity. The proportion of prepositions and subordinating conjunctions could be a candidate "prox" for this particular "trin".

In reality, this one-to-one relationship between "trin" and "prox" was not deployed. Page, his co-workers and those who emulated Page (see later and in Appendix A), developed a procedure in which a sample of the essay set was first of all marked manually. These marked essays were then analysed by computer to determine the value of the metrics chosen by the researchers. Using the values of coefficients for these metrics and the manual marks, a weighted linear function was then developed to regenerate or emulate the manual marks from the metric values. Lastly, this weighted linear function was used to mark all the remaining essays of the essay set. In essence, the manual markers marked on the basis of a parcel of "trins", while the computer marked on the basis of a parcel of metrics.

Page found that when the procedure was carefully conducted, the sample essays carefully selected and the statistics were correct (especially when choosing a set of metrics from a pool of metrics) the automated marking provided a level of performance that equates to experienced markers.

Page in 1966 produced a table of inter-marker correlation and challenged the reader to determine which of the five markers was the computer.

In 1966, and again in 1968, Page published the explicit list of 30 metrics used in his software. Tables 2.5.1.1.a and 2.5.1.1.b shows Page's lists of metrics used in 1966 and 1968 respectively. By 1994 Page's published list of metrics had grown to 34 (with the use of a constant) but this list was not so explicit as his earlier list.

Number	Metric description
1	Title present
2	Average sentence length
3	Number of paragraphs
4	Subject-verb openings
5	Length of essay in words
6	Number of parentheses
7	Number of apostrophes
8	Number of commas
9	Number of periods
10	Number of underlined words
11	Number of dashes
12	Number of colons
13	Number of semicolons
14	Number of quotation marks
15	Number of exclamation marks
16	Number of question marks
17	Number of prepositions
18	Number of connective words
19	Number of spelling errors
20	Number of relative pronouns
21	Number of subordinating conjunctions
22	Number of common words on Dale wordlist
23	Number of sentences end punctuation present
24	Number of declarative sentences, type A
25	Number of declarative sentences, type B
26	Number of hyphens
27	Number of slashes
28	Average word length in letters
29	Standard deviation of word length
30	Standard deviation of sentence length

Table 2.5.1.1.a: Page's 1966 set of metrics

Table 2.5.1.1.a is the list of 30 metrics that Page originally used in his software. On examining this list further a few points worthy of comment arise. Firstly there is no indication of what constitutes a declarative sentence of type A or type B. Although spelling errors are included in this of 30 metrics, grammatical errors are not. One metric, number 1, is binary as a title can only be present or absent.

Metrics 2, 28, 29 and 30 are all calculation. Metrics 2 and 28 refer to average length of sentences and words respectively while metrics 29 and 30 refer to the standard deviation of word and sentence length respectively. All the other metrics, that is 25 in number, are in fact simple counts of various features of the essay.

Thus in the first list the 30 metrics are clearly stated and any reader could use the same list to regenerate the same research findings. This list provides good starting point for any researcher conducting research work in the area of automated essay marking.

However Page's 1968 list is not clear as to what the metrics actually are. In Table 2.5.1.1.b this author has taken the liberty of putting (in round brackets) what a few (three) of the metrics may mean. This author has annotated Table 2.5.1.1.b with words contained in parenthesis to indicate possible interpretations for four of the metrics present (metrics 4, 9, 12 and 30).

Number	Metric desription
1	Constant
2	Fourth root of number of words
3	V(2)
4	Senlen (= sentence length?)
5	V(8)
6	V(10)
7	V(11)
8	V(12)
9	Punct (= punctuation?)
10	W(6)
11	W(7)
12	Prop (= propositions?)
13	W(18)
14	W(22)
15	W(27)
16	W(31)
17	X(1)
18	X(2)*V(1)
19	X(3)
20	Ques
21	X(6)
22	X(7)
23	X(13)
24	X(14)
25	X(19)
26	X(35)
27	X(43)
28	X(49)
29	Y(5)
30	Paren (= parenthesis?)
31	Z(1)
32	Z(2)
33	Z(3)
34	Z(4)
35	Z(8)

Table 2.5.1.1.b: Page's 1968 set of metrics

This author has been unable to find out the specific nature of 29 metrics of those presented in Table 2.5.1.1.b refer to. This author hazards a guess to the meaning of metrics 4, 9, 12 and 30. Metrics 1 and 2 are explicit in meaning.

One may understand the research pressures and potential commercial pressures that have lead to this non-information by Page of the 1968 and succeeding publications.

In the mid 1990's Page made reference (DeLoughry, 1995) to using some 4,000 metrics in his style algorithm. There is not even a glimmer of meaning as to the nature of these 4,000 metrics are. The manner in which the 4,000 metrics is mentioned makes this author think that these metrics would be used as a fixed list. That is all 4,000 metrics would be used in the weighted linear model used for style marking.

Latterly, in 1996, Page's software used a pool of the statistically best metrics from a pool of 290 possible metrics for each and every essay set. Again there is no clear mention of what metrics are included in the set of 4,000 (of 1995) or in the 290 metric pool (of 1966). Page must have reconsidered his approach to metric selection during 1995 and 1996. There appears to be no change in the basic algorithm, only changes in the metrics selected and the method of selecting these metrics. Page must have reviewed his list of 4,000 metrics in the light of which sub-set of metrics appears to have a common role when marking many and varied essay sets. This author surmises that the set of 290 metrics used as a pool in 1996 is, in fact, the most frequently used 290 metrics. Another view of why these 290 metrics were selected may be on the basis of impact in the algorithm for marking style. If a metric continually confers a very low contribution to the weighted linear model then Page may have decided to ignore it use in the future.

In 1966 Page revealed that his software consisted of some 42,000 lines of code and also he revealed the augmentation used with to his software. This augmentation was in the form of:

- General purpose dictionaries,
- 'Taggers' to analyse the part-of-speech of each word,
- Parsers to analyse sentences,
- Special topic related sub-lists, and
- Heuristics.

Page's software actually included the statistical sub-routines required to select the best fitting 50 metrics from the pool of 290 metrics. Therefore there was no need for Page to utilise additional third-party software.

Throughout all four decades spanned by his work in this research area Page has constantly improved his algorithm and his software. At the outset the software was written in Fortran and finally programmed in C++. Page's basic approach has not publicly altered. The manner of submission of the essays has also dramatically changed over the four decades.

Originally both the essays and the software were submitted using punch cards. Currently the mode of essay submission is direct computer entry, via word-processing software.

As well as providing an overall mark, or a holistic mark, Page has maintained for 40 years that the computer may also give marks on the five essay traits that he himself had identified.

These essay traits are Ideas, Organisation, Style, Mechanics and Creativity. The algorithm for marking these five essay traits parallels the same algorithm as for the overall or holistic mark.

Over four decades the work of Page has been doubted and ignored, but never refuted. Both Johnston and Kaplan (and others) actually used the same essay sets that Page, himself, used. In fact Johnston and Kaplan make a point of mentioning that they used the same essays as Page did. However, neither researcher made any comparison between their own particular research findings and that of Page. This author is left wondering why on both occasions such an omission has occurred. This author ponders what would have been the effect on the academic community had these inter-researcher comparisons had been produced and then piblished.

Several other researchers have followed Page's approach to automated marking of essays for style and these researchers' work is now discussed in turn in the following sub-sections.

Several commentators have written up the work of Page and his various co-workers. People like Thomas J DeLoughry (1995) and Debra Viadero (1995) have attempted through their commentaries to publicise automated essay marking in general and more specially Page's Project Essay Grade software (also known as PEG).

2.5.1.2 Bishop (1970)

Robert L Bishop, working at the University of Michigan Ann Arbor, developed software based on Page's approach. In particular, Bishop judged reading difficulty by computing average sentence length and percentage of polysyllabic words. The computation was based on the existing readability formulas developed by Rudolph Flesh, Robert Gunning and Irving Fang.

The software could flag those paragraph or paragraphs and the sentences therein that required modification in order to improve readability. The identification of "*dull, indirect and verbose*" (p. I.8) writing could be achieved by testing for the overuse of articles, the number of adjectives, the number of passive verbs and "*certain mechanical structures*" (p.I.8). Further this software could flag misspelled words, clichés and punctuation faults as well as making suggestion on words that may be required to change.

2.5.1.3 Gajar (1989)

Anna H Gajar a professor at Penn State University Park Pennsylvania, specialises in students with special learning requirements. Gajar had a computer analyse essays from students with and without special needs. However the analysis did require human intervention in the form of specifying thematic units, called T-units. The definition of a T-unit used by Gajar is "*a single independent clause that may or may not be accompanied by one or more dependent clauses*" (p. 125). This definition is attributed to K W Hunt, who developed it in 1997.

By measuring the length of T-units, the number and type of clauses in and the type of clauses in the T-unit Gajar was able to produce a measure of syntactic maturity.

By measuring fluency factors, also referred to as "indices of production", which are partly based on discrete factors and calculated factors, Gajar was able to establish vocabulary based metrics of fluency and diversity. The discrete factors were: number of words, frequency of words, length of words, number of sentences, length of sentences, type of sentences and the number of paragraphs. The calculated metrics used were Herdan's K and Carroll's token ratio.

Herdan's K functions are not defined in this particular paper by Gajar but are said to be independent of essay length and capable of producing figures for richness (or density) and diversity of vocabulary.

Carroll's token ratio is determined dividing the number of different words by the square root of twice the total number of words. In comparison to other similar measures (although not specified by Gajar), Carroll's token ratio is claimed to be less affected by the overall essay length.

2.5.1.4 Johnson (1966)

In 1996 Valen E Johnson, a professor at Duke University, applied the heavyweight statistical technique of Bayesian Analysis to some of the data that had been already used by Professor Page. The data used by Johnson was composed of six markers' results on 496 essays from Page's 1993 data collection. By applying Bayesian Analysis, Johnson found that the computer marking was better than three of the six human markers, and comparable to the remaining three.

The technique underpinning Bayesian Analysis is to determine which metrics represent a particular grade of essay, where this representation is binary in nature, that is either the metric is present or absent. The resultant set of metrics should all have meaningful statistical or probability measures to identify and hence discriminate between the grade. Coefficients are then used to weight the chosen or selected metrics. Often considerable effort with large training essay sets is needed before the selected metrics can be identified.

For the particular essay set used, Johnson developed the following ten metric linear model to determine an essay mark. The equation used is:

Mark awarded to essay = -4.5 + 0.53*(average word length) + 0.22*√(number of words)

+ 4.2*(% commas) + 2.4*(% prepositions) + 0.016*(average sentence length)
- 0.06*(% spelling errors) + 0.30*c_1 – 0.28*c_2 – 0.23*c_3 – 0.041*c_4

Where c_1 c_2 c_3 and c_4 are possible errors generated from the grammar checker that was used in the analysis, and they have neither simple meaning nor any direct meaning.

Only six of the ten metrics are objective measures taken from the essay. Unfortunately Johnson did not compare his statistical results and linear model with the respective results produced by Page for the same data. This author would have welcomed the opportunity to have seen this comparison.

2.5.1.5 Kaplan (1998)

Randy M Kaplan's research into automated marking of essay for style used the same 1,014 essays used by Page that were sourced from Praxis Series. Just like Page, Kaplan used marks based on a six-point grading scheme. These 1,014 essays were augmented by the 300 essays that were marked by additional markers, in such a way as Page did for his research.

Randy Kaplan and his five co-workers Wolff, Burstein, Lu, Rock and Bruce Kaplan, tested five different models for style marking. The five models that were tested included:

- M1 - RightWriter, which is a software grammar checker,
- M2 - The linear counts of words,
- M3 - The fourth root of the word count,
- M4 - Combination of the linear and fourth root of the words as in model 2 and 3, and
- M5 - Combination of models 1, 2, 3.

The M1 to M5 were labels assigned to these five models by Kaplan, and they are used in the Table 2.5.1.5 that follows below.

The choice of using RightWriter arose from earlier research work undertaken by Kaplan, in which four software grammar checkers were tested for accuracy in predicting essay marks. RightWriter performed best of these four grammar checkers. There are 18 metrics arising from the use of the software package RightWriter.

In terms of performance model 5 was not unexpectedly the best and model 1 was the worst.

The term 'unexpectedly' is used because model 5 is a combination of the first three models. The first three models are all discrete measures in themselves (treating the parcel of 18 metrics of model 1 as a single measure) and it is a known that the more measures used to explain a phenomenon then the better the resulting statistical model becomes. The order of performance was model 5, 4, 3, 2 and 1 in that sequence.

Kaplan produced statistics that showed how the five models performed with the percentages for both the exact matches between computer and human marks, together with +/-1, +/-2 and +/-3 grades (marks) different. Table 2.5.1.5 below shows the results that Kaplan obtained.

Model	Exact Match [as %]	+/- 1 [as %]	+/- 2 [as %]	+/- 3 [as %]
M1	56	42	2	0
M2	67	32	1	0
M3	69	30	1	0
M4	69	30	1	0
M5	67	32	1	0

Table 2.5.1.5: Kaplan's five models for style marking

None of the models were more than +/- 2 grades different from the human assigned grades.

For models M2 through M5 there was a 99% match from exact match plus +/- 1 grade. At worst model M1 showed a 98% match arising from the same two categories.

However, Kaplan did not make any comparisons with Page's results. This author would have welcomed such a comparison being made, after all they both used the same essay set.

Initially it is surprising that the fourth root of the word count (model 2 above) modelled the essay scores better than the linear count of the words. In addition to investigating these five models, Kaplan produced a graph of Average Essay Score against Average Essay Length. This graph clearly showed that the longer the essay the better the grade awarded to that essay. The effect of essay length on grade awarded is greatest for small essays and diminishes as the essay becomes longer.

The average essay length to obtain a grade 2 on a six-point scale was 100 words. For a grade 4 the average essay length was 400 words, which is a four-fold increase in length for a doubling in score. To obtain a grade 6 the average essay length rose to 1100 words. In comparing the tripling of the grade score from 2 to 6, the average essay length increased 11-fold from 100 words to 1,100 words. However when one considers the fourth root of the number of words used the following data is produced.

The fourth root of 100, 400 and 1,000 is 3.2, 4.5 and 5.8 respectively. These fourth roots are almost a linear relationship with the corresponding essay grades of 2, 4 and 6. This author wonders if this is just an accidental result arising from the six-point grade used to mark the essays in the first instance.

2.5.1.6 Larkey (1998)

Leah S Larkey conducted a research study in 1998, which developed an effective combination of the main types of algorithm for automated marking of essays for style. In her study Larkey used:

- Text metrics – derived from Page's approach,
- Bayesian independent classifiers – similar to Johnson's approach, and
- k-nearest-neighbours classifiers – echoing Landauer's approach.

Larkey used five different essay sets in her study. Three of the essay sets covered these subject areas, namely Social Studies, Physics and Law.

The two remaining essays sets were labelled "G1" and "G2", and consisted of general questions taken from college examinations for students seeking entrance to graduate study programmes. The Social Studies essay required the essayist to cover certain facts. In the Physics essay the question required the essayist to enumerate and hence discuss the various energy transformations in a given situation. The question in the Law essay required the essayist to produce an evaluation of the legal argument that had been given in the question. The essay set labelled as G1 required the essayist to present a logical argument, whereas the G2 essay set required the essayist to evaluate an argument coupled with a specific scenario. A good essay for the G2 question would be based on expression rather than what the essayist had covered in their essay. For Social Studies, Physics, Law and G1 a good essay had to cover certain points.

In seeking to use both Bayesian classifiers and k-nearest-neighbours Larkey was forced to use large training or tuning essay sets in each of the five different essay sets. Table 2.5.1.6 below shows the size of the training sets, the size of the actual test sets and the number of grade points in each of the five different essay sets.

	Training set size	Test set size	Number of grades
Social Studies	233	50	4
Physics	586	80	4
Law	223	50	7
G1	403	232	6
G2	383	225	6

Table 2.5.1.6: Larkey's essay set size

Each essay in the training set, and for that matter the test set, had already been marked by human markers. For the two essay sets labelled G1 and G2 the marks awarded by two human markers were averaged to produce the manual mark. The other essay sets had an unknown number of human markers.

Although the training set sizes of the five sets of essays would appear to the reader to be sufficiently large, Larkey herself felt that her results could be better if she had used even larger essay sets. It is worthy to note that Larkey used a limited number of grade points, either 4, 6 or 7. In comparison a system that was based on percentages would have 100 grade points. By using a limited number of grades Larkey may have set herself a slightly easier problem to solve.

Larkey had two objectives for her study. The first goal was to evaluate Bayesian or binary classifiers with both the weighted linear model and the k-nearest-neighbours approaches and hence to then build a composite model from all three approaches. The second goal was to improve on the standard method of evaluating inter-marker performance, which made use of Pearson's product-moment correlation.

For the first goal of testing the three approaches for marking essay style some pre-processing of the essays was required. This pre-processing was a two parts process. The first part was the removal, or the purging, of stop-words or function words from the essays. In must be noted that there were some 418 different such words on the stop-word list. The second part of the pre-processing was to stem the words that remained after the deletion of the stop-words.

In the case of identifying potential Bayesian classifiers a stemmed word had to appear in at least three essays. The set of Bayesian classifiers that produced the highest correlation with the manual marks was used to mark the test essays. However simple this appears, in fact, a great deal of work had to be completed in selecting sets of classifiers before the optimal set that produced the highest correlation could be found. Larkey reported using as many as 680 stemmed words (called "features" in the parlance of such statistical methodology) in the Bayesian classifiers.

For the k-nearest-neighbour classification Larkey had to find those essays (this is the "k" in the "k-nearest-neighbour) in the training set that were as close as possible in similarity to the essay being marked. The mark awarded to the test essay was then the average mark of those k essays that were found to be similar to the test essay.
Thus the value of k may be zero, one or many. Larkey used a probabilistic retrieval system to determine the level of similarity between the essays in the training set and the essay being marked.

Larkey used eleven text-based metrics to characterise the essays. The metrics used were similar to those used by Page and were of two types, namely discrete and calculated.

The following eight discrete metrics were used:

- The number of characters,
- The number of words in total,
- The number of different words,
- The number of words longer than 5 characters,
- The number of words longer than 6 characters,
- The number of words longer than 7 characters,
- The number of words longer than 8 characters, and
- The number of sentences.

The remaining three metrics were those that had to be calculated:

- Average word length,
- Average sentence length, and
- Fourth root of the number of words.

The statistical software known as SPSS was used to identify which metrics to select in the linear regression model to mark the test essays for style. At the same time the use of SPSS gave the coefficients for the chosen metrics. Larkey in fact examined three linear regression models, namely using only the eleven text metrics, using only the Bayesian classifiers, and using the combination of all three approaches (text metrics, Bayesian classifiers and the k-nearest-neighbours).

In summary Larkey found the combined linear regression model performed the best of the three models. This result showed that Larkey, in effect, achieved her first goal of validating the use of the Bayesian classifiers as a means of marking essays. Knowing that the combination approach operated just as effectively as two human markers should be regarded as a bonus. Across the five sets of essays used the contribution from each of the different approaches was not constant and in two essay sets, Social Studies and Physics, the contribution from k-neatest-neighbours was in fact nothing.

Considering Larkey's second goal, the problem with using correlation is that it does not specifically inform the reader of just how many essays received the same mark from the different markers.

Indeed a good correlation may hide the fact that not one essay in a set had received the same mark from the different markers. Larkey proposed a scheme, which measured the exact agreement between the different markers.

This proposal ended up with both the exact match and adjacent matches being used, where the term adjacent refers to the grade awarded by one marker that is only +1 or -1 different from the grade awarded by the other marker. The effective correlation from considering exact matches only (about 0.4 to 0.6) was generally doubled when exact and adjacent matches were considered together. In the case of the Social Studies essay set a correlation of 1.0 was reported. Although Larkey did not clearly state what methodology was used in her quest to achieve her second goal, this author strongly suspects that a methodology such as Cohen's Kappa (or a similar or a derivative methodology) was in fact used here. (Cohen, 1960; Cohen, 1968).

2.5.1.7 Slotnick (1972)

In 1972 Henry B Slotnick, while he was Assistant Director at the organisation called the National Assessment of Educational Progress, attempted to produce a theory of how the marking of style could be achieved by computer. To do this Slotnick conducted a principal component analysis of six independent factors, where each factor was analysed at two levels, high and low. The computer marking of essays was used to determine if an essay was to be deemed as "high" or "low" for each of the factors. The six factors were fluency, spelling, diction, sentence structure, punctuation and paragraphing. Objective measures were examined to determine how or why an essay was classified as high or low for each of the factors. The same essays were then examined by human markers. These markers were asked to determine the classification of essays by the same six factors.

For this research work all the 476 essays used were typed into computer-mediated files. In his software, written in Assembler for an IBM 370, Slotnick used 34 metrics. Moreover of the 34 metrics used only 26 appeared once, and only once, in all of the six factors.

Slotnick placed the 26 metrics he used under the six factors as follows:
Fluency ~ 9 metrics, all of which increase in value with increasing essay size.
Number of different words, total number of words, number of declarative sentences, number of (logical, spatial and temporal) words, total number of (logical, spatial and temporal) words.

Spelling ~ 6 metrics, all, naturally, concerned with spelling and independent of other metrics.
Number of misspellings of (difficult, common and uncommon) words,
total number of (difficult, common and uncommon) words.

Diction ~ 2 metrics, reflecting range of vocabulary and quality of choice of words.
Mean of word length and standard deviation of word length.

Sentence Structure ~ 3 metrics.
Number of sentences, mean of sentence length, standard deviation of sentence length.

Punctuation ~ 4 metrics, the number of punctuation marks used: not the correctness of use.
Number of (colons, semicolons, parentheses and quotes) punctuation marks.

Paragraphing ~ 2 metrics
Number of paragraphs, mean paragraph length.

Slotnick noted that the factors of Sentence Structure and Paragraphing were weighted negatively, in other words the bigger the metric(s) value the more the essay mark awarded was depressed. All the other metrics were positively weighted: in other words the bigger the value the better the mark awarded.

However eight metrics of these 34 metrics were not used.
These are:

- Frequency of interrogative sentence,
- Fequency of imperative sentences,
- Standard deviation of paragraph length,
- Number of relative pronouns,
- Number of gerunds,
- Number of past participles, and
- Number of commas.

Is this a feature of the original choice of metrics made by Slotnick? Or, is it a feature of the particular essay set used by Slotnick? If another essay set were to be analysed by Slotnick's software which of the 34 metrics would be used, and what weighting would be applied?

It is asserted that the first doctoral thesis in the area of computer based essay marking was produced by Slotnick in 1971. Slotnick's thesis was entitled "*An Examination of the Computer Grading of Essays*" but was not published.

2.5.1.8 Whalen (1971)

Thomas E Whalen applied multiple regression models to 71 essays and their scores. These essays were sourced from students at Junior High School level (in other words seventh grade students) on their California Language Test. These 71 essays were processed by a software system that was in itself a modification of the software system called "Project Essay Grade" or "PEG" for short. PEG was the software system developed by Professor Page. Whalen developed three regression models: overall writing ability, "mechanical proficiency" and "standardised" language ability.

In these three models 26 metrics were used, of which 17 were metrics that had been used before the remaining 9 were new metrics.

The 17 metrics previously used

The list of 17 metrics Whalen classified as being "previously used" is in two groups: discrete and calculated.

The 13 discrete metrics used were:

- The number of paragraphs,
- The number of parentheses,
- The number of punctuation marks (commas, colons, and semicolons),
- The number of quotation marks,
- The number of questions marks,
- The number of prepositions,
- The number of connective words,
- The number of spelling errors,
- The number of relative pronouns,
- The number of subordinating conjunctions, and
- The number of essay words that were found on the Dale word list.

The 4 calculated metrics used were:

- The average word length,
- The average sentence length,
- The standard deviation of word length, and
- The standard deviation of sentence length.

The 9 "new" metrics

Whalen classified as "new" the following metrics in two groups – 6 metrics in a group labelled "stylistic" and the remaining 3 metrics in a group labelled "mechanics".

The "*stylistic*" metrics were:

- The type-token ratio, which equals
  the number of different words divided by the number of words in total,
- The occurrence of three specific words ("so", "and" & "then"), and
- The occurrences of various forms of the verb "to be".

The "*mechanics*" metrics were:

- The number of capital letters,
- The number of capitialisation errors, and
- The number of usage errors.

Whalen's study showed that the use of the 26 metrics were useful in the production of a style mark using a computer. This study endorsed Page's approach. In fact, Whalen referred to metrics as "proxes" just as Page did. However, in each of the three regression models that Whalen developed the relative weightings, or positions, of the 26 metrics changed dramatically.

In essence, this study of Whalen confirmed that using a computer to produce marks for essay style is as successful as a human marker. This holds provided, of course, that the human marker is well qualified to be a marker of essay style.

2.5.1.9 Burstein & e-rater (1998 2001)

Dr. Jill Burstein, and her various co-workers Educational Testing Services (ETS), developed software that could produce style marks for essays. The software developed is called Electronic Essay Rater (or as it is more commonly called e-rater). More details on e-rater are found at http://www.ets.org/research/erater.html.

The software is applied equally to the training set of essays to build the mathematical model as to the individual essays in the test essay set.

In essence e-rater software package is structured around five modules. The five modules are listed as:

1. Syntactic variety,
2. Organisation of ideas,
3. Vocabulary usage,
4. Mathematical model building, and finally,
5. Assignment of the final mark or score to the test essay.

Syntactic variety is based on the identification of various clause types, such as infinitive, subordinate, complement and so on. To identify clauses e-rater has to assign, or tag, the part-of-speech to each and every word in the essay but with particular emphasise on verbs. The measure of syntactic variety includes various ratios of the different syntactic structures per essay and per sentence.

Organisation of ideas is based on a variety of cue words, such as 'however', and other terms coupled with certain syntactic structure features. By using cue words, cue terms and so on e-rater is able to partition the essay into a series of arguments together with an identification of the type of argument in that partition. From this series of partitions the measure of the volume of ideas and how these ideas are weaved together is generated.

Vocabulary usage is determined by the actual words used by the essayist. After removal of stop words, or function words, the remaining words are processed into what Burstein calls a vector-space model. As will be described in the section 2.5.2.5 this vector-space model is used to determine content marks for essays.

Mathematical model building is achieved by the use of stepwise regression to produce a weighted linear model that takes as input the metric values from the first three modules and the marks from the two human markers.

Assignment of the final mark is produced from the linear model developed in the previous module.

The operation of these last two modules is described later in this section.

In 1998 Burstein used essays produced for the Graduate Management Admissions Test (commonly known as GMAT) and the Test of Written English (commonly known as TWE). See http://www.gmat.org/ for further details on GMAT and http://www.toefl.org/ for further details on TWE.

There were 15 essay sets used, 13 of which came from the GMAT programme and the two of which came from the TWE programme. The TWE essay sets were from essayists who were non-native English speakers. There were two types of GMAT essay sets. Eight of the GMAT essay sets were based on topics requiring the essayist to answer argument focused essay assignments, while the remaining five GMAT essay sets were on issue focused essay assignments.

There were some 640 essays in each of the 15 essay sets. All the essays were marked on a six-point grade scale (the GMAT web site has details of this six-point grade scale). The accuracy claimed for e-rater ranged from 87% to 94% agreement with the human markers. The measure of accuracy used was an exact match or "+/- 1 grade" between the mark awarded by e-rater and the agreed mark awarded by two human markers. The term 'adjacent' is often used for term "+/- 1 grade".

The underlying principle behind e-rater is that for each grade point all the essays awarded that grade point will exhibit very similar features to each other essay.

To operate successfully e-rater requires training on a set of essays that have been previously marked by experienced markers. According to Burstein, the optimal size for the training set is 265 essays that have been marked by at least two experienced human markers. In particular the training set of essays must cover all the six-grades, that is grade 1 to grade 6, of the marking schema. Thus a training set will contain 15 grade 1 essays, and 50 essays for each of the remaining grades 2 through 6 inclusively.

In addition a number of essays, generally five, that would be awarded a zero are also needed as part of the training set. So, effectively a seven-point scale is being used for the e-rater software. The training set is analysed for many features of essays. The purpose of this analysis is to determine which features of an essay map onto each of the six-point (effectively a seven-point with the inclusion of a zero score) grade schema.

There are some 60 metrics used in this analysis. After the analysis a step-size linear regression is applied to the training set to produce a mathematical model for the six grades.

To mark an essay the same metrics are measured. These measurements are matched to the six essays that are the closest in similarity found in the training essay set. Matching in this sense is based on the cosine correlation.

The cosine correlation is used to weight the impact of each of these similar essays' marks into the mark awarded to the test essay. These weighted marks or grades for the six previously marked similar essays are fed into the mathematical model and hence the mark is produced for the essay under consideration.

From February 1999 to 2001 some 750,000 GMAT essays have been scored by e-rater in combination with a single human marker. Using accuracy based on exact and adjacent matching between e-rater marks and human awarded marks, an accuracy of 97% as a minimum is being obtained. Where there is a difference of greater than +/- 1 grade point in the marks awarded then a second human marker is employed to resolve the discrepancy in the marks. The claimed accuracy of 97% between e-rater and human markers is said to be similar to that obtained between two human markers.

At the present time (Autumn 2002) word-processed essays are routinely being marked by e-rater.

2.5.1.10 Writer's Workbench (Reid and Findlay, 2002)

Bell Laboratory developed a software package consisting of fifteen programs called the Writer's Workbench (WWB). This software was used to help document creation in general. By 1984 Kathleen Kiefer and Charles Smith, both of Colorado State University, had adapted this software for use in composition classes.

Stephen Reid and Gilbert Findlay decided that after three years and after some 6,000 students at Colorado State University had been using WWB the time had come to determine if WWB could be used to mark essays.

There was only one essay set used in this study. It was taken from the first-year entrants in their placement examination. All the essays were marked on a 9-point scale. Forty-four essays were taken for this study in order to evenly cover the 9-point scale, these particular essays had been manually marked by a minimum of three independent markers with the marks awarded within +/- one scale point.

The human markers were given a description of what would be expected for each of the nine points of the marking scale. In addition, markers were allowed to award one scale point higher than the matching description provided the essay was written in a fluent stylistic prose.

WWB operates on 27 metrics that are listed in the Table 2.5.1.10. Of these 27 metrics only the first nine (asterisked and shaded in the table) were statistically significant for the essay set.

Number	Metric Description
1*	Essay length
2*	Spelling errors (ignoring repetitions of the same error)
3*	Kincaid readability
4*	Average word length
5*	Percentage of content words
6*	Average sentence length
7*	Percentage of long sentences
8*	Percentage of pronouns
9*	Percentage of short sentences
10	Percentage of abstract words
11	Percentage of the forms of "to be"
12	Percentage of nouns
13	Percentage of normalisations
14	Percentage of vague words
15	Percentage of adjectives
16	Percentage of conjunctions
17	Percentage of compound / complex sentences
18	Type tokens
19	Percentage passive voice
20	Percentage of simple sentences
21	Percentage of adverbs
22	Percentage of complex sentences
23	Hapax Legomena
24	Percentage diction
25	Percentage of compound sentences
26	Percentage of prepositions
27	Percentage of subject openers

Table 2.5.1.10: Writer's Workbench style metrics

At the end of this study Reid and Findlay proposed those WWB programs that could potentially be used to mark essays for (or after modification to WWB software for improved) classroom use. However Reid also gave a warning that care should be exercised in using any style marking software to avoid misdirecting or mis-advising the would-be essayist.

2.5.2 Approaches for marking content

2.5.2.1 Landauer's approach using Latent Semantic Analysis

(Foltz, 1996 1998;Foltz, Laham and Landauer 1999 2001)

The approach taken to marking essay content is basically an extensive co-location pattern of words, called Latent Semantic Analysis (otherwise known by the abbreviation LSA). This technique was applied to the automated content marking of essays in the late 1990's by Thomas K Landauer, Peter W Foltz, and Darrell Laham at the University of Colorado at Boulder. The software is branded as "The Intelligent Essay Assessor" and it is commercially available from http://www.knowledge-technologies.com or http://lsa.colorado.edu.

The original approach was developed by Landauer and his then co-workers back in 1990, to permit the comparison of the semantic similarity between different pieces of textual information.

The principle behind LSA is that there is an assumption that there is an underlining thread in the word usage for a set of similar documents in the same context. This is the latent part of LSA.

LSA generates a matrix of co-occurrences of each word in each document. Since LSA avoids the problems of synonymy (that is different words having same meaning), then LSA facilitates matching between two documents even if there are no words in common. By its own methodology LSA requires considerable training, or tuning, in a new context before any piece of text or an essay can be analysed.

The example cited by Foltz (1996) is on the topic of the Panama Canal. For this specific context three sources of text were required, namely 21 pieces of text of over 6,000 words, several paragraphs taken from three encyclopaedia of a little under 5,000 words in total and excerpts taken from two books of about 17,000 words. So, something in the order of 30,000 words taken from a variety of sources were needed to build the semantic space required for the Panama Canal.

The LSA training created one 100-dimensional semantic space consisting of about 600 text units and about 4,900 unique words. In this semantic space there was a requirement to label each sentence of each piece of text used in the building of the semantic space. Thus there is a serious amount of training required for LSA before any meaningful analysis of any document of essay may be undertaken.

In the analysis part of LSA each and every sentence of the document, or essay, is matched against the semantic model created from the training. This matching process is statistical or mathematical in nature. For a sentence to be identified as being semantically similar to any sentence in the semantic space then the cosine of the angle, or dot product, between the vectors representing these sentences tends to 1.0. A perfect match has the cosine value of 1.0, whereas the value for a perfect no-match is 0.0.

Following Foltz's example cited in the training for the semantic space the target sentence:

"*Only 42 marines were on the USS Nashville*" Foltz, 1996 p.199

was matched with document labelled "MF" sentence "2.1":

"*USS Nashville arrives in Colon Harbor with 42 marines*" Foltz, 1996 p.199

The quoted cosine value for this match was 0.64, which represents a strong match.

When LSA is applied to marking the content of an essay then, obviously, there may be several sentences to be matched. In this situation the mean value of the cosine generated from all the sentences of the essay are used to determine the grade that is awarded to the essay. This is provided, of course, that some amount of calibration of previously marked essays is done to provide the scaling required to reliably award essay marks.

For formative use of LSA the markers may prepare structured feedback in terms of comments or questions. Thus the essayist may be given cues, comments or prompts to improve their essays for re-submission or for future essays. But it must be noted that the markers have to create the feedback for each and every essay topic.

Where they have been encouraged to use submit-resubmit cycles, essayists have been known to use the LSA provided feedback to improve their essay grades (Foltz, Laham and Landauer 1999). In the example cited by Foltz, Laham and Landauer the average mark for the cohort of essayists' first attempt was 85%. After submission-resubmission cycles the final average mark was 92%. The range of mark improvement was from 0% indicating no improvement to 33%, and this was obtained after an average of three submission-resubmission cycles.

The use of Intelligent Essay Assessor (IEA) provides the examiner with further pieces of information to aid the marking process.

The first is that LSA readily detects plagiarism. If a high cosine value were produced between the essay being considered and one of the texts that had been used to train the semantic space then the implications arising from such a good match are all too obvious. It is highly likely that plagiarism has occurred.

Secondly a similar process just as readily would detect where an essayist has based their essay on rote recall.

Thirdly, by using the Intelligent Essay Assessor software based on LSA then those essays that exhibit unusual characteristics may be flagged for human markers to examine and mark.

The characteristics that cause the essay to be flagged for human attention are essays that:

- Show high creativity,
- Have unusual syntax,
- Violate expected essay formats,
- Violate expected structure essay.

The flagging of essays for human marking also heightens the awareness of markers to essayists who may be experiencing difficulties and therefore need to be helped. Difficulties span the range from poor in content to poor in expression.

On the positive side the flagging of essays for manual marking will permit markers to identify and nurture those essayists who write superior essays and to identify and nurture essayists who possess superior intellect.

There have been several reviews on Landauer et al's work on the Intelligent Essay Assessor software. Some reviews are serious (Holmes, 1998 and San Jose Mercury News of 16th April 1998) while at least one has been a somewhat lame attempt at being frivolous or humorous (Computing 30th April 1998 p 164).

2.5.2.2 ETS' approach using natural language processing (Kaplan, 1992)

The original approach that was developed by Randy M Kaplan in 1992 used a pattern-matching procedure developed from natural language processing. This pattern matching was augmented by the use of a grammar to describe the pattern, whilst the grammar was supported by the use of a small context specific dictionary. The initial work was conducted on short answers or simple free text answers of about three to five words long.

Taking a sample of answers such that all possible answers were covered, Kaplan generated a scoring key in the form of a grammar that "describes the language of the responses". This grammar was rule-based and was composed of semantic classes. A semantic class represented some identifiable part, either a single word or a sequence of words, of the answer. Kaplan labelled these identifiable parts as "response elements".

The sample of answers was in effect a training set for the three phased software package.

The first phase was used to generate the semantic classes from the response elements, and was further used to generate a vocabulary list, a lexicon, of unique words.

The second phase took the outputs from the first phase and using the training set of answers generated the scoring key.

In the third and final phase the responses were processed into a format compatible with semantic classes from the first phase and were then scored with the scoring key that was generated in the second phase.

Kaplan used university volunteers as markers to test his software. These volunteers were evenly drawn from both junior and senior undergraduate students. Half the volunteers were majoring in the subject area of social sciences whilst the other half were studying in the humanities, biological sciences and natural sciences. This author is perplexed as to why Kaplan used this particular selection of markers to test his software.

For every possible context, a sample of essays had to be manually marked. This sample was then used to train the software and to generate the required infrastructure of grammar and dictionary or lexicon to support the automated marking.

Then, and only then, was the complete essay set marked. This approach has now cumulated into an operational package called eRater.

2.5.2.3 InQuizit's approach using natural language processing (Weinstein, 1998)

Around 1984 the software used for InQuizit was initially developed by Dr. Kathleen Dahlgren and Prof. Ed Stabler. InQuizit, which was announced about 1998, uses natural language processing techniques to analyse the language used in essays by following the principles of linguistics.

Application of this software was on university level essays in subjects such as Chemistry at UCLA. InQuizit operated at a throughput rate of about 130 essays per minute.

After receiving funding of $10M from various sources (and seeking more funding), the developers were planning to release the InQuizit software in July 1998 for sale at $100 per copy.

However, by July 2002, InQuizit still exists but no longer offers any essay marking software. It offers advanced browser software that appears to be based on the software that was used to mark essays.

2.5.2.4 Allott's approach using data structures

(Allott, Fazackerly and Halstead, 1994a 1994b)

N Allot, with P Fazackerley and P Halstead as co-workers, undertook research into on single sentences. The basis for their algorithm was a projection of natural language processing onto dendritic structure made of simple nodes.

Each node had the characteristics which included:

- an identifier,
- an activation value,
- a threshold value,
- a list of parents,
- a list of children, and
- a list of evidence.

Allott, Fazackerly and Halstead devised three types of nodes. These were the Evidential node, which represented low level knowledge at the level of the word. The Abstract node which represented "*the abstract sense of several nodes*", and the Compound node which represented "*compound sense of several nodes*".

It must be noted the "Abstract" node type equated to a logical OR, while the "Compound" node type equated to a logical AND.

Allott, Fazackerly and Halstead's software requires "training". The knowledge base used to model the required answer was built from pre-marked answers. These researchers put forward a claim that a model that offers 0.85 correlation can be constructed in less than five minutes. Some more time would be required to develop the model further for a correlation of 1.00 to be achieved. In operation a correlation of 0.65 was achieved, with a throughput of over 60 answers in one second.

There are many other researchers, mainly working in the area of automated free text responses, who are using similar approaches to the researchers whose work is discussed above. This author cites the work of Mason and Grove-Stephenson (2002) and Mitchell et al (2002) as examples of other researchers' work in the area of automated marking of free-text responses that parallel his own research work. Mason and Grove-Stephenson presented at the Sixth International Conference on Computer Assisted Assessment, as did Mitchell et al.

For those readers who are interested in computer assisted assessment in general this is an excellent series of conferences. In this conference series many papers have been presented in the area of automated free-text response marking and a few on automated essay marking.

2.5.2.5 ETS' e-rater and c-rater

The following paragraphs explain the use of firstly e-rater and secondly c-rater as described by (Burstein et al, 1996 1998 and Burstein, Leacock and Swartz, 2001.

E-rater

There is use of e-rater to award marks on the basis of the content of the essays. It does so by the application of mathematics to the training set of essays.

The basis for this approach is to realise that an essay is basically a collection of words. By building a word profile from every one of the training essays then a bank of words is obtained where the frequency of words is indicative of the essay topic under consideration. For example if the topic for the essay was on the operation of the human heart, then the word bank produced would be expected to have frequent use of words associated with this topic. In other words the actual words in the word bank together with their frequency of occurrence would be very specific for that domain or context.

Therefore to mark an essay for content becomes a task of matching the word bank representing the training essays against the word bank derived from the essay under consideration. In the word bank produced from all the training essays there will an indication of how to differentiate between good essays and bad essays across a six-point scale of grade categories.

After removing stop words or function words a mathematical formula is applied to determine how each different word is used to facilitate this differentiation of essays into good versus bad essays.

A mathematical formula is used to determine how representative a particular word is in each score category. The formula is:

$$W_{i,s} = (freq_{i,s} \,/\, max\_freq_{s}) * \log(n\_essays_{total}/n\_essays_{i})$$

Where $W_{i,s}$ is the weight for word I in the score category s,

$freq_{i,s}$ is the frequency of word I is score category s,

max_freq_{s} is the frequency of the most frequent word in score category s,

n_essays_{total} is the total number of essays used in the training set,

n_essays_{i} is the number of essays in the training set with the word i.

The process of assigning a mark to an essay involves finding six of the nearest matches from the training essays and averaging the marks previously awarded to those essays. This average is then assigned to the test essay under consideration. To determine a match the cosine correlation is measured between the test essay and the training essays.

C-rater

Probably due to the limitations of using e-rater for marking essay content, Burstein and her co-workers developed the software package called 'c-rater'. This is a (prototype) automated scoring package developed as an offshoot of the automated marking package called e-rater. C-rater, shortened from 'concept-rater', operates on short answer text responses that one would expect to be used in modern on-line assessment packages or in the review formative assessment section of the chapters in better quality textbooks.

The c-rater prototype uses natural language processing to mark the candidate's (essayist's) answers. Burstein et al claims an 80% agreement between c-rater and a human marker. The mark awarded by c-rater is limited to a binary 'credit' or 'no credit'. The award is made on a single correct answer provided by the examiner.

The parentage of c-rater is e-rater and another unnamed (unspecified) software package that was developed in 1996. This unspecified software package was used to classify free text responses where a response was in the range of 15 to 20 words in length, either as a sentence or as a fragment of a sentence.

The awarding of marks was based on the classification of the text against previously marked responses. For this study Burstein developed a domain specific lexicon and grammar that used 200 responses from a set of 378 responses. The lexicon was built from 1-word, 2-word and 3-words terms found in the training set of 200 responses, whilst the grammar was developed from the matching of the lexicon unto structural features of the training set of responses.

The time taken to develop the lexicon manually was two-person weeks. Burstein made an estimate that should this process be automated then it might take 8 to 10 hours to develop the lexicon and a further 8 to 10 hours to complete the pre-processing and post-processing required to generate the grammar rules.

Initially the results were acceptable being 90% for the training set and test set of responses in unison and 81% for the test set only when it came to matching human assigned awards. By improving the lexicon by augmenting the initial set with 'metonyms' the accuracy increased to 96% and 93% for the same categories respectively. A 'metonym' is defined by Burstein as a word that may replace another word when both words have a relation that is specific to that domain.

The impetus for augmenting the lexicon came from Burstein's analysis of the causes of inaccuracies or failures of accuracy. She identified four reasons for inaccuracies. These four reasons, together with each reason's reported percentage contribution to the overall volume of inaccuracy, are listed next. They are

- the lexical gap (40%),
- the human grader misclassification (1%),
- the concept-structure problem (30%), and
- the cross-classification (17%).

The use of metonyms is reported to have reduced the level of inaccuracies due to the lexical gap in the lexicon. However, Burstein does not indicate as to how effective the reduction of lexical gap inaccuracies was. It is worth noting that these four percentages in the list above add up to 88% in total. The source or sources of the remaining 12% of inaccuracies is not further detailed by Burstein.

2.5.3 Inter-marker problem – human versus computer

Table 2.5.3 below serves to summarise the results of inter-marker correlation reported by several researchers. This table includes details (*in Italics*) from the earlier table of human versus human inter-marker correlation (Table 2.2.2 applies here).

Researcher	Number of Markers	Style (S) or Content (C)	Overall Correlation Max	Overall Correlation Min	Computer Correlation Max	Computer Correlation Min
From table 2.2.2						
Cast 1939, 1940	*12*	*S*	*0.779*	*0.009*	-	-
Fajardo 1985	*46+*	*S*	*0.96*	*0.88*	-	-
Finlayson 1951	*6*	*S*	*0.824*	*0.591*	-	-
Kniveton 1996	*2*	*S*	*0.72*	*0.19*	-	-
Jacoby 1909	*6*	*C*	*0.973*	*0.516*	-	-
Nyberg 1980	*6*	*S* *C*	*0.949* *0.952*	*0.592* *0.609*	- -	- -
Wiseman 1949	*4* *4*	*S* *S*	*0.85* *0.73*	*0.72* *0.53*	- -	- -
New Data						
Allott 1994	-	C	-	-	0.77	-
Larkey 1998	2	S	0.88	0.87	0.88	0.86
Landauer 1999	-	C	0.87	-	0.86	-
Page 1966	6	S	0.61	0.44	0.61	0.44
Page 1994	6	S	0.743	0.389	0.743	0.545
Page 1995	6	S	0.778	0.550	0.778	0.716
Page 1997	8	S	-	-	0.926	-

Table 2.5.3: Summary of Inter-Marker Correlation (Automated)

The rows in *Italics* are presented here with the data from human versus computer marking in order that the previous manual inter-marking may help contextualise the new data and to help illustrate the performance of automated marking. The reported inter-marker correlation for automated marking has a range extending from 0.44 to 0.927. This still gives a wide range in performance, but the range for automated marking is very much smaller.

The comparison of like-for-like inter-marker correlation regarding human marking versus automated marking respectively reveals a minimum of 0.009 and 0.44, and a maximum of 0.9 in both cases. This shows that there is no significant difference in maximum reported inter-marker correlation. However, there is considerable difference at the reported low end of the range. This difference ranges from virtually no correlation in human marked essays to a not too unacceptable 0.4 for automated marking.

2.5.3.1 Allott

The research of Allott and co-workers Fazackerley and Halstead has already been discussed in this chapter. By intensive manual operations Allott achieved an excellent correlation value of 0.85 between human and computer marking. The value of 0.85 was achieved by using seen scripts that had been marked and by optimising his software to obtain the highest correlation possible. However the correlation fell to a still respectable value of 0.65 when dealing with unseen scripts.

2.5.3.2 Larkey

In the discussion section of her work, Larkey claimed high correlation occurring in the 0.7's and 0.8s, and that these correlation values are comparable with those correlation values that have been produced by Landauer and Page. However Larkey stated that the performance of the marking depended on the essay type as well as on the quality of the human markers.

2.5.3.3 Landauer

In 1999 Landauer published the research work done by his team and ETS using the Latent Semantic Analysis software, or LSA as it is more commonly known as. Landauer used two sets of ETS sourced essays from the ETS GMAT standardised essay tests.

The first set consisted of 695 essays in which the essayists were required to express their opinion. The second essay set consisted of 668 essays in which the essayists had to develop an argument. For the first essay set the correlation between the two ETS human markers was 0.86, which was exactly the same correlation as one ETS marker and Landauer's LSA. For the second essay set the ETS human marker's correlation was 0.87, which was only very slightly better than one ETS marker correlation of 0.86. It is therefore, safe to assume that in the case of both essay sets marking by the LSA software produced marks which were reliable.

The correlation between LSA and the second ETS marker was not mentioned, but this author assumes that it would be equally as impressive as the results quoted above.

In another LSA experiment, three human markers had an average correlation of 0.73 while the LSA average correlation with the three makers individually was 0.80. Thus the LSA software appeared to perform better than a highly acceptable human performance. Landauer does not reveal how the LSA fared against the three markers when they were considered as a group (as Page does). However in recognising that the treating of human markers as a group enhances the reliability of marking then, perhaps, the group performance was on a par with the LSA performance.

2.5.3.4 Page

Ellis Batten Page (Page, 1966) produced a table of inter-marker correlation across five markers, or judges as he refers to them. In this experiment 138 essays, written by school children (the essayists) in school grade 8 to 12, were marked by five markers. This table is asserted to be Page's first experimental results ever. This author repeats Page's first table of inter-marker correlation below as Table 2.5.3.4.a:

Marker	A	B	C	D	E
A	-	0.51	0.51	0.44	0.57
B	0.51	-	0.53	0.56	0.61
C	0.51	0.53	-	0.48	0.49
D	0.44	0.56	0.48	-	0.59
E	0.57	0.61	0.49	0.59	-

Table 2.5.3.4.a: Page's 1966 inter-marker correlations

The values of correlation ranged from a low of 0.44 to a high of 0.61, which is a slightly lower range of performance than most of the other researchers. When Page wrote this paper in 1966 he invited his readers to determine which of the five markers was the computer. As far as this author knows Page's invitation still stands.

In 1994 Page reported on his 1990 research on the National Assessment of Educational Progress (NAEP) essays. The NAEP programme collected thousands of essays from three years (1984, 1988 and 1990) and for each of these three years three school stages of children (4th, 8th and 12th grade). There were 599 essays in this study, of which 410 essays were used in the training essay set while the remaining 189 essays representing about a third of the complete essay set were used as the test essay.

Page produced a table of inter-marker correlations between six human markers and the computer. These correlations are shown in Table 2.5.3.4.b is presented below:

Marker	Computer	J-1	J-2	J-3	J-4	J-5
J-1	0.728	-	-	-	-	-
J-2	0.664	0.487	-	-	-	-
J-3	0.743	0.605	0.519	-	-	-
J-4	0.564	0.530	0.607	0.648	-	-
J-5	0.545	0.516	0.465	0.487	0.460	-
J-6	0.712	0.581	0.505	0.508	0.389	0.488

Table 2.5.3.4.b: Page's 1994 inter-marker correlations

The mean computer correlation was 0.659, while the human marker mean correlation was 0.545. So the computer marker performed significantly better than the human markers. In comparing Page's first and second tables the computer marking was clearly superior in performance to the human marking. There appeared to be little difference in mean human performance between the two tables.

Page, himself, undertook statistical calculations that showed that his software out-performed three human markers when they were considered as a one group.

In 1995 Page reported the results when he had completed his first blind test of essay marking for style. He, and co-worker Nancy S Petersen (1995), took 1,314 essays that were collected from 33 North American states as part of the 1994 Praxis Series: Professional Assessment for Beginning Teachers. Educational Testing Services (more usually referred as ETS) sent the marks from two human markers with the essays.

ETS assigned 1,014 essays for Page and Peterson to use as their training essay set with the remainder to be used as the test essay set. Page did not know the ETS essay marks. Since ETS accepted that human markers are not reliable ETS arranged for 300 essays of the training set to be further marked by an additional four markers.

The results of the inter-marker correlation for the six human markers and the computer are shown in the Table 2.5.4.4.c below:

Marker	Computer	J-1	J-2	J-3	J-4	J-5
J-1	0.732	-	-	-	-	-
J-2	0.778	0.649	-	-	-	-
J-3	0.740	0.748	0.585	-	-	-
J-4	0.748	0.705	0.684	0.674	-	-
J-5	0.737	0.596	0.656	0.643	0.666	-
J-6	0.716	0.550	0.668	0.594	0.649	0.635

Table 2.5.3.4.c: Page's 1995 inter-marker correlations

This time the mean computer correlation was 0.742, while the human marker mean correlation was 0.646. Although both correlation values had improved, yet again the computer marking performed better human markers. Again, Page undertook the statistical calculations that revealed once more that the performance of computer-based marking was superior to three human markers taken as one group.

By 1997 Page was reporting an inter-marker correlation of 0.926 with a group of eight human markers. Given that 1.000 represents perfect correlation and essay marking is not an exact science then a correlation value of 0.926 generated by computer style marking must be considered as near perfect.

2.6 Proving the robustness of automated marking

This may be seen in a quartet of very contrasting papers. These papers range from marking style to marking content. One is a 'debate', while one reports on a deliberate attempt to deceive the marking software.

The first two papers were authored by Donald E Powers (Powers et al, 2000 2002), and the same set of co-workers namely Burstein, Chodorow, Fowles, and Kukich at Educational Testing Services (ETS), probably engender in their readers the most revealing insight to automated essay marking. Both papers extensively make use of the software package called e-rater. This should not be too surprising as Powers and his co-workers are those who are behind the development of e-rater itself.

The first paper (Powers et al, 2000) set out to examine the validity of using automated essay marking against essay marking by human means. The bulk of this particular paper is concerned with the restatement of the performance of e-rater against human markers across various aspects of essay marking. In summary the effective performance of e-rater is no less than the effective performance of human markers.

The one major feature of performance that is quoted is the accuracy of marking. On the six-point grading scale that appears to be traditionally used in the schools in the United States of America, accuracy is defined as being either:

- The exact grade match only, or,
- The exact grade match together with a +/- 1 grade difference.

The +/-1 grade difference is referred to as an "adjacent grade". Using the second definition of "exact plus adjacent grades" then figures of 97% or better for agreement are often cited, serving to indicate that there is very little significant difference between the two marking methodologies.

Other measures of performance such as the standard deviation of grades and the mean grade obtained from using both the marking methodologies are comparable and they also engender the assumption, or the belief, that there is little significant difference between e-rater marking and human marking.

Where non-agreements, or inaccuracies, most occur is at the extreme ends of the six-point grade scale, that is the low end (the grades 1 and 2) and at the high end (the grades 5 and 6). At the low end e-rater tends to give award slightly higher grades while at the high end e-rater tends to award slightly lower grades.

At the extreme ends of the scale it could be that some, as yet undetermined, subjective factor or factors are affecting the manual judgements being made. Or it may be that some higher sentient factors must be present in the human markers that are not being modelled in the e-rater software that causes the non-agreement. These subjective and / or sentient factors are tending to make human markers depress a low grade and inflate a high grade.

The use of a six-point grade scale for marking essays does make it statistically easier to achieve an acceptable performance than does the United Kingdom's traditional 100-point scale, that is otherwise known as percentages. In the United Kingdom the awards made to candidates are often expressed in terms of a grade or a band such as A, B, or C and so on. Nevertheless very frequently a mark based on percentage is produced by the marker that is then reported as that percentage mark's equivalent grade or band.

It should be borne in mind that e-rater, just like any other software, does not understand text and cannot fathom any meaning from the text. Software only 'reads' text at the character level not at any higher cognitive level.

For this author there are two significant aspects arising from this paper. These aspects are first, in the acceptance that neither marking methodology is absolutely correct and second, is what Powers et al envisages as the route ahead for the next stages in the development of the e-rater software to improve it performance.

In regard to the first of these two aspects, the manual marking of essays has never been absolutely accurate. Therefore the basis of the calibration of the e-rater software is in itself flawed. E-rater has to be trained on a set of previously marked essays before any test essays may be marked, and it has to be trained for each and every different essay topic. Only two types of essay theme are tested by e-rater, that of "analysing an argument" or "discussing an issue".

Powers et al make the recommendation that if e-rater is used for large-scale operational deployment then e-rater should be used in partnership with at least one human marker. That is each essay being assessed is in fact assessed by e-rater and a human marker before an agreed grade is awarded to that essay.

This author reminds the reader of the definition of an essay that is attributed to Stalnaker that was presented in section 1.1 of this thesis. This definition clearly sets out the fact that there is no clearly right essay.

Therefore there is no "gold standard", or 'model', essay to mark another essay against. And who is sufficiently competent among us to produce a non-trivial essay that no one may find fault with, or have suggestions on how to improve the essay?

In regard to the second of these two aspects, the future developments in the e-rater software, Powers et al list a series of areas for potential development. This list mentions the following areas such as:

- Increasing the range and type of essay features used,
- Identification of common or general features to make e-rater universally applicable,
- Consideration of non-linear regression and tree-based regression to build the model,
- Consideration of using neural nets, and so on.

The second paper (Powers et al, 2002) is a totally different to the first paper as it is concerned with deliberate attempts to discredit e-rater. People of various levels of expertise in (computational) linguistics were invited to fool, dupe, or to deceive e-rater into awarding an essay higher or lower grades than it deserved to have received.

Two expert human markers from Educational Testing Services (ETS) were deployed to manually mark each essay in order to create an agreed baseline to determine how successful the attempted deception of e-rater was.

Since the measure of e-rater success, or accuracy, is its agreement with human markers, where such agreement is the combination of exact match in the grades awarded to essays or adjacent (that is +/-1 grade) grades awarded to essays. Therefore successful deception of e-rater means that these invited essayists had a target of being at least +/-2 grades different from the agreed mark produced by the human markers.

Four essay topics were selected by ETS for this experiment, of which two were from the category of analysing an argument and two were from the category of discussing an issue. These four essay topics were chosen from all the available ETS essay topics as being the most representative of the two categories. The customary six-point grading schema was to be used in this experiment.

27 people accepted the invitation to participate in the experiment to discredit e-rater. These 27 essayists range from a few professors to a sizeable number of undergraduate students and graduates. Although none were paid, each essayist was promised feedback on how successful their attempted deception was. There were two monetary prizes of 250 US Dollars for those managing to deceive e-rater the most.

Each participant was randomly given two essay topics from the four selected for this experiment, one from each category. This is the same as real essayists receive when they normally undertake ETS tests, except that the essay topics for real essayists are taken from a much larger pool of essay topics, not just four.

However these participants differ greatly from the usual essayists that ETS marks. The usual essayists for ETS assessment are those mid-point in their academic career. Often the essayists are non-native English speakers. These participants were treated differently from the usual ETS essayists as these 27 participants were given the following information, which are denied to the normal essayists:

- A detailed description of how e-rater operates,
- A list of the particular cue words that e-rater operates on,
- Copies of the scoring guides for the essay topics,
- Sample marked essays for the same essay topics,
- Allowed unrestricted time to create their essays.

This additional information is greatly advantageous for these participants in their challenge to deceive e-rater.

Each participant was tasked to create four essays on a basis of a pair of essays for each of the two essay topics they were given by ETS, although not all the participants did submit their four essays. In each of the pair of essays per topic each participant created an essay that intended to attract a better than deserved mark, and created an essay that intended to attract a lower than deserved mark. For each of the four essays the participants submitted they had to give notice of their reason(s), or rationale, for how they thought that their essays would deceive the e-rater package.

The two manual markers were made aware that these essays were deliberately written for the experiment, but they were not aware of the essayist's intentions. The average manual grade was compared with the e-rater grade and the discrepancy in grades awarded, if any, was determined.

The 27 participants submitted 63 essays, a number somewhat short of the 108 essays if each participant had submitted four essays each. The performance of the manual marking was that for 52% of the essays, the grades awarded were in exact agreement. This level of agreement rose to 92% when exact grade agreement was augmented by adjacent (+/-1 grade) agreement. The mean scores for the two human markers was 3.22 and 3.26 and standard deviations were 1.45 and 1.54. The correlation between the two markers was 0.82. These four statistics namely percentage agreement, mean grade, standard deviation and correlation for this experiment are in the range of values that is normally produced from pairs of experienced ETS markers. So, it is reasonable to assume that the manual marking is reliable and represents a safe baseline to determine the extent the essayists were able to deceive e-rater.

In this experiment the agreement performance of e-rater fell to 65% from its usually high level of agreement of about 95%, when agreement for exact plus adjacent grades are taken together. The e-rater software produced grades that were in exact agreement with the average manual grades in 34% of the essays. This level of agreement rose to 65% when the exact plus adjacent (that is +/-1 grade) were taken together. The normal correlation between e-rater and manual marking is normally 0.80, but in this study the correlation halved to about 0.40. Therefore the participants had been successful in duping, or deceiving, e-rater into awarding incorrect grades to a significant number of their 63 essays.

Of the 63 essays received for this study only 54 essays had their intentions for duping declared. 30 essays were intended to be awarded better grades than they deserved and 24 essays were intended to be poorer than deserved. The remaining nine essays had no intention declared.

Therefore how effective was the duping?
E-rater awarded 39 essays higher grades than the agreed manual grade, 14 were awarded less and the remaining one essay received the same grades and hence there was no difference in e-rater and agreed manual grade.

Of the 54 essays that had been declared to attempt duping e-rater into awarding grades that were higher or lower than deserved, only 36 managed to dupe e-rater in the direction intended by the essayist. Nine essays duped e-rater in the *opposite direction* to what was intended by the essayist and the remaining nine essays failed by default as e-rater awarded the same grade as the agreed manual mark.

The table below shows how effective the various essayists were in duping e-rater. The shaded parts of this table show differences that are greater than +/1 grade.

	Duping Intention	Grade difference	Number of Essays
63 Essays	30 ~ higher	3+	3
		2+	5
		1 – 1.5+	12
		0.5+	6
		0	4
	24 ~ lower	2-	3
		1-	3
		0.5-	4
		0	14
	9 ~ none given		

Table 2.6: Results from attempting to dupe e-rater

From Table 2.6 out of the 30 essays with the declared intention of duping e-rater into awarding higher than deserved grades, only 26 achieved their intention. Only eight essays (25%) were clearly successful in their duping attempt, that is at least two grades higher.

Again from Table 2.6 out of the 24 essays with the declared intention of duping e-rater into awarding lower than deserved grades, only ten achieved their intention. Only three essays (~12%) were clearly successful in their duping attempt, that is as least two grades lower.

It would appear that for this set of 27 experts in linguistics it is about three times as easy to construct an essay that will be awarded at least +1 grade than it is to construct an essay that will be awarded –1 grade. This does not infer that is was easy to dupe e-rater. Given that the participants are linguistic experts, and that they had been given considerable information concerning e-rater and its mechanics as well as unlimited essay creation time to construct their essays into deceiving e-rater they only managed to effect a weighted average duping of e-rater of +/-1 grade.

In spite of the advantageous circumstances of the participants, this is not the achievement of a large-scale, or whole-scale, duping of e-rater. Only the leading linguistic experts managed to produce the winning deception strategies.

Under normal circumstances that ETS operate under an essayist would stand little chance of deliberately duping e-rater in its current configuration. Accidental duping by, for example the creation of a superior essay, will always be a problem. Therefore the deliberate duping, or deceiving, of e-rater into awarding a false grade is not so easy as the critics of automated essay marking have believed. Nevertheless where ETS employs e-rater its operational policy is to have each essay marked by at least one human marker, thus further preventing the deliberate or the accidental duping of e-rater. This author presumes that manual marking will always out-rank (but not necessarily out-perform) automated marking in the ETS organisation whenever and wherever grade discrepancies occur from using the different marking methodologies.

From this highly orchestrated attempt at deliberate deception there is a two-fold payoff for ETS and the e-rater software package.

The first payoff is at an operational level in knowing that e-rater can be deceived into awarding the wrong marks for an essay, and knowing how e-rater was in fact deceived. Knowing how the deception was achieved will identify the changes in the software that have to be incorporated into the e-rater package in order to make it a more robust marking methodology.

In other words some of the weaknesses in the current e-rater package will have been highlighted by this experiment. At the same time e-rater could be re-tested with the same deception essays to check that the required robustness has been achieved after the specific changes in software have been completed.

The second payoff occurs at a higher level, or strategic level, for both ETS and e-rater in that when the deception caused by these specially invited essays have been defeated, or negated, then automatically e-rater will be shown to operate more robustly. Thus some of the sources of criticism of automated marking of essays would be silenced, especially where that criticism is directed towards the fact that automated assessment of essays in not safe. "Not safe" in the sense that the marking software may not award the correct marks for essays and that an essay may be so constructed as to deliberately be awarded the wrong mark, especially if it is higher than deserved.

Essayists will have more confidence that e-rater has awarded their essays the correct grades, especially when they realise that e-rater grades are all checked manually at least once.

The author now wishes to draw the reader's attention to the review work conducted by Wresch (1993), and Hearst (2000), the remaining two papers of the quartet mentioned at the start of this section.

Wresch (1993) of the Department of mathematics and Computing, Stevens Point, University of Wisconsin, produced a paper in which he reviewed the last 25 years or so of automated essay marking. Most of this paper is based on Page's work, and how it was received by the academic environment at large. In fact Wresch based the title for this paper after the title of Page's first paper on this subject written for *Phi Delta Kappan*. Wresch also included some alternative approaches of other researchers and evaluated how these other approaches fared. Wresch states that the research done by Page did produce acceptable results in terms of performance. The followers of Page's approach such as Slotnick were also credited by Wresch as producing acceptable performance.

The other approaches that Wresch commented upon are listed here:
1988 - Finn – Standard Frequency Index (SFI),
1990 - Hellwig – Semantic Differential Scale (SDS),
1990 - Reed – Use of Writer's Helper,
1992 - McCurry – Alaska Assessment Project.
Each of these four approaches is now briefly outlined.

Finn's approach was based on the belief that better students used a bigger vocabulary and as the students progressed in their studies all would grow their vocabularies. So Finn used an approach using words counts derived from 1,000 textbooks used in America in 1969. A word that appeared once in every ten words, for example 'the' would have a SFI of 90, while a word with an SFI of 80, such as the word 'is', appeared once in every 100 words. Wresch does not quote any performance figures for Finn's approach.

Hellwig's approach is based on a scale that is based on his 'feel' of 1,000 commonly used words expressed in terms of 'potency' and 'evaluation' on a scale he calls the Semantic Differential Scale (SDS). What is meant by the two terms 'potency' and 'evaluation' is not clearly explained by Wresch's summary.

Hellwig devised a marking formula using the number of words in the essay less the SDS. Wresch quotes the performance of Hellwig's methodology as being 74% agreement between the computer and manual marks.

Reed's research study was to quantify what improvement there would be in essayists using software called Writer's Helper as compared to just using word-processing packages, especially when revising their essays. Wresch reports that Reed found that essayists that used Writer's Helper software scored 5.5 (on average on a six-point grade scale) while those essayists who did not scored 3.9. Thus by providing formative feedback information on essay style does significantly improve the essayist's performance.

N McCurry and *A McCurry* as project administrators working with a statistician and a software developer conducted the Alaska Assessment Project to provide information to the United States Department of Education on how it was improving the English education of native Athabascans in Alaska. They conducted multiple regression on those metrics the software had been programmed to identify with the manually assigned scores. Thus this study identified 24 Page-like metrics, with a correlation of 0.96 between the computer and manual scores.

While Wresch positively notes the success of the various researchers, he does include a few negative notes from other researchers, namely Macrorie in 1969 and Hellwig's work.

Wresch notes that Macrorie's work summed up the opposition to automated essay marking. In essence Macrorie states that essays are too complex to be marked by computer and that making corrections and assigning marks are not the best approaches in marking essays. Wresch wryly notes that Macrorie's remarks both eliminate the use of the computer marking and also eliminates human marking!

Wresch notes that although the approach used by Hellwig appeared to work it only worked for one set of business reports. McCurry's attempts to replicate Hellwig's work in the Alaska Assessment Project fail according to Wresch.

In his paper Wresch includes a section on "Grader Education". In this section Wresch examines the work of Page and Slotnick in terms of how their work commented on the performance of manual markers. Wresch suggests that Page's research commented on what the markers saw in the essays and not what was actually in the essays, while Slotnick's research work suggests that norms or standards could be set for student essays. Wresch notes that Slotnick does not propose what would be a standard for any student essay. At the start of this section on "Grader Education" Wresch echoes the research work done by Don Coombs in 1969 in which Coombs questions the focus of the research work in automated essay marking and suggests that the focus is really research on the experienced manual markers' cognitive processes.

Towards the end of his review of automated essay marking Wresch makes a suggestion that the marking software may equally both evaluate the essayists and evaluate the markers. The marking software may have a role in standardising or normalising manually produced marks!

The last paper of the quartet is a debate organised by Matti A Hearst of the University of California, Berkeley. This debate was hosted by IEEE Intelligent Systems as part of its "Trends & Controversies" series (Hearst, 2000). This debate was structured into four parts:

- Beyond Automated Essay Scoring – Kukich, ETS
- The Intelligent Essay Assessor – Landauer, Knowledge Analysis Technologies
- Automated Grading of Short-Answer Tests – Hirschman, The MITRE Corporation
- To Grade or Not to Grade – Calfee, University of California, Riverside

The first three parts are written from the standpoint of the writer's organisation. The last part of the debate is in effect a commentary on the first three parts.

This author feels that this is not a true debate as it allows three different organisations, possibly even in competition with each other, to "advertise their wares". The author would welcome a true, or real, debate in which both the supporters and detractors of automated essay marking, regardless of what the purpose of such marking is for, are encouraged to express their viewpoints. Such a debate, especially with the opportunity to answer these viewpoints, may focus on the key elements of the question of automated essay marking. The identification of these key elements will define the future of automated essay marking, and may clearly expose the fears and the worries (and perhaps the prejudices) of the detractors. By knowing what are the detractors' key points the supporters then have clear targets for future marking software developments.

Finally in this section this author summarises a balanced review of grading software produced by Thomas Hamel (1988), a Professor of English at Black Hawk College.

Hamel published a paper in which he sets out to give a user's view of automated marking. Over a period of 18 months Hamel tried several pieces of grading software, including one that he designed himself. Software ranged from spreadsheets to databases. Hamel found the various software packages to be objective, helpful, efficient and slightly more complicated and more time-consuming than using paper. He also found that grading software '*ruthlessly*' (Hamel's own word) exposed any prejudices that he may have harboured towards (some) students.

Most of the software required similar information, such as course information, lists of students, the marks given to students and so on.

Hamel identified problems in using software for grading, some of which were related to computer operations: access to the computer, starting the computer, loading programmes and data, making back-ups, printing reports and so on. There is also mention of problems of compatibility and portability. In 2002 the problems of compatibility and portability are very much reduced than in 1988 when Hamel published his paper.

However other problems that were identified related to the effect on the marker in using the software. These are listed below.

- Changing the method of how he marked essays to suit the technology,
- Synthetic precision of a number in place of an elastic A B C with attendant + and -,
- Loss of the "fudge factor" such as "class participation".

Hamel balanced these problems with the benefits of using grading software, such as:

- The realisation of just how subjective other marking methodologies can be,
- Students welcomed the objectivity in numeric grades,
- The marking process became less obscure to the students,
- Students benefit from more objective feedback,
- Taking more care in the assigning of a mark.

The essence of Hamel's review is best shown in his last paragraph, which is quoted below. He does not make any opinion but gives strong incentive to think about the use of grading software.

> *"Like many computer operations, grading programs oblige us to think in a different way. Used with care, however, they offer significant advantages, especially for students."*
> Hamel, 1988, p. B 3

Chapter 3: Research Methodology

3.1 Basic approach

This chapter seeks to set out what the author has gained from the literature survey, moves on to how the algorithms for style and content operate and finishes up with short descriptions of those statistical measures that are appropriate to this research work.

An extensive literature survey was conducted to determine the extant methodologies for automated essay marking. The literature survey however, produced very little information on automated essay marking in total and what was found is almost exclusively on style marking. Further the style marking information given in the literature tends to be somewhat short in detail, but long on vagueness. The survey will remain ongoing after this particular PhD research project is finished since the author intends to carry on research work in this field. The survey however did uncover information about the associated performance in terms of accuracy, acceptability, and usability of these methodologies. Where possible the rationale and history for these methodologies were also determined.

The results of the initial literature survey were used to define the starting point for developing methodologies for both Style and Content.

The development of the author's algorithms for Style and Content has followed a quasi-Newtonian / quasi-Heuristic development.

The Newtonian approach is to examine the data, or facts, and hence to devise a theory to explain these facts. By means of a series of experiments the theory is then tested and the results are then reviewed in order to improve the theory. The cycle of experimentation and review repeats until no further improvement to the theory is possible.

The Heuristic approach complements the Newtonian approach in this research project. Heuristics is a system of rules used to explain, or to develop a pattern of behaviour. By experience of a system a simple set of rules is devised to account for the system's behaviour. These rules are revised as the experience of the system grows. As the experience grows the volume of the rules and the complexity of the rules increase.

This bifurcated approach was decided on by the author, as initially there are few facts and few rules to work with. As the algorithms used in this research project were being developed the author had to devise rules of increasing complexity where these rules were being tested on experimentation.

The stages identified by the author in development of his research are shown in sequential order below. They are –

1. The application of the algorithm to the test data,
2. The projection of the results which will be expected to be produced by the running of the test data,
3. The re-design and upgrading of the software that implements the algorithm,
4. The testing of the software with the test data, and comparing actual results with projected results,
5. The refinement of the algorithm, which is encompassed by stages 1 to 4, until no further algorithm improvement is achievable.

To greatly simplify the research objective of this project the author took the decision, with the agreement of the author's supervisors, that the research would be limited to essays of a text-only nature. No tables, figures, graphics or any other non-textual elements of essays would be considered. Therefore it was necessary that all non-textual elements be deliberately purged from all the essays used before marking could proceed.

The author was fortunate in obtaining a range of twelve essay sets. Ten of these essay sets proved suitable for content marking. The remaining two sets were extensively used for algorithm development, especially in the early history of this research project.

It was regrettable that all essay sets obtained had been marked for Content only.

The author was unable to obtain marked essay sets for Style. The author has made numerous attempts to obtain essay sets that had been marked for style, including *offering to pay* for such sets, but to no avail.

In the author's mind automated essay marking should be for style in nearly all the essay assessments and for content as appropriate. Without marked essay sets for style then the development of style marking algorithm is effectively suspended. Nevertheless this chapter tries to cover what would happen if this crippling absence was not present. The reader should thus be assured that the author does not hide behind this absence as an excuse for excluding style marking.

3.2 Text extraction methodology

The decision to start this research work with essays as word-processed files reflects the fact that it would be natural for essayists to use the word-processing technology they are familiar with in their studies. This is the currently institutionally preferred method used by students (essayists) for creating essays. It is thus more acceptable for essayists to produce their work as document files rather than to impose some sort of bespoke essay creation process to merely satisfy any automated assessment software. Such assessment software should fit in with the skills' set and skills' level of the essayists; it should not force the essayists to change his or her own essay creation protocol(s) or methodologies. In short any essayist producing essays destined for automated marking should not have to learn how to use specialist software merely for the convenience of automated marking.

Having made the initial decision to work with essays as word-processed documents during this research project, the author then required these documents to be manipulated in such a way as to extract the text as American Standard Code for Information Interchange, more commonly known as ASCII, plain character based files. After reviewing the various format options for extracting the text from word documents to save as a separate text file, the author decided to develop a bespoke package for this extraction.

Due to its extremely high market presence, the family of word-processing packages selected for use in this research project was Microsoft Word™® for Windows ™®, and main versions of this particular package. This particular package has been adopted by the Robert Gordon University for its standard word-processing software.

The author as a starting off point initially explored the document formats options available to the essayist. There are several document format options available which are all accessible from within the various versions of Word using menu sequence such as "File > Save As > Save as type" format options. Of these available document format options, only the text related ones were considered as being suitable for investigation in this research project.

There is a range of text based document formats that are possible with Microsoft Word97 ™®. These options are listed below. This list is in the order as presented in the Word97 word-processing package and includes:

- Text only
- Text only with line breaks
- MS-DOS text
- MS-DOS text with line breaks
- Rich text format
- Unicode text
- MS-DOS text with layout
- Text with layout

There are however, two problems arising from the use of any of these methods.

The first problem occurs when any essay contains one or more non-textual elements. The removal of these non-textual elements from the essay by using any of these formats may result in uncontrollable or unpredictable effects in the extracted text file. Therefore the use of any of these excellent formats could cause problems for the downstream or subsequent program components of SEAR. The easiest point of origin for this research is that of files containing textual elements only. This avoids extraneous complications and so makes the research easier. The proper inclusion of non-textual elements is a topic for further research (see Chapter 6) after this current research project is finished.

The second problem is that the presence of all of these available text formats would require further file processing before it is realistic to submit the extracted text files to the Style and or the Content algorithms.

In the course of the development of both of these algorithms, especially the Content algorithm, further processing was indeed required subsequently. This requirement was to express the file of extracted text in the format of one, and only one, sentence per line of text. By using a bespoke extraction methodology the author exercised full control of the overall extraction process as well as control of all additional processing requirements that could be potentially incurred by the Style and Content algorithms.

However, the decision to use a bespoke extraction methodology created the problem of how to extract the plain text from the many versions of the selected family of word-processing package. This problem is compounded by the fact that the document file format is not the same across the different versions of the word-processing family. In addition using different Microsoft Windows ™ ® operating systems but with the same version of Word also produces different document file sizes.

The method used to illustrate the relative document size was to create a small ASCII plain text file, actually the text of the poem *Desirata* by Max Herman which he composed in 1929. This plain file was imported into each word-processor / operating system combination tested. The font type for each such file was set to Tahoma, and the font size was set to 10 point, font colour set to black and no text enhancements used. The file was then saved.

The base value for relative sizing was set to the value of 100 and assigned to the size of the smallest document size recorded - that is Word 6 running on Windows 3.1. For comparison the original plain ASCII text is included in the Table 3.2.a below.

Word	Operating System	Relative size
ASCII plain text	All MS Versions	16
Word 6	Windows 3.1	100 ~ base
Word 6	Windows 95	105
Word 95	Windows 95	160
Word 97	Windows NT	270
Word 2000	Windows 2000	195
Word XP	Windows XP	227

Table 3.2.a: Relative document sizes

From the Table 3.2.a it can be seen as a general trend that the more modern or the more sophisticated the word-processor being used, the larger the size of the resulting document. Exactly the same general trend is seen with the resulting document size when using a more modern operating system.

There is a further problem with using different versions of Microsoft Word ™®, which is that of the internal document structure. The author experienced no little difficulty in obtaining both legal and accurate information of the internal structure of the various versions of Microsoft Word.

This diagram Figure 3.2 below is intended to illustrate the fact that the text of the essay is in fact a continuous ribbon of characters sandwiched between data and metadata ('metadata' is data about data) held on the document.

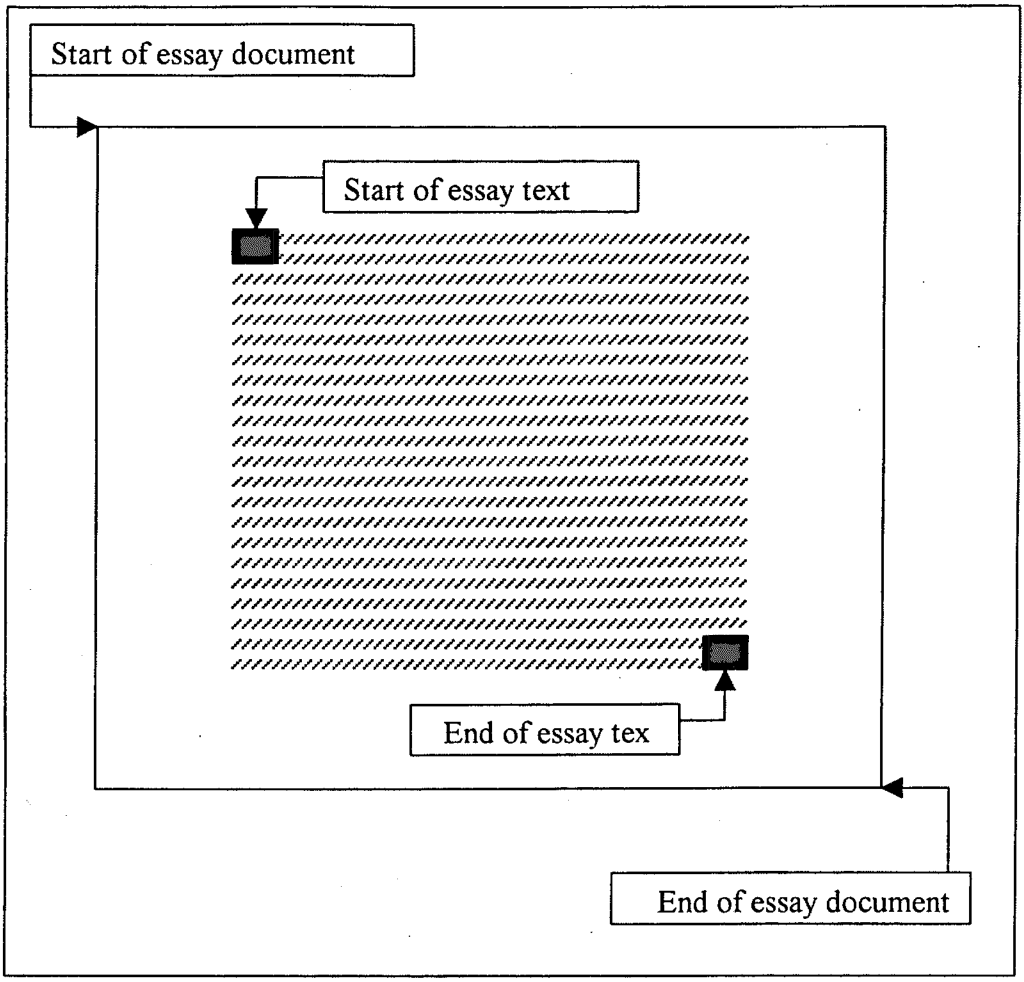


Figure 3.2: Word document schematic

Although the data and metadata is effectively outwith the scope of this project perhaps the reader would benefit from knowing what some of the non-text parts of a word document hold. Data is items like the creation date, the author, and the file location. Other data held is editing time, number of versions and the date last edited. Metadata holds data on the font(s) used, where their colour, size, starting and end positions, also where and what text enhancement(s) are used such as **bold**, *italics* and underline. Metadata includes data on non-textual components such as table(s). There is a record kept of the various edits that has been done on that document at that editing session. For all the versions of Microsoft Word™® this sandwich structure is maintained. However the starting position of the essay text is not constant across the different versions but the marker, or flag, 'end-of-text' used to indicate end of the document text is nearly constant across the versions as being "0D 00" in hexadecimal.

Therefore the author had himself to investigate the internal structure of the various Word document versions. Table 3.2.b below shows the salient structure features obtained from this investigation.

Word Version	Version code at byte 49	Text starting position d ~ decimal / h ~ hexadecimal	End of text marker character sequence
4.3	06 ;	2944d / B80h	0D 09 00
95	06 >	1280d / 500h	0D 00
97	06 >	1536d / 600h	0D 00
2000	06 >	1536d / 600h	0D 00
XP	06 >	2560d / A00h	0D 00

Table 3.2.b: Internal Document Features

These values had to be determined by the author. The author made use of a powerful, yet small, software tool that is called Debug. This tool is supplied, free of charge, with Microsoft operating systems. For each version of Microsoft Word, Debug was used to determine the differences in the actual document file produced by the author making simple variations in the document.

In the software the author had created for the extraction of text from various Word versions the first requirement is to determine if the essay file is of a Microsoft Word document. If it proves to be of this format then the next step is to determine which version of Word was actually used by the essayist. Each essay is considered in turn. There is an expectation, but no presumption, on the behalf of the author to expect that all the essays, in any or all the essay sets, will be produced by the same version of Word. Thus each essay could potentially be stored in any Microsoft Word document format. Therefore no overarching, or blanket, approach to text extraction would be acceptable when dealing with any essay set in particular nor would it in fact be practical to adopt such an approach.

It chapter 6 there is mention of the incorporation of non-textual elements into the essay and the use of different text enhancements with or without the various font effects of typeface, colour and size. The same process of investigation, probably starting with the use of Debug, will have to be done to establish the specialist marker(s) for these various features listed above.

Of course, there are other fully acceptable word-processing packages in competition to Microsoft Word. The same method of investigation, again probably starting with the use of Debug, would be used to establish the internal document structure of these competitive word-processing packages.

3.3 Wordlists used in both the Style and Content algorithms

Whenever a specific wordlist has been encountered in the literature survey, for example Dale words; or in the case of wordlists that the author knows exist before he started this research project, for example Roget's Thesaurus; then every possible effort was made to obtain that specified wordlist. Once it was obtained then it was converted it into a format suitable for this research project.

It must be noted that the author is not an expert in linguistics.
From this lack of linguistic expertise there are two significant problems that have been encountered.

Firstly, the wordlists that have been included in the collection used for this research project may be inappropriately included. They may not be appropriate, or they may have been superceded by more recent version(s) and or by more comprehensive versions, or even may have been superceded by alternative wordlist(s) that have been produced by other compiler(s).

Secondly, there may be other wordlists and, possibly, more appropriate wordlists that should have been included. However, the author is not aware of them at the time this research project was undertaken.

Nevertheless, for the purposes of this research these two significant problems are not too detrimental to the methodology. Should alternative wordlists be desired, then when they are obtained and converted into the required format, they would then be easily incorporated into the existing collection of wordlists. In this way these additional wordlists would be available to be employed by the algorithms for style marking and for content marking. Inclusion of alternative wordlists or new wordlists would be a technical operation.

All the words held on the collection that is comprised of the main wordlists and the additional wordlists are free from known spelling errors. Where applicable, all alternative spellings, such as United Kingdom and American, are included thereby maximising the utility of the wordlist.

Each of the main wordlist files in the wordlist collection has been obtained from a single source that is unique for them. If alternate versions are known only one version is used.

All the **main wordlists** are held in the same simple format. This format is:

- Each wordlist is held in its own folder within the general folder of wordlists,
- Each folder contains separate files of words, based on the first letter of each word,
- Each file contains one word per line ~ optionally there is additional information prefixing each word,
- All words are held in ascending alphabetical order.

This particular format was selected as being the easiest to work with in terms of the development of both the style and content algorithms and the storage of the wordlists, but it results in a lower throughput when marking essays. Should any commercial considerations arise from this research project then, of course, a more throughput effective storage format may be designed.

In the case of one specific wordlist, Roget's Thesaurus, this format is augmented by pre-fixing each word occurrence with a code for the part-of-speech and the topic number. This particular augmentation is discussed later on in this chapter. Each occurrence of part-of-speech, topic number and the actual word is only listed once. When a word has different part-of-speech and topic number then the order of listing is Noun, Verb, Adjective and lastly Adverb; and where there is more than one topic number entry for a part-of-speech listed for a word then the order of listing is ascending order of topic number.

In addition there are three **specialist wordlists** each stored as a single file in ascending alphabetical order within each file. These three files are detailed as follows.

- The first is a file of "stop words", which is currently a list of 411 words, and this list is alternatively known as "function words",
- The second file is a file of pronouns, which is currently a list of 75 words,
- The third file is a file of prepositions, which is currently a list of 81 words.

All three of the above wordlist files have been constructed with a view to their potential use in the marking of style. This is particularly true of the latter two files. At the present time only the "stop words" specialist wordlist file serves any real purpose in the marking of content.

The three specialist wordlists are different from the other files of the wordlist collection. These particular files have been created by combining words from a number of different sources. Many of these sources were collected from the Internet, with the rest being sourced from paper-based media such as dictionaries and books on English grammar.

Regardless of the original version of each of the main and the specialist wordlists, the author undertook various deliberate actions to maximise the effectiveness of each of the wordlists in the collection. All these deliberate actions are concerned with the alternative spelling of words and all of these actions served to inflate or to increase the size of each wordlist. By such deliberate inflation arising from the inclusion of alternative word spellings, the author seeks to make each wordlist independent of the original spelling, thereby avoiding any problem or problems that may be caused by the author's adoption of a particular version of English spelling.

The first deliberate action was to include versions of United Kingdom (UK) and American (A) spelling of words. A few examples should be sufficient to show the effect of this particular action: col**our** (UK) and col**or** (A), harb**our** (UK) and harb**or** (A), neighb**our** (UK) and neighb**or** (A).

The second deliberate action was to include different versions of word endings.
Again a few examples should suffice to demonstrate the effect of this particular action: normal**ise** and normal**ize**, custom**ise** and custom**ize**.

Finally, the third deliberate action was to include both non-hyphenated and hyphenated versions of words together with non-apostrophe and apostrophe versions of words. Again a few examples should be satisfactory in fully displaying the effect of these differences: antiflammatory and anti-flammatory, nonhyphenated and non-hyphenated, wont and won't.

Should an additional wordlist or additional wordlists be discovered which are pertinent to automated essay marking research then, except where there is a specific requirement to do otherwise, the same three deliberate actions to maximise effectiveness will be undertaken.

Should it be necessary to revoke any, or all, of the deliberate actions in generating alternative word versions, then any such "clean up" alterations should be a simple and straightforward technical operation. This is a consequence of deliberately using a simple format for holding the various wordlists.

Further, for the purposes of this research project, the author suggests that some incompleteness in the collection of wordlists would not appreciatively adversely affect the research methodology, nor would it invalidate any of the research findings.

The collection of the computer-mediated wordlists associated with this research project has been achieved from conducting various searches of the literature and various web sites that are relevant to this research project. The wordlists that are held on the computer mediated files used in this research, have been created either from paper based lists or have been downloaded from various locations on the World Wide Web. Paper based wordlists were converted into computer mediated files by the author.

The range of wordlists collected varies in size that is in the number of words that are included therein. An indication of the range can be seen from the smallest being, Edwards-Gibbon's which has 627 words to the to the largest, being Roget's, which has 35,554 unique words and 69,713 words overall when the multiple entries for words are included in the word count.

The creation date of the collected wordlists also varies. The oldest is the Thorndike-Lorge wordlist version of 1944 (the first version was created in 1921), which contains 1,000 words, while the youngest, Grady Ward's Roget's words was presumed by this author to be circa 1993. Although technically the youngest Ward claims to have based his wordlist on the 1913 version of Roget's Thesaurus.

There are some commercially available wordlists but, unfortunately these are too expensive for the author to afford at the present time of this research project. Furthermore many of these commercially available lists are very large and as such are more suited to being employed in any commercial products such as dictionary creation than for purpose of this research project.

Using larger wordlists should be regarded as being "more of the same" and would not improve the findings of this research substantially. For example, consider the Bank of English, which is hosted by Collins <http://titania.cobuildcollins.co.uk> (as at August 2002). It has some 450 million words in its corpus. However, access to it is controlled by both costs of licensing and output arising from Collins' own software. Access to their main corpus is, naturally, not permitted to those outside the Collin's commercial remit.

When anyone wishes to create, or compile, a new wordlist, or corpus, there is the problem in the collection of the required materials – newspapers, magazines, media broadcasts, books and so on.

The initial problem facing compilers of corpa is the selection of those items to be used as the source materials. The compilers then have to assemble the words in their selected sources into the one corpus or wordlist.

Sorting, tagging, classifying and related activities are then undertaken by compilers. Once the compilers have completed their self-assigned tasks, their work is then released to the public or the sector designed by the compilers as their audience. This simple sounding sequence of tasks actually masks what in fact is a very long-term commitment on behalf of the compilers.

Even in the days of modern computers, this sequence of creating a new wordlist, or corpus, will take a considerable amount of time and effort. Therefore there is always going to be a delay, often considerably, in the production of the corpus. As a result, the longer the time interval taken to publish any wordlist means that there is the opportunity for some level of divergence between the vocabulary in vogue as present in the source materials and the actual vocabulary in vogue at the current time of release. So, whatever the reason for the creation of the corpus in the first place, the corpus is to some extent obsolete and out-of-date simply as a result of time taken for the assembly of the wordlist or corpus.

For an example, it is stated that Dr. Roget started work on his world famous thesaurus in 1805 when he was aged 26 years old, yet it was first published in 1852 and Dr. Roget kept on working on his thesaurus until he died in 1869 aged 91.

It is at this point that it would be the time to examine some statistics on wordlists in order to gain some insight into wordlists in general.

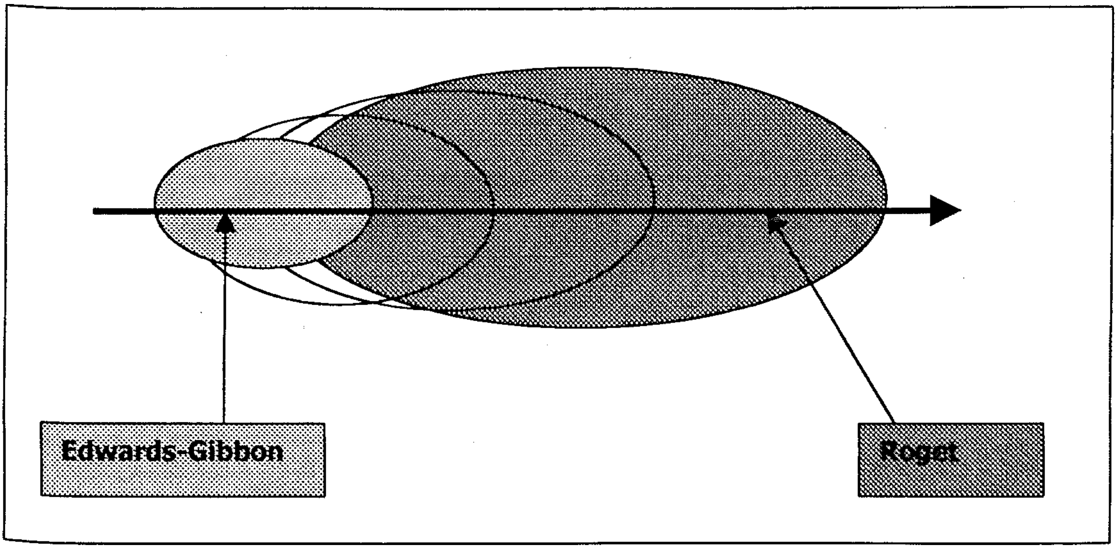


Figure 3.3.a: Wordlist relationship

This diagram serves to show the reader that there is a volume relationship between the different wordlists that are to be outlined later on in this sub-section. Edwards-Gibbon wordlist (1973) is based on the vocabulary of early primary school children and Roget is aimed at adult use. Then unlabelled ellipses are to represent the volume of different vocabularies between the two end ellipses to represent how an individual's vocabulary changes with time (the arrow).

From the same diagram is may be inferred that at each stage of vocabulary development some words are always there regardless of stage, new words are added, and other words are dropped (the un-shaded parts are the represent words that have fallen from an individual's use). Words may be dropped for a while and then be re-included at a later stage. For example the word 'pussy-cat' – (very) young children will frequently use this word, a teenager seeking to impress would not, yet parents and grandparents will re-use this word (perhaps only in the presence of their grandchildren!).

So, an individual's vocabulary is a dynamic volume of words that changes with time, with education and with experience. The wordlists used in this project try to capture vocabularies that are aimed at specific target ages, groups and so on. The author does not infer that the wordlists that have been collected for this project are fully representative of the whole range of vocabulary.

Frequency of word usage (Thorndike, 1952)

Table 3.3.a, and the graph in Figure 3.3.b, below illustrate the frequency of the usage of words by the general public.

Frequency (per million words)	Number of different words		**Frequency (per million words)**	Number of different words
>=100	1069		**25**	85
50 – 99	952		**24**	72
49	36		**23**	98
48	35		**22**	110
47	33		**21**	112
46	38		**20**	131
45	42		**19**	132
44	38		**18**	145
43	39		**17**	172
42	30		**16**	188
41	40		**15**	189
40	45		**14**	200
39	41		**13**	231
38	47		**12**	294
37	50		**11**	316
36	62		**10**	340
35	53		**9**	441
34	64		**8**	522
33	55		**7**	593
32	65		**6**	684
31	54		**5**	890
30	74		**4**	1064
29	74		**3**	1442
28	77		**2**	2503
27	88		**1**	5209
26	76		**0.2 – 0.9**	9202
			<0.2	1358

Table 3.3.a:

Distribution table of frequency of word usage per million words

Both the table and graph show that some 2,000 words are in very common usage, that is words with a frequency of occurrence equal to, or greater than, 50 times per one million words. As the frequency of usage falls the number of words with that frequency increases. This reflects the richness of the English language and shows the precision with which members of the general public use their vocabulary. Other deductions about socio-linguistics may be made from this Table 3.3.a and the graph in Figure 3.3.b especially when coupled with other data; however these deductions are out-with the scope of this research project.

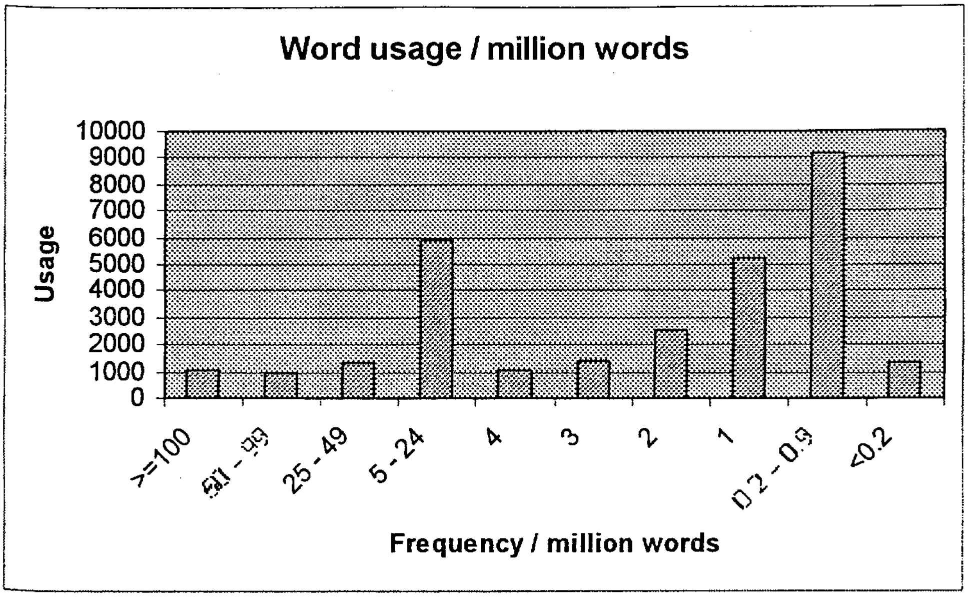


Figure 3.3.b:

Distribution graph of frequency of word usage per million words

For the purpose of this research project the author used the 1,000 most popular words identified by Thorndike (Thorndike 1952), as part of his collection of wordlists. This particular is identified as the Thorndike-Lorge wordlist.

For each wordlist that has been included in the collection a short overview is presented in the next section. This overview explains some of the salient features of each wordlist and follows the summary Table 3.3.b below.

Letter	Dale	Thorndike-Lorge	Edwards-Gibbon	Ward	LOB	Roget	TOTAL
A	125	58	26	73	508	4,550	5,340
B	246	66	57	42	343	3,475	4,229
C	276	69	58	66	731	6,744	7,941
D	142	47	27	45	443	4,696	5,400
E	70	34	10	44	368	3,146	3,672
F	169	62	43	61	338	3,231	3,897
G	114	39	24	20	199	1,990	2,386
H	173	50	34	35	248	2,399	2,939
I	42	24	9	34	304	3,722	4,133
J	32	6	7	4	69	521	639
K	30	12	7	6	49	374	478
L	118	44	25	45	262	2,152	2,646
M	129	53	26	52	334	3,244	3,838
N	59	33	12	22	139	1,213	1,478
O	61	27	14	28	182	1,624	1,478
P	191	52	41	63	547	5,654	6,548
Q	12	5	1	3	29	394	444
R	122	40	25	47	459	3,979	4,672
S	401	118	75	98	841	7,855	9,388
T	185	68	49	60	384	3,265	4,011
U	32	9	7	11	96	2,365	2,520
V	18	8	2	8	106	1,266	1,408
W	150	60	40	43	241	1,600	2,134
X	0	0	0	0	1	37	38
Y	22	9	7	5	31	142	216
Z	0	0	1	0	3	94	98
TOTAL	2,919	993	627	915	7,255	69,732	**82,429**

Table 3.3.b: Overview of the Wordlist Collection

Table 3.3.b shows that the most common starting letter is the letter 's' and the least common is 'z'. The distribution of frequency of the first letter of the word does not follow a normal distribution, as there are peaks at 's', 'p' and 'c', although these peaks are of decreasing size.

Table 3.3.c below shows the distribution of words by length across the wordlist collection.

To improve clarity where the frequency is zero the entry is left blank.

Length (letters)	Dale	Thorndike – Lorge	Edwards – Gibbon	Ward	LOB	Roget
1	2	2	2	2	2	1
2	33	26	23	30	54	65
3	299	119	111	105	242	559
4	747	312	205	247	773	1808
5	676	219	128	183	1039	2702
6	515	165	76	169	1213	4029
7	313	88	41	115	1152	4692
8	168	37	20	66	957	4983
9	99	17	11	44	762	4884
10	46	8	9	26	485	4151
11	12		1	9	300	3122
12	6			2	149	2056
13	5			1	73	1268
14				2	34	676
15					6	323
16					1	124
17						66
18					1	20
19						7
20						4
21						3
22						
23						1
24						
25						

Table 3.3.c: Word Length Profile

Table 3.3.c reveals that the larger the wordlist becomes there is a drift from a modal word length of 4 to 8. This is probably due to there being a common core of smaller words in all the wordlists. So, as the wordlist increases in size then the increase in the words listed must be sourced from more less commonly used words.

There are different ages represented by the wordlists in the collection. As the target age, or the effective age, associated with the wordlist increases then, again, there is a drift in modal length from 4 to 8.

For all the wordlists there appears to be a normal distribution across the word length profile. In addition the maximum word length appears to increase as the size of the wordlist increases, and likewise the maximum word length increases as the effective age increases.

This table of word length profile was used by the author to set the maximum word length for words to 25. There are only 4 words of length greater than 20. These four words, whose length is given in brackets, are Chromato-pseudo-blepsis (23), Chromatopseudoblepsis (21) (the same word but without the hyphens), Indistinguishablility (21) and Otorhinolaryngologist (21). These three different words are in emboldened in Table 3.3.d below.

Therefore the author decided on the use of a maximum word length of 25 which would not lead to the problem of truncating words that are longer than 25 characters. Using a maximum word length of 25 characters would not lead to waste in the provision of computer memory for storing longer word lengths that are not likely to be encountered.

Although not appearing on any of the wordlists used in this project, the author has found other long words. With regard to the imposed limit of using a maximum of 25 lettered words those words in the range 20 to 25 letters are:

The 22 lettered words are:

electroencephalography,
honorificabilitudinity,
polytetrafluorinethylene.

The 23 lettered words are:

polytetrafuoroethylene,
transubstantiationalist.

Only one 25 lettered word has been found by the author, namely:
antidisestablishmentarian.

However the author did find four words that are longer than 25 letters. These words are:

antidisestablishmentarianism [28 letters],
floccinaucinhilipilification [28 letters],
dichlorodiphenyltrichloroethane [31 letters],
pneumonoultramicroscopicsilicovolcanoconiosis [45 letters].

As an aside the author expects that there are considerably more than four words longer than 25 letters, but finding them is not in the scope of this research project.

The details of the current wordlists available in this research project are given in the following paragraphs.

3.3.0.1 Dale Wordlist: about 3,000 words

The date of creation is about 1948. In order to construct this list Dr Edgar Dale of the Ohio State University surveyed children in the 4th Grade at school. The Dale wordlist is thus a list of some 3,000 words and contains those words that were recognised by 80% of the children that were included in Dale's survey.

Only four percent of the words are more than two syllables long, and a further four percent excludes proper nouns and words formed from simple words. This is rather a low figure when it is realised that most words in common use are of three syllables or more.

The Dale Word List was frequently used by the earlier researchers into the automated marking of essays for style. This usage is shown by these researchers directly referring to the use of "the Dale Word List" (or "Dale Words") or indirectly by the researchers' oblique reference to "the list of common words".

Source: *The Technique of Clear Writing By Robert Gunning,* Appendix B (Gunning, 1968).

3.3.0.2 Dale-Gibbs Wordlist

The list is about the same size as the Dale word list and its content are similar to those of the Dale Wordlist. As the volume of differences between Dale and Dale-Gibbs wordlists is so small this list was not deployed in this research project. If it proves necessary to deploy the Dale-Gibbs wordlist, such deployment would be rapidly achieved.

3.3.0.3 Thorndike-Lorge Wordlist: about 1,000 words

Edward Thorndike and Irving Lorge created their original list in 1921, then enlarged it in 1931 and published the particular version in 1944 used herein. Some 120 written materials were used as sources in the creation of the earlier versions. These particular materials were suitable for use as reading materials for younger children; some were used in their entirety, while others were only partially used. For the wordlist published in 1944 a revised approach was used. In this version more sources were used, about 300, and these sources encompassed a wider range of written materials.
Source: *The teacher's handbook of 30,000 words.* (Thorndike, 1952)

3.3.0.4 Edwards-Gibbon Wordlist: about 600 words

R P A Edwards and Vivian Gibbon created the second edition of this graded vocabulary (for the Leicestershire County Council) in 1973. The first edition was created in 1964.

This wordlist was generated by collecting the "spontaneous writing" of 5+, 6+ and 7+ year-olds in the two terms spanning September 1961 to March 1962. Some 2,120 children of mixed ability and mixed backgrounds were used. This group was made up of 820 5+ year-olds, 794 6+ year-olds and 506 7+ year-olds. These children were drawn from some 45 schools ranging from single class village schools to large country, urban and sub-urban infant schools.

Teachers collected the spontaneous writing produced by the children. There were no set topic assigned to direct the children's writing. After collecting the children's writing, the teachers then listed each word, proper names being omitted, with its frequency for each sample of writing. Summaries of the listed words were made at the school level. These school summaries were then collected and collated. Words that were used by less than 1.5% of the children were purged from the wordlist. Of the 2,788 words held on the overall wordlist, 658 words were left on the 5+ years-old wordlist, with 789 words for the 6+ wordlist and 1,343 words for the 7+ wordlist.

These wordlists are organised on the basis of a popularity index, rather than the more usual orderings of pure alphabetical or pure frequency. The popularity index was formed from the mathematical product of the percentage of children using that word multiplied by the average use of the word by frequency. This index attempts to produce a better match between the children's vocabulary and the wordlist than the match that is found using the style of previously organised wordlists. The use of the popularity index does affect the ordering of words.

Perhaps an example taken from the reference will show how the popularity index alters the listing of words. Consider the two words 'and' and 'go'. The word 'go' was found to be top of the ranking, when ranked by the greatest use amongst the children. However the word 'and' was used by less children, but the frequency of use by these children was five times that of the word 'go'. Therefore by the use of the popularity index the word 'and' is ranked higher than the word 'go'.

By generating the popularity index Edwards and Gibbon were able to better fulfil the purposes for the collection of children's writing. The first purpose was to help the creators of reading books for children match the vocabulary that these children actually know. The second purpose was to provide teachers with a means of assessing their pupils' vocabulary.

Various sub-lists of words were produced from the overall wordlist. Such sub-lists included the list of words only used by 5+ year-olds and the list of words used by 6+ year-olds that are not used by 5+ and 7+ year-olds.

Source: *Words your children use*, Edwards-Gibbon, 1973, ISBN 0-222-0122-4

3.3.0.5 Ward Wordlist: about 900 words

This is a short list purporting to be of the 900 or so most frequently used words that Grady Ward found in his research. The author found that in this list quite a large number of words were repeated. Ward constructed his list from a "wide variety of common texts", a term used by Ward himself to describe his sources.

Source: The institute for Language, Speech and Hearing, Sheffield University, <http://dcs.shef.ac.uk/research/ilash/Moby/>, as at November 2000.

3.3.0.6 Lancaster-Oslo/Bergen (LOB) Wordlist: about 7,000 words

Work on the creation of this wordlist was started in 1970 in the University of Lancaster. In the subsequent seven years the work continued under the stewardship for Geoffrey N. Leech. In 1977 the work was completed by Knut Hofland & Stig Johansson; a joint effort by the Department of English in the University of Oslo with the Bergen's Norwegian Computing Centre for the Humanities. The idea of creating the LOB corpus was to produce a "British English" corpus to complement the Brown corpus, which is based on "American English".

The LOB corpus creation mimicked the sampling methodology used to create the Brown corpus. This is to maximise the match between the two corpa. For the LOB corpus some 500 text samples were used and each text sample was of some 2,000 words in size. These 500 text samples were drawn from a predetermined profile of 15 different types of written materials that were in circulation to the general public. The year that these 500 text samples were sampled was 1961. This particular year and the predetermined profile of the types of written materials samples were selected to maintain correspondence with the Brown corpus.

For inclusion in the LOB corpus a word had to be present in at least five different text samples and to have been used at least ten times.

Source: Word frequencies in British and American English (Hofland, 1982)

3.3.0.7 Roget Wordlist: about 70,000 words

Numbers of Words: Total 69,713; Unique 35,544; Multiple entries 34,269

This computer-mediated version of the Roget's Thesaurus was created by Grady Ward in 1993. Ward based his version on the 1911 version of Roget's Thesaurus, augmented by many additions to bring the content of the thesaurus up-to-date.

Source: The institute for Language, Speech and Hearing, Sheffield University, <http://dcs.shef.ac.uk/research/ilash/Moby/>, as at November 2000.

This is the only wordlist in the current research project where there are multiple occurrences of the same word. This is completely due to the underlying purpose of the Thesaurus. In the English language a word may have multiple meanings or uses. Further in the English language very often a word may have several equivalents, though with slightly different nuances of meaning. In the Roget's Thesaurus based wordlists each occurrence of a word is included in the wordlist.

Two examples taken from the Penguin 1966 version of Roget's Thesaurus should show these twin sources of multiplicity.

First consider the word "cross".

This word occurs in the following eight Roget topics, or 9 if the entry for topic 222 is counted for both noun and verb:

hybrid	(43noun),	counteract	(182verb),
cross	(222noun, 222verb),	bane	(659noun),
decoration	(729noun),	angry	(891adjective),
religious faith	(973noun),	talisman	(983noun).

Second consider the adjective 'angry', which is listed as Roget topic 891 adjective.
The twenty adjectives listed in 891 are:

angry, displeased, serious, impatient, cross, waxy,
ratty, wild, wroth, wrathy, ireful, irate,
indgnant, incensed, infuriated, shirty, furious, apoplectic,
rabid, mad.

Yet this list does not include two adjectives in common use, namely 'livid' and 'seething'!

Just to complicate matters further, there are seven adjectival phrases equivalent to 'angry': "not amused", "worked up", "het up", "hot under the collar", "in a temper", "in a taking", "hopping mad".

Roget's Thesaurus contains both words and phrases. In the wordlists the author created from the computer-mediated version of Roget's Thesaurus, no phrases have been included in the content marking algorithm. However, in anticipation of extending his further research work after the completion of the current research project the author has prepared lists of the phrases taken from Roget's Thesaurus. Lists of phrases are based on the same format as for wordlists.

First Letter	Nouns	Verbs	Adverbs	Adjectives	TOTAL	Biggest Word (length)	POS / Number
A	2,581	769	174	1,026	4,550	Anti-flammatory (17)	N / 662
B	2,080	694	57	644	3,475	Beggar-my-neighbour (19)	N / 840
C	4,075	1,183	74	1,412	6,744	**Chromato-pseudo-blepsis** (23)	N / 443
D	2,357	1,180	53	1,106	4,696	Disproportionateness (20)	N / 24
E	1,539	785	88	734	3,146	Electromagnetism (16)	N / 157
F	1,674	590	55	912	3,231	Favourably-minded (17)	J / 602
G	1,220	293	20	457	1,990	Gastroenterologist (18)	N / 662
H	1,321	322	71	685	2,399	Hemidemisemiquaver (18)	N / 413
I	1,596	615	93	1,418	3,722	**Indistinguishablility** (21)	N / 464
J	357	81	3	80	521	Jack-in-office (14)	N / 887
K	265	59	2	48	374	Kleptodipsomania (16)	N / 503
L	1,193	307	35	617	2,152	Latitudinarianism (17)	N / 984
M	1,982	497	48	717	3,244	Miscellaneousness (17)	N / 78
N	702	124	46	341	1,213	Notwithstanding (15)	D / 30
O	783	345	45	451	1,624	**Otorhinolaryngologist** (21)	N / 418
P	3,395	855	102	1,302	5,654	Paleoanthropological (20)	J / 124
Q	229	67	8	90	394	Quadrifoliolate (15)	J / 96
R	2,126	985	42	826	3,979	Refigerator-freezer (19)	N / 387
S	4,143	1,532	154	2,026	7,855	Self-gratification (18)	N / 374
T	1,896	531	88	730	3,245	Transsubstantiation (18)	N /140
U	337	210	57	1,761	2,365	Unquestionableness (18)	N / 474
V	763	160	32	311	1,266	Valetudinarianism (17)	N /656
W	760	310	84	446	1,600	Without-foundation (18)	J / 546
X	28	0	0	9	37	Xantho-cyanopia (15)	N / 436
Y	75	31	16	21	143	Yieldingness (12)	N / 762
Z	78	1	1	14	94	Zoohygiantics (13)	N / 370
TOTAL	37,555	12,526	1,448	18,184	69,713		

Table 3.3.d: Summary of the Wordlist based on Roget's Thesaurus

3.3.1 Accuracy of content of wordlists

All words in each of the wordlists are believed to be as accurate as possible. In other words all known spelling errors have been expunged.

Checks were made on the accuracy of converting from both paper sourced and Internet sourced wordlists into the format required for this research project. When any discrepancies where found they were immediately resolved. In the event of any further discrepancies in accuracy being found they will be resolved immediately. However there is very little that the author could do to check that the source materials are in themselves accurate.

There is one known "false positive" problem in these wordlists. This is that should any incorrectly spelt words are included in any of the wordlists this results in an inadvertent inflation of the volume of wordlists and gives rise to the opportunity of the occasional 'extra' count when matching the words in an essay against the wordlist or wordlists.

Conversely there is one known "false negative" problem in these wordlists. This is that should any words be missing from any of the wordlists then when matching the words in an essay against the wordlist or wordlists then there will be a shortfall in the count.

However, both these problems are, of course, systematic in their nature, therefore not one essay will receive differential effect when being marked for style of content. Thus, all essay sets will be equally exposed to the same false positives and false negatives

For the purpose of this research project, the author suggests that some inaccuracies in the collection of wordlists would not appreciatively adversely affect the research methodology, nor would it invalidate any of the research findings.

3.3.2 Completeness of content of wordlists

Checks were made for completeness of converting from both paper-sourced and Internet-sourced wordlists into the format required for this research project. When any discrepancies in completeness where found they were immediately resolved. Again, should any further discrepancies be found, they will be resolved immediately. However there is very little that the author could do to check that the source materials were in themselves complete. Therefore, there may be a word or words omitted from and the false inclusion of a word or words in, none, some or all of the wordlists in the collection.

Removal of errors from wordlist(s)

Should any errors of accuracy or completeness be found then the errors will be corrected immediately. All subsequent marking will be based on the then corrected wordlist or wordlists. The volume of errors in the collection of wordlists should be a reducing one – the more errors found then the less there will be to find. Since these wordlists are as accurate and complete as possible at their creation then the rate of error detection should be low or zero. As this error rate is so low, then there should be little need to first correct the errors and hence immediately to remark previously marked essay sets retrospectively.

3.3.3 Use in the Style algorithm

The substitution, the exclusion or the inclusion of any of the main wordlists already in the wordlist collection would not affect the methodology for the marking of essays for style. As it currently stands this methodology operates on the basis of a selection of the collected wordlists to be used in the generation of the style metrics. The inclusion of additional wordlist would not alter this methodology but would only expand the basis on which the possible selection would be made from.

The three specialist wordlists, namely "stop words", 'pronouns' and 'prepositions', may be acceptable for inclusion in the methodology for style marking. Their inclusion would be for the purpose of producing metrics such as the overall count of those essay words that are found in all of these lists or of producing three metrics giving the count for those essays words that are found in each one of these lists. In addition to these overall metric counts, the count for any specific word that is to be found in any of these specialist wordlists may be worthy of inclusion. For example the number of times the word 'they' is found in an essay. This will provide additional word specific metrics.

There is a part of the methodology for marking of essays for style that involves the potential use of none, one or more of the wordlists in the collection. The use of any specific wordlist or wordlists seeks to reflect the essayist's use of any expected vocabulary, or in other words does the essayist use "common words" appropriate for that level of education.

All the wordlists show virtually the same pattern of distribution of words by their first letter. This, while in itself not a surprising pattern, serves to indicate that the general distribution of words by their first letter would not adversely affect preference in the selection of one wordlist or wordlists over other wordlists.

In other words, the balance of distribution of words by their first letter is not significantly different when changing from one wordlist to another.

However, preference should be given to those wordlist or wordlists that best match the expected vocabulary of the essayist. As an illustrative example, for those essayists who are aged 5, 6 or 7 then the Edwards-Gibbon's wordlist should be given the highest preference from within the current wordlist collection. In this particular instance Roget's wordlist should be given the least preference since it offers the best match between wordlist and their expected vocabulary. On the other hand, for those essayists who are aged in their late teens or early twenties the reverse would make a better selection, that is preference should be given to the Roget's wordlist from within the current wordlist collection and not the Edwards-Gibbon's wordlist. This is for exactly the same reason, namely that of achieving the best match between essayist and vocabulary expected of those essayists.

Should there be a mismatch between selected wordlist, that is expected vocabulary, and the vocabulary used by the essayist then one of two mutually excluding effects will occur. Either a too advanced vocabulary is selected and the essayist will systematically have a low score on vocabulary usage; or a too limited vocabulary is selected and the essayist will systematically have a high score for vocabulary usage. In both of these situations any, or all, of the value of vocabulary based metrics used in the style marking algorithm would be reduced, and possibly invalidate it from inclusion in the metric pool.

3.3.4 Use in the Content algorithm

Only one of the main wordlists is used in the content marking algorithm. This is the Roget's wordlist, in particular the wordlist that has each word prefixed by two elements, namely part-of-speech and number. Part-of-speech and number are held, as far as possible, to be true to the organisation of Dr. Peter Mark Roget's original 1852 publication. Parts-of-speech used by Roget, and hence used by this author, are Nouns, Verbs, Adjectives and Adverbs. The number that is used is the topic number, which is also alternatively known as the head-number or heading-number or category-number or root-number as assigned by Roget. This author has deliberately forgone using the higher levels of the Roget classification, namely the six Class entries, the twenty-four Sections and the ninety-seven Sub-Sections, to concentrate only on topic number, part-of-speech and number. This is done as the author felt that the inclusion of the higher levels of the Roget classification would not ensure any significant improvement to the content marking algorithm. In Roget's thesaurus domains are seen in the topics, where they are delimited by the use of semi-colons.

The use of domains as in the Roget classification has also been ignored by the author as, again, such use offers little potential value to content algorithm. The higher levels of Roget classification and the use of domains are more important when using the Roget 's wordlist as a thesaurus (its proper use), especially when one is using any "top-down" procedures such as synonym word searching.

From its inception in 1805 and its first publication in 1852 Roget's Thesaurus has enjoyed the publication of dozens of editions from a variety of publishers. As examples Roget's Thesaurus has appeared in its original format published by Penguin in 1966 and published by Roydon (1972). Other publishers have produced different formats: HarperCollins' (1996) and in an "A-Z" format Oxford University Press' (book 1997) and as a CD-ROM. Unfortunately, as will be shown later, different editions are not compatible. It is however still a widely used and highly respected reference source. The author is not aware of any publication similar to Roget's or any equivalent that exists for any non-English language(s).

Roget's wordlist is the only wordlist the author has obtained that has the facility to identify synonyms. This facility is very important in the content marking of essays as it allows the content marking algorithm to mark alternative wording as provided by essayists. In a crude manner the use of the prefix in Roget's wordlist in marking content loosens what could otherwise have proved to be a very strict word matching algorithm for content marking, and awards the deployment of a complex procedure for alternative word matching.

Oxford University Press has published a thesaurus in both paper and computer formats both in a radically different format to the Roget's Thesaurus. This radical format is a thesaurus based on the "A-Z" indexing of words – a format familiar to the users of dictionaries.

HarperCollins publication "Roget's International Thesaurus" mimics the general approach devised by Roget, but it is not in the same format. HarperCollins' format is based on a "synopsis of categories" – their own description of the format. This format has 15 classes, 1,073 categories within these classes, and then numbered paragraphs within the categories. The parts-of-speech used in this format includes the original four - nouns, verbs, adverbs and adjectives - augmented by three more – prepositions, conjunctions and interjections.

The classification is based on a two-part number held in a decimal format X.Y – where X is the category number and Y is the paragraph number within that category.

However, for this research there are several problems in using any of these radical formats detailed above. The first problem is that they are only available in paper format. Secondly these alternative thesaurus' non-standard classification scheme are felt by this author to be too radical to be deployed since most educated people should be familiar with Roget's Thesaurus in its original format. After this research project is finished the author intends to re-consider the decision on which thesaurus format to use.

Example of use

The use of a simple example should show how Roget's wordlist is used in the content marking schema: "The cat sat on the mat."

First, all the three stop words, namely both the "the"s and the "on", are ignored. Second, the assignment of both the part-of-speech and topic number is made to the remaining words. This assignment (from the different versions) gives the following Table 3.3.4 below:

Word	Penguin	Roydon	Ward	Collins	Oxford
cat	N 365 cat	N 366 animal	N 366	311.21 animals, insects	cat
sat	V 311 sit down	V 184 settle	* 000	173.10 quiescence	(sit)
mat	N 226 floor-cover	N 652 cover	N 215	295.9 covering	mat

Table 3.3.4: Roget based assignments

In the content marking schema these parts-of-speech and topic numbers are then stored against these three words. The table above shows an unexpected problem that occurs when using Roget's Thesaurus – different parts-of-speech and topic numbers occur when using different editions or versions!

The different assignments would not be too much of problem if the actual content of the different topics were the same – but they are not!

Cat Overall there are 20 different words for word 'cat in the five thesauri.
Penguin 9 entries: *cat, grimalkin, puss,* kitten, tom, *tomcat,* gibcat, mouser, feline.
Roydon 8 entries: *cat, puss,* pussy, kitten, *grimalkin,* gibcat, *tomcat,* mouser.
Ward 7 entires: *cat,* feline, *puss,* pussy, *grimalkin,* gibcat, *tomcat.*

Collins 20 entries: *cat,* feline, *puss,* pussy, pussycat, moggy, mog, tabby, *grimalkin,* kitten, kitty, kitty-cat, kit, kitling, *tomcat,* tom, gib, *gibcat, mouser,* ratter.

Oxford 15 entries: *cat,* feline, domestic cat, pussy, pussy cat, *puss,* moggy, *grimalkin,* tabby, *tomcat,* tom, ginger tom, kitten, mouser.

There are only 4 entries in common across all five thesauri for the word "cat"!

These are cat, grimalkin, puss and tomcat.

The overall set of the 20 words alphabetically is:

cat, feline, gib, gibcat, grimalkin, kit, kitling, kitten, kitty, kitty-cat, mog, moggy, mouser, puss, pussy, pussycat, ratter, tabby, tom, tomcat.

<u>Sat</u> Overall there are 22 different words for 'sat' in the five thesauri (excluding tense).

Penguin 5 entries: sit, squat, kneel, recline, couch.

Roydon 13 entries: place, situate, locate, localize, put, lay, set, seat, station, lodge, quarter, post, install.

Ward 5 entries: place, localise, lay, set, seat.

Collins 4 entries: sit, set, perch, roost.

Oxford 5 entries: place, position, put, situate, deposit.

There are no entries in common for the word 'sat', even allowing for the different tense!

The overall set of the 22 words alphabetically is:

couch, deposit, install, kneel, lay, localise, localize, locate, lodge, place, perch, position, post, put, recline, roost, quarter, seat, sit, situate, station, squat.

<u>Mat</u> Overall there are 7 different words for cat in the five thesauri.

Penguin 6 entries: carpet, *mat,* doormat, rug, drugget, matting.

Roydon 4 entries: cover, drugget, *mat,* doormat.

Ward 2 entries: *mat,* rug.

Collins 4 entries: rug, carpet, carpeting, *mat.*

Oxford 4 entries: *mat,* doormat, rug, carpet.

There is only 1 entry in common for the word 'mat' – and that is the word 'mat' itself!

The overall set of the 7 words alphabetically is:

carpet, cover, doormat, drugget, mat, matting, rug.

The author, thus, has had to adopt one version as the version to use. The basis for this selection, the Ward version of Roget's Thesaurus, is that it is the only computer mediated Roget's Thesaurus that the author had obtained.

In the actual marking for content any essay that can map basically onto this pattern of N366 - sat – N215 will be marked as being a valid alternative to the model answer. In terms of the number of possible alternatives there are 3,080 alternatives in the active voice – generated from 20 different words for "cat" times 22 different words for 'sat' (allowing for different tense) times seven different words for "mat" – 20 * 22 * 7 = 3,080. When the passive voice is included the number of possible alternatives increases to over 6,000.

If there are so many alternatives from just a very simple sentence, consider the possible alternatives that may occur in a thousand-word essay. The examiner must provide a clear answer schema, and must provide a model answer. The content data structure must allow for all of the alternatives being acceptable, but without the examiner having to be explicit in expressing or specifying what the alternatives are.

3.4.1 Development of style algorithm

The literature search revealed that the research into marking of essays for style had originated as early as the 1960's. It also revealed that during the intervening period spanning from the 1960's to the present day there had only been one basic algorithm generated. This is the weighted linear model, as shown in the figure below:

$$\boxed{StyleMark = \sum_{x=1}^{N} A_x Metric_x}$$

Figure 3.4.1.a: Weighted Linear Model

For this algorithm **N** is a set of metrics, and **A** is the co-efficient for each metric in the set.

The term 'metrics' is here used to encompass measurable features in the text.

These metrics are text features such as the number of words used, the number of different words used, the number of sentences, the average word length and so on. For further details see Appendix A of the actual metrics for marking style quoted by various researchers and Appendix B for this author's overall list of potential metrics for measuring style.

There is a very large number of such metrics but there is no standard set of metrics used. There does, however, appear to be some commonality of metrics used. The number of metrics, **N**, used varied with each researcher and this variation ranged from a very small number to a very large number. To illustrate this Johnston (1996) use six fixed metrics whereas Page (1966) used over 30 fixed metrics. Moreover **N** can be a selection of the best fitting metrics from a pool of metrics, for example a selection of the statistically best 50 metrics from a pool of two to three hundred possible metrics as used by Page in 1996.

Algorithm summary

A) First of all a sample of the essays is taken from the essay set.

It must be noted that the size of the sample is determined by the number of metrics to be considered. Indeed, statistically the sample size has to be equal to, or better than, twice the number of metrics to be considered.

B) The sample is then marked manually, usually by only one examiner.

C) The software is used to measure the complete set of metrics considered in each of the essays in the sample.

D) The co-efficient for each metric is determined by the application of statistical techniques such as multiple regression. Usually statistical software is used to determine these co-efficients. Should a variable metric list be deployed, for example a selection of X metrics from a pool of Y, it is at this stage that the best fitting metrics are selected. The selection of those metrics is determined by the application of statistical techniques, generally multiple regression or analysis of variance (usually shortened to 'ANOVA').

E) The linear weighted equation developed in stage D is now applied to all the essays in the set.

The author's initial literature survey revealed several surprising facts, and subsequently a few interesting questions were posed.

The facts revealed are listed below.

It was found that all the algorithms used very similar metrics. Some 15 to 20 metrics were commonly used. These are shown as the shaded rows in Table 3.4.1 below. The author uses the term "commonly used" to describe these metrics, which are used by at least three researchers out of the eight found in the literature survey. These researchers alphabetically are Bishop (1970), Christie {this author} (1988), Gajar (1988), Johnson (1996), Larkey (1998), Page (1966a, 1966b, 1968, 1944), Slotnick (1972) and Whalen (1971). For further details of the metrics used by these eight researchers see Appendix A: Table of Metrics for Style Marking.

It was very noticeable that one particular metric was always used. This is the metric termed "average sentence length". The author wonders why indeed was this metric used by all eight of the researchers. This author is not in a position to answer this yet.

He further ponders if this metric should indeed be used for all automated marking of essays for style.

The literature search revealed that only essays written in English were used. Although it is true that many different essay sets have been used by these eight researchers, there is no evidence in the literature that any non-English language essay sets have ever been used. However this author is fully aware that non-English language essays are also manually marked for style, and for content. There is a strong presumption held by this author that the automated marking of essays for style and content is just as feasible in French, German and other ASCII character based or Latin character based languages as it is in the English language. In Chapter 6 the author returns to the topic of non-English languages.

Only ASCII files, in other words plain text files, were used and no word-processed essays were ever used. While this is understandable for those researchers who were working in this field in the period before word-processing became common place, it does not explain why later researchers subsequently appear to use only ASCII based essay files.

Finally it was noticeable that all algorithms reported results as good as, if not better than, human markers.

<u>The questions raised are:</u>

The facts revealed above led to the author formulating several challenging questions, the answers to which are of great importance to the future of automated marking of essays for style.

One of the first questions which came to the author's mind was whether it was possible to establish a "common set" of metrics from those used by previous researchers. This appeared a possibility especially bearing in mind that one metric was already in common use, and an average of 22 metrics was found to be deployed across all eight researchers. There was a high degree of commonality within the 75 metrics used in total by these eight researchers. This is shown in Table 3.4.1 below.

Frequency of Occurrence (used by the 8 researchers)	Number of Different Metrics, with that frequency	Cumulative Total
8	1	1
6	2	3
4	2	5
3	12	17
2	11	28
1	47	75

Table 3.4.1: Frequency of metrics in automated style marking

This question gave rise to other questions. If it were possible, then:

- How universally applicable would such a common set be?
- What would be the implications and ramifications of using word-processed essays?
- Were word processed essays to be used what, if any, common set metrics could arise?

It should be realised that when word-processed essays are considered then additional metrics for marking style are possible. Examples of such metrics would be Font types, Font Sizes, Colours, Text enhancements, and so on. These are metrics which are not found in essays held as ASCII files. The inclusion of these word-processing based metrics strongly reflects the power of word-processors in making text more readable and more presentable, thereby impinging on style style.

The author has generated a list of numerous metrics that could be considered as a pool, or bank, of metrics to be employed in the marking of style. This list is found at Appendix B: Table of Potential Metrics for Style Marking.

The author has developed software to utilise what appears from his research to be the "common set" metrics. This software could be expanded to include some or all of the objective metrics listed in Appendix A and Appendix B, as well as that additional metrics yet to be devised, or discovered. To include metrics arising from word-processing use will require substantive development of the software.

It is regrettable that the author found no style-marked essays available to be used in this project. The consequence of this is that this branch of the author's current research is in suspension. Although future research and development is certainly possible, and indeed desirable.

The author has created two appendices on metrics that underpin both current and future research work on automated style marking. These appendices occur at the end of the thesis as Appendix A and Appendix B.

Appendix A – This is a summary of eight research workers who have published style marking metrics. This appendix is in three parts, organised on the basis of frequency of use by researchers. The three bands of frequencies are those greater or equal to three, those equal to two and those equal to one.

In considering the first part the 17 metrics that have been used three or more times by the researchers, one metric, namely "average sentence length", is used by all eight researchers mentioned in this appendix. Of the 17 metrics listed three are average word length, average sentence length and average paragraph length and two are the standard deviation of word length and the standard deviation of sentence length, all the other metrics being simple counts.

In considering the second part, the smallest of the three parts, lists the 11 metrics which have been used exactly twice by the eight researchers, one is a calculation, the Flesch formula, and the others are, again, simple counts.

In considering the third, and final part, lists the remaining 47 metrics which have been used exactly once by each researcher, there is an increase in the variety of type of metric. There are five formulae, the Fang, Fog, Gunning, Lorge and Winnetka. There is one percentage, the polysyllabic words. One standard deviation is listed being that of paragraph length. All the rest are either simple counts or simple frequencies.

Appendix B – The range of possible metrics that could be used for marking style.
This appendix both compliments and enlarges on the Appendix A. Its purpose is to list all the metrics that the author has found or devised. This listing should be viewed as being the current high-water mark of possible metrics, and should not be considered as being exhaustive. After all, as mentioned earlier, Page is reported (DeLoughry, 1995) to have been considering some 4,000 metrics!

A "Common Set" of Metrics for the marking of style.

This author (Christie, 1998) is contending that there exists a "common set" of metrics for automated style marking. This set of metrics may be based on, or about, the 17 metrics that are listed in the first part of Appendix A. For their inclusion in the "common set" then the deployment of these 17 metrics must result in a level of style marking performance that is at least as good as experienced manual markers.

The concept of having a "common set" of metrics, however, remains not-proven and will remain so until a range of previously marked essay sets are available for the author to further his research.

However the three-stage approach the author is intending to take with the metrics is to consider all the metrics, either gathered from the literature or self devised, known to him. These metrics are listed in Appendix B.

After several essay sets have had a weighted linear model produced for them the first stage is to rank the contribution each metric makes to the weighted linear model. This stage will use all the metrics for each marked essay set and will use statistical procedures to produce this ranking.

The second stage is to progressively truncate the model till the onset of the failure to match the performance of manual marking for each of the essay sets.

This is followed by the third stage which is to examine all the truncated models in order to ascertain the metrics in common use. These metrics are then ranked to produce a proposed common metric set for marking style. Once the common metric set has been identified then the author will remark the essay set using this particular model and hence seek to determine the performance of the common set against the performance of both the manual marking and the performance from using all the known metrics. Topping up the set of common metrics with additional metrics drawn from the next common metrics to achieve all-round performance may be required.

The author suspects that a common set of metrics may be produced in this way. If this suspicion is correct, the author fully expects that for each essay set the coefficient for each metric will not be the same value. Therefore the need for 'training' the weighted linear model will never be removed, but the number of sample essays required would be fixed at twice the number of metrics in the common set, whatever that number turns out to be.

In terms of presenting the style marks from the SEAR software system the author intends to mimic the format of the SEAR reports for content. Appendix E is an example of the reports for content. In the reports for style marking the 'Usage' and 'Coverage' elements would be omitted, as they are irrelevant for style marking. In any case the whole essay would be considered for the marking of style. The term 'entity' would be replaced by the term 'metric'. The facility of exporting the style report into a spreadsheet would also be made available.

The export of the reports for style allows the examiner, and other appropriate staff, to further process the results. The capability of sorting the individual essays into some order would permit both feedback to both marking and teaching staff and the detection of academic misconduct. Plagiarism is a rising scourge in the academic environment, and thus, by sorting the style reports (or for that matter content reports) then those essays exhibiting a highly similar pattern of metrics (or entities) flag potential candidates for some manual investigation.

As a conclusion to this sub-section the author presents a problem to the reader in the form of a diagram. This diagram, Figure 3.4.1 is shown below and opens up the question of the optimum number of metrics required for automated marking of essay style.

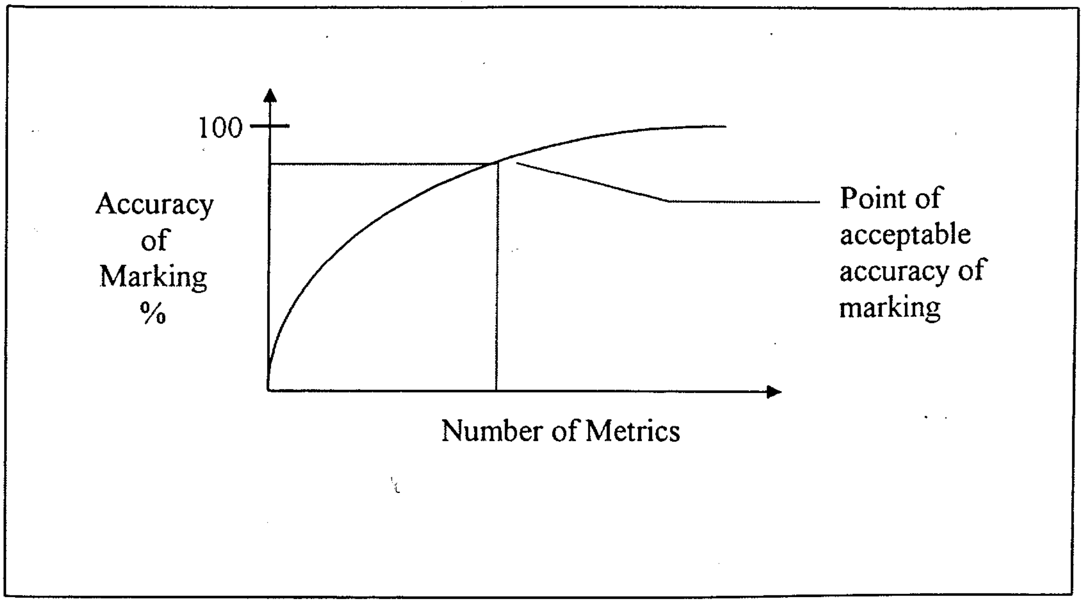


Figure 3.4.1.b: What number of metrics to use?

As the number of metrics used in the style marking algorithm increase then there should be an increase in the accuracy of marking style. This raised a question of marking performance for this author. What is the acceptable level of accuracy for marking, both manual and automated?

The easy answer for automated marking is that is must at least be as good as manual marking. This answer is naïve as is the question. The problem is with manual marking. How does any marker determine their own accuracy of marking? Who would set the level of acceptability? The author would find it incredible for any marker to announce at an examination board that his of her marks are not accurate.

Setting the performance for manual marking to be perfectly accurate is an expectation, but is it a reality. To ensure that each and every single mark is perfectly accurate would take considerable time and effort, both of which are in short supply in the current academic environment. Once there is an answer to this question for *manual marking* then the actual metrics needed to achieve that specific level of performance for manual marking will be rapidly determined for automated marking to match manual performance at least.

3.4.2 Development of content algorithm

The literature survey identified two different approaches to the development of an algorithm relating to the marking of content in essays. One approach is developed by Alliot and the other approach developed by both Landauer (the Intelligent Essay Assessor {IEA}) and ETS (eRater software developed by Educational Testing Services {ETS}, USA). The reporting of these approaches are possibly not detailed enough to have the underlying algorithms. The approach that was taken by Alliot (Alliot, 1994) is designed for use with short phrases or sentences. The other approach was taken by eRater and IEA (ETS; Landauer, 1998 respectively) and was used for extensive essays.

Alliot's approach is based on a simple data structure operating on a range of small text, best regarded as objective short answers, that is short phrases of free text usually in the form of a short simple sentence.

IEA and eRater are both based on Latent Semantic Analysis (LSA), where many texts are analysed for their word association. This association is then used to mark essays. However, this approach lacks operational flexibility as it is predicated on the substantive text analysis required to develop the word association.

3.4.2.1 How content algorithm was tested

The content algorithm was developed and tested in two main phases. The algorithm necessitated the construction of a data structure the purpose of which is to hold the content schema. This data structure was virtually completely developed during the first phase. This was a necessary precursor to the development of the initial content marking algorithm.

It was decided to use two unmarked essay sets as well as several "independent" pieces of small unmarked text during the first phase. Dr R A Butler provided an essay set comprising a suite of five essays based on the history of Robert Gordon, the founder of Robert Gordon's College and the Robert Gordon University both located at Schoolhill Aberdeen. Ms S Earl provided the second essay set comprising a pair of essays based on the early history of train transport. The independent pieces of small text range from the trivial "The cat sat on the mat." to small paragraphs. The emphasis during this first phase was the development of the underlying data structure together with the development of the outline content marking algorithm rather than for the full testing of essay sets. Clearly, there had to be two-way interactive development of the content data structure and the outline content marking algorithm, as neither could be developed in isolation from the other.

During the second phase the unmarked essay sets previously used in phase one were used to fully test the content marking algorithm. These essay sets were augmented by marked essay sets. Marked essays sets were sourced from the Robert Gordon University, which provided 10 essay sets, and from the Open University, which provided a further two essay sets. The main testing of the algorithm was performed using the marked essay sets. As a result of the process of testing, the content data structure, the content marking algorithm and the combination of the two were finely tuned in this phase. In order to further verify marked results two independent human markers were employed to additionally second mark some of the essay sets provided by the Robert Gordon University. The two independent experienced second markers were Ms Caroline Norton (a lecturer at Aberdeen College) and Dr. Mohammad Sharif (a lecturer at Napier University). The results obtained from computer marking, and first (and where it occurs second) human marking are found in Chapters 4 and 5. The agreement, or lack of agreement, between the first and second markers is also discussed in Chapters 4 and 5 of this thesis.

3.4.2.2 The format of the content data structure

The format of the content data structure is given in Table 3.4.2.2 below:

ID	Element	Description
1	EntityID	Entity reference number
2	PartID	Question part / sub-part identification
3	EntityType	Type of this entity
4	Detail	Actual content information or relationship type
5	ParentID	Entity number of the parent entity
6	Child-A	Entity number of first child
7	Child-B	Entity number of second child
8	Child-C	Entity number of third child
9	Child-D	Entity number of fourth child
10	Number	Roget's topic number
11	POS	Roget's part-of-speech
12	Threshold	Value to be attained before awarding mark
13	Mark	Marks to be awarded
14	Touched	Used to record when this entity is matched
15	FeedBackID	Feedback reference number

Table 3.4.2.2: Format of the content data structure

Description of the elements of the content data structure

Each element of the content data structure is described in the following paragraphs. The term "entity" is used to refer to a row of information of the content data structure, where each entity is made up of the 15 elements listed in the table above. When expressed using this data structure the content schema is expressed in a format, or structure, that computer scientists would recognise as a linked-list.

EntityID

This element is purely used as a reference number to identify each entity in the schema.

PartID

This element is used to hold the identity of the question part or sub-part as devised by the examiner. For example, this element may hold values such as 1, 4a, 9ci or 9c2. This identity is used in the reporting of the marks to the member or members of staff and to the individual student.

EntityType

This element is used to indicate to which of the six entity types this particular entity has been assigned. The six types of entity are in two groups: data and relationship. In the data group the only code is "f" which represents "fact". In the relationship group the remaining type codes are "a s o e i" which represent, in order: And, Simplex And, Or, Exclusive Or, and Not.

Fact: This is a piece of data held as a single word. For example "blue", "123" or "Christie". There is a provision for any fact type entity to have up to four child entities. However, the author currently does not use this provision but the provision is left for either revisiting the current usage regime or for future work.

And: The presence of this code requires that all the child entities are present, not necessarily in order. For example And(James Christie) would match "James Christie" and "Christie James".

Simplex And: The presence of this code requires that all the child entities are present, but in the given order. For example Simplex And(James Christie) would match "James Christie" but would not match "Christie James". "Simplex And"s are also used for storing quotes in the content schema.

Or: The presence of this code requires that at least one of the child entities is present. For example Or(James Jim) would match "James" or "Jim" or both.

Exclusive Or: The presence of this code requires that at least one of the child entities is present, but not both. For example Exclusive Or(James Jim) would match "James" or "Jim", but would only count one of these options.

Not: The presence of this code excludes the given detail. For example Not(Jim). This would prevent "Jim" being accepted.

Detail

This element holds either a copy of the relationship entity type, that is it holds "a s o e n"; or a single word at present. As a potential future development of the SEAR software this may change to short phrase of a few words at most.

ParentID

The author uses this element to hold the linking relationship between the various data entities. The ParentID element is used to identify to which entity this particular entity is upwardly linked to. There is the use of a special ParentID value, that of "0". At any top-level of the schema the value assigned to the ParentID is "0". Therefore in a content schema in which there are several parts (or sub-parts) there will be many ParentIDs that are set to "0".

Each entity will have one and only one parent. There must be at least one child entity present for any entity to be assigned as being a parent. As a maximum, only four child entities are possible.

Earlier paper design exercises, conducted by the author, lead to the choice of having only one parent entity. With no parent entity the required data entity linking would be impossible to create. Using more than one parent entity led to two issues that had to be addressed. The first issue was that very often the second (third and so on) parent entity was not required, that is it was redundant. The second issue is that when the second (or more) parent entity was occasionally used it led to an overcomplicated content marking algorithm. Most of the entities present in the linked lists used in computer science have one parent, and generally two child entities.

Should the need arise the decision to have a limit of one parent may be revisited. To incorporate a second or more parent entities should be a straightforward technical exercise.

Child-A, Child-B, Child-C, Child-D

In the author's design up to four child entities can be associated with any one parent. When there are not enough child elements the un-used ChildID value is set to "zero". It was through the construction of various paper design exercises that the author decided to use four child entities in this data structure.

If fewer than four child entities were used then the author found that more linking parent entities were required, thereby raising the complexity of the linked list holding the content schema without any increase in performance. In using more than four child entities, the author found that very often there was redundancy in using the larger number of child entities and the extra complexity involved in this did not enhance software performance.

Should the need arise the decision to improve a limitation of four child entities may be revisited. To incorporate more child entities should be a straightforward technical exercise. If the decision were made to use fewer child entities the easiest technical solution would be to assign "0" to the un-wanted child entities with no change to the existing content marking algorithm. Otherwise it would required to modify the data structure and hence to modify the existing content marking algorithm.

Number

This element holds the Roget's topic number that has been assign to the detail. The range of values is from 0001 to 1000 inclusive. In addition there is use of a non-Roget topic number, that of 0000. The use of this value is for those details that are not to be assigned a Roget topic number. Proper names, for example the word "Christie", are an example where producing a Roget topic number assignment would not be appropriate.

POS

This element holds the Roget's part-of-speech that has been assign to the detail. The values that this element may hold are N, V, D, J and *, where N represents Noun, V represents verb, D represents adverb and J represents adjective. The value "*" was designed by the author to further indicate that the detail element holds a value not assigned a Roget part-of-speech.

Threshold: There are two types of use depending on Entity Type.

Threshold - Data Entities

This is the minimum value that has to be reached before any marks are awarded for this part of the content schema.

Threshold - Relationship Entities

Depending on the relationship type of the entity this value takes a varying special significance.

Mark: There are two types of use depending on Entity Type.

Mark - Data Entities

This element contains the maximum mark that is to be awarded for this part of the content schema. The examiner assigns this value.

Mark - Relationship Entities

Depending on the relationship type of the entity this value takes a varying special significance.

Touched: There are two types of use depending on Entity Type.

Touched – Data Entities

This is used to record when this entity is matched as the essay is being marked. The number of entities flagged will be used to calculate the percentage of schema covered by the essay when the marking of that essay has been completed. The initial value is set to "N". This value changes to "Y" when there is a match between the entity and some part of the essay that is being marked.

Touched - Relationship Entities

Depending on the type of the entity this value takes a special significance that varies.

FeedBackID: There are two types of use depending on Entity Type.

FeedbackID – Data Entities

When the provision of classified or structured feedback is made operational then this element will provide the link between essay and content schema. The author intends to develop this technical aspect after the main research project is finished. The provisional view of this technical development is that the author will create a numbered list of classified feedback comments, which would be generated by the survey and analysis of feedback comments already being used in essay marking.

FeedbackID - Relationship Entities

Depending on the relationship type of the entity this value takes a varying special significance.

The author took the deliberate decision that the format of the content data structure would be flexible enough to facilitate change as this research project developed. This in-built flexibility was required so as to avoid the need to stop development of the content marking algorithm and hence to develop from nothing or very little again. The current format of the data structure could undergo a fair amount of alterations as the need arises. New elements may be added, or existing elements re-formatted or removed. All alterations to the data structure should be viewed as being more changes of a technical nature rather than requiring a severe re-design. The manner is which the software package SEAR has been designed and coded also has the flexibility to be changed easily.

This author intends to keep the flexibility of content data structure as an ongoing characteristic in any work undertaken after this research project finishes. When the research recommences on the style-marking algorithm the flexibility already in the style-marking algorithm will also be retained and maintained. The same approach to flexibility also extends to include the SEAR software through the design and coding phases.

This data structure is linear in nature, although it is not strictly linear in true mathematical terms. It is linear in the sense that the bigger the content schema then the larger the number of entities required to hold the content schema. Adding some extra parts to the content answer schema will result in an increase in the number of the entities in the content data structure.

The author has set 475 as a maximum limit for the number of entities in the content data structure. This maximum is set for the current configuration of the computer hardware hosting the SEAR software, which has 24MB of RAM memory. The largest number of entities used in the essay sets in this research project is in the range of 300 to 350, which is well short of the artificial limit of 475. Should it be necessary to raise the maximum then either extra RAM memory would need to be found or it would be necessary to convert the method of holding the data structure from a RAM based memory storage to storage on a magnetic storage on a hard disk. The change from RAM memory to hard disk would be achieved by the means of a technical conversion. Post conversion the maximum number of entities would be limited by the available space on the hard disk. The conversion from RAM to hard disk will result in a decrease in throughput for SEAR.

Appendix C has some examples of abridged content data structures. Considering the examples given in Appendix C the ease of changing the information in the content data structure should be readily seen. Changes of information in the content data structure may arise from the essay examiner requesting such changes or from any comments made by the second marker or external examiners.

Appendix D is an example of how the author intends the essay examiner to set up the content schema. This example illustrates how the content schema covers the identification of parts and sub-parts, and shows the allocation marks. The rows represent related facts and tabbing of the facts indicate dependence between the rows of facts. The author intends that essay examiners will use the free software package called Notepad, which is provided by several different versions of Microsoft Windows ™®.

This layout is easy to create and easy to alter. The author intends, post PhD, to automate the creation of the content data structure from the content schema should SEAR ever become an operational piece of software.

3.4.2.3 The process of creating a content data structure from an answer schema

The manual process of creating the content data structure from the given content answer schema is a relatively straightforward process. There are several stages in the creation process. Stages 1 to 4 may be accomplished on paper or may be input directly using the computer, although the paper version offers the greater flexibility is use for drafting and re-drafting. The six stages of this process are detailed below.

The **first stage** is to begin with the overall structure of the answer schema and create the relationship entities and data entities required supporting the answer schema overall. In particular part(s), sub-part(s) and so on are allocated space on the content data structure.

The **second stage** begins with the consideration of the first item on the answer schema and creates the required data entities and relationship entities to support the first item. The second stage continues by considering the remaining items of the answer schema, until the answer schema has been covered. The second stage is recursive in nature in order to cover all the possible valid data and all the possible valid relationships. It thereby covers all the fine granularity of the answer schema. Stop or function words are not considered as adding value to the content data structure and are therefore deliberately excluded from the content data structure.

The **third stage** is to assign an identity number (called EntityID) to the entities.

The **fourth stage** is then to assign Roget's topic number and part-of-speech for those entities for which such assignments are appropriate, such as data. Where there is no Roget topic number that can be assigned then the topic number is set to "0000" and the part-of-speech is set to "*". All the entities that represent relationships have their topic numbers and parts-of-speech set to "0000" and "*" respectively.

The **fifth stage** is to express the paper version as a computer-based version. This stage is not required if the first four stages were accomplished on the computer directly.

During this stage the special coding required for the various relationships are entered into the appropriate entities' elements of Threshold, Mark and FeedBackID.

The **sixth, and final, stage** is to test the supplied model answer against the completed content data structure. All the adjustments that are required to maximise the fit between the model answer and content data structure are undertaken during this stage.

3.4.2.4 The process of marking content

The process of marking content is simple and straightforward. Because the extracted text of the essay is held as one sentence per line, each sentence is processed individually in turn.

At the start of marking any essay the appropriate variables have their values reset to their default or original values. These variables are illustrated by counts of number of words in the essay, that is the Touched elements of the content data structure. This ensures that each essay is free from any memory effects arising from any previously marked essay. Each sentence is processed by first identifying, and then ignoring, the stop or function words as these words add little value to the content. A count of stop words is maintained to aid diagnostics later on. The second step in the process is to match entities, both the data entities and the relationship entities. The last step in the marking of the essay occurs when all the sentences have been processed. This step awards marks to the essay, determining its the coverage and usage, and so on, and at the same time records these various values. The overall final step is to update the running totals that are being kept. These totals are for the number of essays marked in this session of marking for this particular essay set and so on.

Each of the main steps is described in more detail in the next three paragraphs.

3.4.2.4.1 Stop (Function) Words

Every word in the sentence under investigation is checked against the list of stop or function words. When a match is found, the total number of stop words found in this essay is increased and the stop word is no longer considered in content marking. The word is not deleted from the essay, it is merely ignored from any further consideration. The purpose of removing the stop words from any further consideration is that stop words do not add any significant value to content. Stop words are not included in the content data structure for exactly the same reason. By ignoring stop words there is gained an incidental operational advantage as the throughput rate of essay marking is increased.

The deliberate choice made by the author to ignore the stop words echoes similar choices make by many web-based search engines and versions of library catalogue software for exactly the same reasons.

3.4.2.4.2 Matching data entities

The remaining words in the sentence are then matched against the content data structure. In particular data entities are considered as these entities form the basis for determining relationships. When a match is found between the word in the sentence and the Detail element of the content data structure then the Touched element of that matching entity, or entities, is set to the value "y". The term "match" covers both exact detail match and, just as importantly, the match between sentence word and the Roget equivalent. By "Roget equivalent" is meant that the sentence word matches a word held in the Roget wordlist that has the same topic number and the same part-of-speech as the entity. Where there is no Roget topic number assigned in the entity then there is no part-of-speech present as well and no match is attempted. The count of essay words used is incremented every time a word has been matched. The count of words used is the basis for the usage figure that is reported later for each essay.

3.4.2.4.3 Finding relationships

Once each sentence has had its words matched then it is the turn to match relationships. The content data structure is matched from the last entity to the first entity. This is because the content data structure is designed and implemented in a top-down linked structure.

If a detail entity has been matched then the entity's Touched element's value of its Parent entity is also set to "y" and the Threshold element's value of the parent entity is incremented.

If a relationship entity has been matched then the Threshold element's value of its Parent entity is incremented. For a relationship entity if the value held Threshold element is equal to, or greater than, the Mark element then the Touched element value is set to "y".

The diagram, Figure 3.4.2.4.3 below elucidates the relationship between the essay, its sentences and the content data structure.

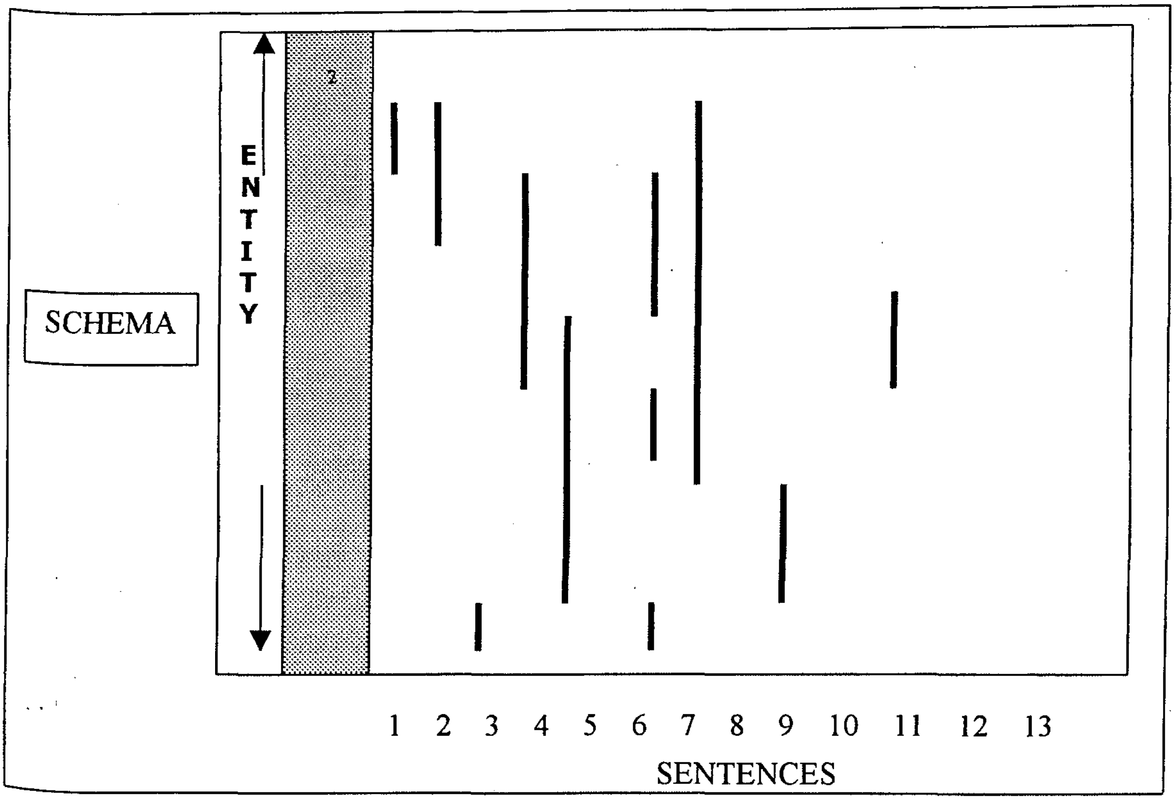


Figure 3.4.2.4.3: Entity versus Essay Sentence

Imagine that the above diagram is a visualisation of the essay after all its thirteen sentences have been processed against the content data structure.

Only a few entities have not been matched by any sentences in the essay, in particular entity number 2 has not been matched. So a high level of schema coverage would be reported.

Sentences number 8, 10, 12 and 13 were not matched with any entities. Therefore these four sentences will make no contribution to the awarding of a mark to this essay. Nearly all the sentences overlap their mapping to the content data structure. Because of this a medium level of essay usage would be reported.

3.4.2.5 Reporting the marks awarded

After each essay has been marked there is a scan from the first entity of the content data structure to the last entity. For each and every entity that has its Touched element set to the value "y" then the marks allocated are added to the running total of marks for that essay. Should there be a sub-part or sub-parts in the answer schema then those will be reported as well as the overall mark.

This report, called the Content Report, also the Coverage and Usage both expressed as percentages; the number of words; and the number of sentences. The two terms of Coverage and Usage are explored in more detail in sub-section 3.4.2.5.1 below. All this additional information gives the examiner, at one opportunity, the ability to see how each essay has been marked in terms of performance and how the awarding of the actual marks has been accomplished across the whole of the answer schema.

A set of additional statistics is produced for the essay set as a whole. These statistics are the number of essays in the essay set, the number of essays marked, the date and time processing started and the date and time processing finished. This report is appended to every time that the relevant essay set is marked for content. By appending the new results and statistics to the previous report for an essay set then all late submissions are shown as being marked out-with the original batch. At the same time appending allows the examiner to check for continuity across the separate processing of the essays. Should an essay set need to be remarked then all late submissions will be combined with the on-time submissions and thereby will be remarked in one batch. The administration of essay submission will continue to show those essay or essays that were late submissions.

A second report is also prepared during this phase. This second report details how entities in the content schema structure have been flagged as being 'touched' for each essay. A 'touched' is shown by the letter 'y', otherwise the character '-' is shown. As well as showing the 'touched' entities, for those relationships that are defined as being of the type "Simplex And" the state of the relationship is shown, by the use of a numerical code juxtaposed after the letter 'y'.

This second report, called the Schema Report, is used to provide feedback to both the student and the examiner. As each entity is listed for each essay in a grid format, then examiner feedback is achieved by using the vertical columns and essayist feedback is achieved by the using the horizontal rows.

This report is also appended to on the processing of any late essay submissions.

Sample Reports

A sample copy of both these two reports are given in Appendix E, illustrating a very simple example of each report. Although the author has not displayed any SEAR based reports for style, the reader is invited to envisage that the reports of marking style will, with a few alterations, mimic the reports of marking content. These alterations are minor – the items of 'Usage' and 'Coverage' are omitted and all the mention of 'entity' is replaced by 'metric'.

Exporting Content and Schema Reports

These two reports may be imported into Microsoft Office™® products for further analysis to satisfy any whim the examiner may have. Microsoft Excel™® may be used for sorting the essay results. Microsoft Excel™®, Microsoft Word™® and Microsoft Notepad™® may be used to print these reports, or these reports may be subsumed into any larger document.

3.4.2.5.1 Usage and Coverage

The author (Christie, 1999) has coined two new terms in the process of marking essay content. The two terms are "Coverage" and "Usage". These two terms are complementary and were coined to give an indication of the performance of the marking of the essay with respect to the content data structure. When considerably more work has been done to make the software more operational then a better feel for or understanding of the significance of the values of coverage and usage would be expected to arise. This is mainly the effect of experience of using the software.

The term 'Coverage' is used to describe to what extent the content data structure has been matched by the essay. High coverage indicates that the essay contained a sufficient number of both appropriate data and relationships such that a high amount of the entities of the content data structure have been 'touched'. Low coverage could mean a number of different things. The first meaning is that the essay is a poor essay in terms of content. There is little relevant content in the essay that could be matched against the content data structure. The low mark thus awarded to such an essay is justifiable regardless of marking methodology. The next meaning has very serious ramifications – it could be that the essay is a superior essay. A superior essay in terms of content would mean that the essayist has used data and or a relationship or relationships not expected in the answer schema provided by the examiner.

For an example of the latter see the example of a superior answer given in the middle of sub-section 2.4.2.3 in the previous chapter. In the superior essay the essayist may use more precise data than what the examiner expected. Likewise the essayist's use of special terms or highly sophisticated relationships may make the essay superior.

The term 'Usage' is used to describe the proportion of the essay that has been used in the marking for content. High usage indicates that most of the essay has been used as a basis for the mark that was awarded. Low usage has the same possible interpretations as for low coverage described in the previous paragraph, namely the essay has a poor level of appropriate content or the essay is superior to what was expected.

In addition to examining the coverage and usage terms on their own, there is a possibility of examining the combination of these terms. To illustrate the relationship between coverage and usage, consider the diagram in Figure 3.4.2.5.1 below.

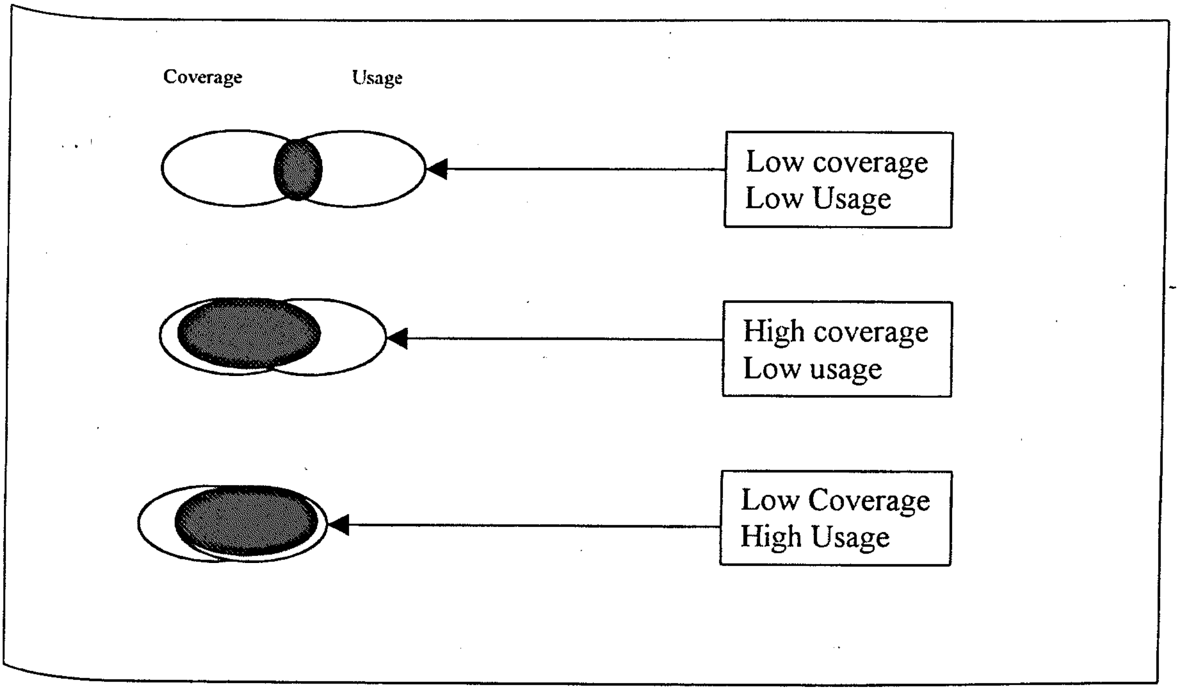


Figure 3.4.2.5.1: Relationship of Coverage and Usage

High Coverage – High Usage

This combination would be generated by an essay that could be considered as a specimen essay or model essay answer. Little improvement in the essay content is possible.

Note that the manner of how the marks have been allocated in the answer schema is reflected in the actual mark awarded. For example suppose there is one critical detail in the answer schema that has been allocated 10% of the overall mark, and this is expressed in a few entities in the content data structure that has, say, 100 entities. Then if the essayist misses this one critical detail ensuring a coverage of 98%, yet the mark awarded will only be 90%. Change the marks allocated to this one entity from 10% to an extreme of 50% (and, of course, making other relevant, related changes) then the coverage would still be about 98%, but the mark awarded to the essay would fall to 50%.

High Coverage – Low Usage

This combination would represent an essay that should be awarded a very good mark. But this is not a perfect essay. To have low usage then the essayist has produced an essay with a number of problems, where these problems may occur singly of in some combination.

These problems are listed as being –

1) The essay is full of extraneous material,

2) There is much repetition or re-statement of the same facts, or

3) The essay has included too much of what is colloquially called waffle.

An example of each would be –

1) Including information on magnetic disks in an essay supposed to be only on magnetic tape,

2) "The cat sat on the mat. Yes, indeed the mat was sat on by the cat."

3) "The black and white cat demurely and quietly reposed on the chequered blue and red worn carpet." for "The cat sat on the mat."

Low Coverage – High Usage

This combination is typified by the shortsome essay. What is in the essay is correct and most of the essay is used to produce the marks awarded. However, the essay is somewhat deficient in length, not in terms of word volume, but in content. For an example of this problem consider the essay "The cat sat on the mat.", when "The black and white cat demurely and quietly reposed on the chequered blue and red worn carpet." is expected.

Low Coverage – Low Usage

This combination is the most problematic of the four. In the author's experience an essay exhibiting this combination could either be a very content poor essay or it is a superior essay. No halfway position is thought possible. Unfortunately the overwhelming balance of probability is that the essay will be content poor. In the author's experience of the marking of hundreds of essays over two decades it is very rare, but exceedingly welcome, to encounter a superior essay. The SEAR content algorithm unfortunately does not yet provide the sensitivity sufficient for the discrimination between these two possible types of essays. An essay exhibiting this particular combination would require the marker to manually mark the essay. This human intervention is required to make sure that the essayist is awarded the correct marks.

Throughout the author's experience he has not, yet, encountered the third type of essay that could be classified as having a low coverage – low usage combination. This third type of essay is referred to as the "bad faith" essay. The author is not personally aware of any colleagues who have received any "bad faith" essays.

"Bad faith" essays

There is always the possibility that a "bad faith" essay has to be marked. A "bad faith" essay is an essay in which the essayist had deliberately created an essay that has the key words present but present in a nonsensical manner.

There are two essays in Appendix F, both prepared by Dr R A Butler for the author. The first is what could be considered an acceptable "on topic" essay on the topic of "Who was Robert Gordon?" The second essay in this appendix is an example of a "bad faith" essay. An essayist is not likely to produce a "bad faith" essay unless they are cold-bloodedly seeking to have a re-sit or want to deliberately waste their time and as a result of this their marker's time.

3.4.2.5.2 Marks awarded

Both the Content and Schema reports generated by SEAR as described are intended to give the maximum amount of simple information. This information is provided for use by the examiner, the essayist, and all additional markers in order to base auditing and other aspects of quality assurance.

As the results produced by the SEAR package may be imported into a spreadsheet, the examiner may then further process the data into whatever format is required.

It is not just the marks that are reported. The mere reporting of the marks awarded to each essay in the essay set is not acceptable to the examiner and the essayist as this provides no useful data on which to base feedback. Feedback needs to furnish answers posed by the teacher, the examiner and the essayist in order to improve performance by all these people.

There is a well known set of three questions that is commonly used in many such management situations, especially when dealing with quality management and / or quality improvement issues. These are: "What went right? What went wrong? What to do next time?" These same three questions are what the author uses for his own self improvement and he also suggests to his students that they ask themselves the same three questions in order to improve their own performance. The author is aware, and fully accepts, that the reader may have his or her own methodology for initiating performance improvement.

3.4.2.5.3 Providing feedback

The author regards the award of any marks as a statement of performance or a statement of achievement or of attainment. This is useful information to the candidate, the education provider, the prospective employers or other stakeholders. Equally the author regards the award of marks as sterile in terms of performance improvement, especially when the diagnostic and formative modes of assessment are being used.

The provision of feedback is required, as it is the basis for the introspection necessary for performance improvement. No introspection means no discovery of deficiencies, problems and faults and hence does not lead to their removal. This lack of introspection also means no route to the generation, hence evaluation, of alternatives. Further there is no opportunity to determine performance against standards. No introspection also results in the non-identification of what exemplary actions or activities have taken place.

The author regards feedback as the true origin for feed forward. Just having feedback does not, in itself, enable performance improvement unless the feedback is a basis for taking action.

3.4.2.5.3.1 Feedback to the examiner

The examiner's feedback shows how each part of the answer schema fared. This covers examiners' questions such as which, if any parts, were always answered correctly, which were never answered, did the more able essayists correctly respond to certain key features present in the answer schema.

Of course the examiner feedback would not be solely limited to using the second or Schema Report, the first report, the Content Report, would play an important role too.

After the initial content marking process, the examiner may decide, either alone or in consultation with second markers or external examiners, to edit the original answer schema. The examiner has the right to modify the answer schema. If any modification is made to the original answer schema then there is software created by the author, as part of the SEAR system, to reset the reports and the marking of the essays to the state as if no marking had ever taken place. Hence a new content marking process may begin again. This reset and then re-mark may be repeated as often as required until the examiner is satisfied with the marks for the essay set.

The examiner must always have the reserved right to reject any or all marks if they are not satisfied. Equally there is an obligation on the examiner to ensure that marks awarded by the software are acceptable.

As the reports generated by the SEAR package are detailed rather than summary reports then the examiner has the facility to identify, and hence target, those essayists who are having problems with the content of the particular essay topic. This seemingly innocent facility dramatically changes the mode of giving feedback from being purely reactive (along the lines of the essayist seeking why they received that mark) to being active (the examiner seeking out the poorer performing essayists to give them extra attention).

Because the marks from the SEAR software system may be exported to a spreadsheet the examiner may sort the Schema Report (and for that matter the Metric Report) into a format that will highlight the performance of the selected parts of the content answer schema. A lightweight or high-level scrutiny would reveal which parts of the schema were always answered, or never answered, and the whole range between these two extremes. By deepening scrutiny to the level of the essayist and the essay it is possible to show the examiner how the good essayist and the poor essayist can be distinguished. This hierarchy of scrutiny somewhat mirrors the performance monitoring that is offered by automated objective marking in terms of item appraisal.

Plagiarism and other forms of academic misconduct

It is mainly from the Schema Report, but with some support Content Report, that there is the opportunity to detect plagiarism. Because the Schema Report shows how each essay matches with the answer schema then, almost by default, it should be relatively easy to use the same information to compare one essay against another. By exporting the Schema Report into Microsoft Excel™® it should be possible to sort the data into an order that shows any close pairing of the essays that may be present. Whenever present any such close pairing, or similar pattern, of entities provides the basis for considering that plagiarism may, stressing the may, be present. It is an obligation on the examiner to elect to exercise any follow up investigation of that group or groups of essays exhibiting too similar a pattern of entities.

Note that the Metric report that will arise from the marking of style will also have sufficient detail present in order that it may be used for plagiarism detection.

There is the opportunity for the detection of pairings to be automated by the use of software. The author is aware of a statistic called the Cosine Angle which is often used to determine how close two vectors are, and this statistic be may deployed on the data held in the Schema Report. Should the use of the Cosine Angle statistic prove to be useful in initiating suspicion of plagiarism then this statistic is relatively easy to automate into a separate report. A second statistic, the Horn statistic (Dolland and Mowrer, 1947), may be used to determine the closeness of essays to each other.

Other forms of academic misconduct may be detected from these two reports and from further developments in reporting facilitated by additional software.

3.4.2.5.3.2 Feedback to the essayist

The essayist's feedback shows those parts of the answer schema where the essayist was awarded marks, and where the essayist was awarded no marks. The feedback to the individual essayist would, of course, have the details of marks, coverage and usage contained in the first or Content Report.

Automated marking should provide the opportunity for the essayist to obtain rapid feedback. By obtaining rapid feedback, the essayist is thus enabled to take action to improve his or her performance much faster than is currently possible under the existing manual marking methods. If the time lapse between submitting the essay and obtaining feedback for that essay is too long then the essayist may have submitted one or more further essays while being ignorant as to their performance.

Too long a time lapse also results in the loss of freshness of the impact of feedback. Would the reader welcome detailed feedback on an essay they wrote several months ago? Would they even remember what that essay was about? The author suggests to the reader that the same feedback delivered within 'hours' of the submission deadline would have more impact on the essayist than a delivery some three weeks later. Most further and higher education organisations are using two to four weeks as the norm for the return of coursework submissions in general.

When the author increases the software's operational effectiveness the software will make use of the feedback element of the content data structure to provide an automated feedback response to the essayist.

Reverting to the suggestion that all essays may be marked for style, but only some for content, leads the author to suggest that the essayist could receive value-added feedback on an essay purely marked for content. This value-added feedback would take the form of giving the essayist feedback on their style as well as for the essay's content. As well as the reader, this author has frequently received essays (and project reports, for that matter) which attract high content marks but are at the same time poor in style.

If a common set of metrics could be found that are suitable for all automated marking for essay style then this author suggests that the summary feedback to the essayist could be both ranged and tabulated, as may be envisaged from the following table 3.4.2.5.3.2 below.

Metric	Recommended Range Minimum – Maximum	Your value
Average sentence length	xxx – yyy	zzz
2nd	. . .	. . .
. . .	. . .	. . .
Nth	. . .	. . .

Table 3.4.2.5.3.2: Standardised style feedback slip

Once a large series of essay sets have been marked for style then it may be possible for the creation of a feedback slip to have the range of 'recommended' minimum and maximum for each metric to be established. Here the term 'recommended' is used to indicate the range of metrics that an acceptable essay will be expected to contain. Using the term 'recommended' rather than the term 'absolute' should help to avoid any essayist becoming to focused on the metrics rather than the actual essay and its style.

Thus each essay's own metric values may be entered on the slip (by automated slip production) so that the essayist will rapidly establish exactly how their essay's style is in relation the common metric set, and more importantly in which area or areas the essay may be improved.

3.4.2.5.3.3 Entity feedback by specific essay

The author has created a piece of software that will display a specific essay entity report in a format that is easier for the examiner and the student to use together. By this simple software the essayist can readily learn where they lost marks as those entities that were not matched by the essay is question are listed out on the computer screen. The examiner should be able to match this list with the original answer schema and thereby the essayist is directly given the aspects on which they lost marks. Should the examiner deem it necessary then he or she may copy the displayed data onto a printable file or print the data directly.

The use of this piece of software should help to protect against the accidental revealing of another essayist's results when the examiner is providing feedback to one essayist as only one essay is examined at any one time.

3.5 Statistics used in this research project

The reason for using statistics is to analyse manual marking versus computer marking. Where second manual marking is available then the analysis is expanded to include manual versus manual marking, and both individually versus computer marking.

Hypothesis

The null hypothesis is that there should be no difference in the mark awarded to any essay regardless of how that essay is marked. This hypothesis applies to style and to content equally.

The author employs a range of statistics in this research project. The range is divided into two groups, where the first, and larger, group encompasses the statistics that are applicable to both style and content. The second group is for application in style.

This second group is included here to indicate what statistical techniques would be used in the application of the style algorithm. There are no actual results included in this thesis as the research of automated style marking is in suspension until manually marked essay sets are obtained for style.

The first group includes descriptive statistics, parametric and non-parametric statistics. The second group includes multiple regression and ANOVA.

Where appropriate graphs will be produced to illustrate the results of the analysis.

The author used Statistical Package for Social Scientists (SPSS) for Windows Release 10 for producing the statistics and graphs, in all cases except for Cohen's Kappa. The statistic known as Cohen's Kappa was produced by the author using Microsoft Excel™® since SPSS has some technical difficulty in handling occasional empty cells in the data.

3.5.1.1 Descriptive statistics

The statistics used include maximum, minimum, average and standard deviation. These statistics are used to generate a holistic view of the data. By plotting the raw mark data the general trend of one set of marks against the other is visible. The use of descriptive statistics sets the scene for the use of more specialist statistics.

3.5.1.2 Parametric statistics

The Pearson Product-moment Correlation Coefficient (more commonly known as "Pearson's Correlation", or even more commonly 'Correlation') is used here as it is the most commonly used statistic for measuring the degree of association between variables. The hypothesis stated in 3.5 above permits the use of the one-tailed version of this correlation statistic.

Pearson's correlation coefficient may only range in value from –1 through 0 to +1. The further the coefficient is from 0 the stronger is the association between the variables. A coefficient with a value close to 0 would indicate that there is no effective association between the variables. The coefficient only indicates the quality of the association between the variables. Negative coefficient values indicate an inverse association, that is as one variable increases in value then the lower the other variable's value becomes. Positive coefficient values indicate a positive association between the variables, that is, as one variable increases in value then the other variable will increase in value as well.

Pearson's correlation is sensitive to the context in which the data was collected. In, say a social science context a value of 0.2 to 0.3 could be considered, or accepted, as being good. However in an engineering context coefficient values of about 0.3 could be considered as being weak with the expectation that a value of 0.7 to 0.9 as acceptably good.

Often a straight-line regression is plotted and or calculated from the data. Pearson's correlation coefficient has no part in the formulation of the straight-line regression equation, but it does show the degree of fit of the regression line with the data.

3.5.1.3 Non-parametric statistics

The Spearman Rank-Correlation is used here for two reasons. The first is that it complements Pearson's Product Moment Correlation. A similar value of correlation from both Pearson's and Spearman's formulae for the same data would act as mutual confirmation of the degree of correlation that is inherently in the data. The second reason for its use is that it directly measures the association between the variables in the data by the ranking of the ordinal values in the data. The hypothesis stated in 3.5 above permits the use of the one-tailed version of this correlation statistic.

The range of correlation values and the significance of that correlation value from Spearman's formula are both directly comparable to Pearson's, as is the interpretation of that significance.

Jacob Cohen's Kappa (Cohen, 1960 1968) was used by the author as an additional measure of agreement. The basis for this statistic is that under certain assumptions there are two statistically relevant quantities, that of the proportion of items for which the two raters agreed and the proportion of items in which such agreement could be expected to arise by chance. The assumptions are:

- The items under analysis are mutually independent,
- The raters are mutually independent, and
- The rating system's categories are independent, exclusive and exhaustive.

Kappa can have a range of values. The maximum value is +1.00. This value of Kappa would arise when the two raters are in perfect agreement. As the degree of agreement between raters falls then the value of Kappa falls. When the proportion of raters' agreement matches, or equals, the proportion of agreement that would occur naturally through chance then the value of Kappa will equal 0. Kappa may go negative in value, and when this happens then the degree of agreement between the two raters is actually less than what could happen by chance. Negative values for Kappa are measuring active dis-agreement between the raters!

Both Pearson's and Spearman's coefficient may have a high value for the coefficient and yet there may be zero exact agreement between the two raters.

There are several references in this thesis where the author mentions that other researchers use the terms 'exact' and 'adjacent' when they refer to accuracy. This author uses the Kappa statistic as a more robust, and value-added, adjunct to the concept of 'exact' and 'adjacent' measure of agreement.

Although this statistic was specifically designed to measure the agreement between only two raters, Joseph L Fleiss (Fleiss, 1971) did extend the statistic to operate on six raters.

Although Cohen did not assign any labels to bands or ranges to values of his Kappa statistic, J Richard Landis and Gary G Koch (Landis, 1977) did assign labels to arbitrary specified ranges as in the Table 3.5.1.3 which follows.

Range of Kappa Statistic	Strength of Agreement (label)
< 0.00	Poor
0.00 – 0.20	Slight
0.21 – 0.40	Fair
0.41 – 0.60	Moderate
0.61 – 0.80	Substantial
0.81 – 1.00	Almost perfect

Table 3.5.1.3: Range Labelling of Cohen's Kappa

In the event Cohen's Kappa statistic was not used, as the spread of marks was limited to a few discrete values. Kappa is recommended for use when (nearly) all the NxN cells in the grid are filled.

3.5.2.1 Multiple regression

Multiple regression is an extension from the simple linear regression that is based on only two variables. There is still a dependent variable, or response variable, but with two or more independent, or explanatory variables. The value of the response variable is computed from a constant term plus terms derived from the different multiples of the explanatory variables. These multiples are known as the 'coefficients.' Essentially this coefficient is a means of showing how a unitary change in the explanatory variable will result in a change in the response variable, assuming that all the other explanatory variables and their coefficients are held constant.

There are a number of problems in using multiple regression. These are:

- Missing and or incorrect choice of explanatory variables,
- Calculation of the coefficients may be wrong,
- Co-linearity, and
- Incomplete explanation for the response variable.

An incorrect choice of variable or variables to explain the response variable may result in the calculation of the coefficients being incorrect. In missing one or more explanatory variables the same inaccuracy results.

Besides the incorrect calculation of coefficients due to poor choice of variables, the actual calculation may be incorrect due to the fact that a mathematical model is being built. In many situations there may be a large volume of experimental data, where each item of data is subject to error.

When a large number of explanatory variables is used in the multiple regression model then the problem of co-linearity asserts itself. Collinearity occurs when two or more explanatory variables are not independent and are in effect redundant variables. When co-linearity occurs then the redundant variables have to be identified and removed from the model. While eliminating known redundant variables is straight-forward, the identification of the same is not easy.

The last problem is that of incomplete explanation of the response variable. Ideally all the explanatory variables should totally explain the behaviour of the response variable. If this is true then the assumption is that the regression model is (probably) correct and safe to use, but not guaranteed to be 100% correct. If the explanation of the response variable is not totally explained by the explanatory variables then there is a difference between the actual response and the predicted response. The existence of this difference means that at least one of the other three problems listed above is occurring.

If any or all these problems occur then the multiple regression model is not safe to use.

3.5.2.2 Analysis of Variance (ANOVA)

When there are one set of marks produced from computer based marking and at least one set of marks produced from manual marking then the question arises of how well the two, or more, sets of marks agree especially the mean of each set. The hypothesis behind automated marking is that there are no differences about the means for the different sets of marks. This is the "Null Hypothesis". ANOVA permits the testing of the "Null Hypothesis" with a measure of how probable it is that the "Null Hypothesis" is true.

(Lack of) Style

Multiple regression and ANOVA are both for use in the style-marking algorithm. As has been stated elsewhere in this thesis the style part of the research project is in suspension these two statistical methods were not deployed.

Chapter 4: Experimental findings [for content only]

4.1 Test data

There were several developmental and ten test essays sets used in this research project, each of which is given a brief description in this chapter. The various essay sets used in this research project were obtained from a number of sources.

The first source was the author who created the following essay sets: test, CatMat, RG, BigBox and iNet. These consisted of several short pieces of text. The second source was the author's PhD supervisors Ms S Earl, who is responsible for essay set SE 1-2 and Dr R A Butler, who is responsible for essay set RAB 1-5. Both of these second source essay sets were used to develop the algorithms for style and content, and both of these essay sets were provided as un-marked essays. The third source was again Dr R A Butler but in his role of examiner of post-graduate and under-graduate students. These consist of the "3202" series of two sets and the "162" series of six sets. The fourth source was The Open University who provided the "OU" series of two sets. Each essay set will be outlined in the following sub-section.

The shorter texts that were used ranged from the sensible but trivial "*The cat sat on the mat.*" to The Robert Gordon University's Intranet welcome message, which is quoted in full below:

> "*The Robert Gordon University has in this new Millennium introduced as a learning support mechanism and teaching aid a university wide intranet.*
>
> *The Intranet Campus exists to provide students with access to learning support materials, anytime anywhere. In addition it is designed to provide a communications system.*"
>
> <http://inet.rgu.com> Accessed on 4th September 2002

Other less sensible pieces of text were also used purely to check software operations, such as "Big Box", quoted below.

There are balls and bricks in the box.
There are five balls which are blue.
Five bricks are coloured red and the other five bricks are navy blue.

For some of these texts the author assigned marks to the texts. In effect, these texts were treated as if they were model answers. The text of the Robert Gordon University's Intranet's welcome message was used in two modes. Firstly, it was used as one complete text and, secondly, its three sentences were used as three separate pieces of text.

4.2.1 Model Essay Answer

None of the development essay sets were marked for content by their originators, but without any marking schema(s). Thus in all cases the author adopted the method of constructing a model essay answer for use in each essay set in order to test the algorithm, the content data structure construction and the marking using this content data structure. The author created the marking schema for the development essay sets.

In each essay set the author constructed the model essay answer in such a way as to produce a mark of 100%, but this mark was only obtained if the content marking algorithm, the content data structure and the software operated as designed and as expected. The use of the model essay answer therefore acts as an internal calibration for each essay set. The practice of using a model essay answer was continued in the marked test essay sets for exactly the same reason. For the test essay sets, that is those essay sets where a marking scheme was already specified, the author had to construct the model answer such that it was very closely based on the given marking schema.

The author reasoned that should there be problems with the marking of the model essay answer then it would be a waste of resources to test the whole essay set until these problems were removed. The model essay answer was therefore tested in isolation from the other essays in that essay set before the rest of the essay set was used. Only when the author was satisfied that the model essay answer had been correctly processed were the essays in the set then processed.

In practice should a mark for a model essay answer not attain 100% the most likely reason is that the content data structure is not optimised. Altering the content data structure to attain a mark of 100% is neither underhand nor deceitful as many scientific and engineering practices require calibration or normalisation to function properly. In automated essay content marking the alteration of the content data structure is less problematic than the alteration of the content marking algorithm and subsequent alteration of the software delivering that algorithm.

When altering the content data structure it is possible to compare the various versions in terms of the attainment of a mark of 100% for the model essay answer. In effect a simplex process for optimising the content data structure is in operation. Altering the content data structure is a relatively straightforward procedure when done manually. Should the process of the construction of the content data structure from the content schema become automated then it should prove possible to rapidly generate new versions of the content data structure as required. A previous version or versions of the content data structure may be kept for reasons of reference, accountability and auditing.

Furthermore in practice the author would request that the examiner supply a model essay answer in order that the examiner be satisfied that the intended automated marking of their essay set is acceptable to the examiner's own standards of marking. In order to be of any practical use the content marking algorithm should not be modified for any one particular model essay answer, as this would constitute an unstable baseline to mark all the essay sets. Changing the content marking algorithm for each essay set would provide a constant, possibly uncontrolled, source of subjectivity that would be difficult to measure or eliminate. Nor would altering the software to suit a particular essay set or to suit a particular model essay answer be acceptable, for exactly the same reason.

Natural evolution of the algorithm and or the software is expected. When operational, therefore, there is a need to log the version of the algorithm – software combination which was used to mark each essay set to create an audit trail if required.

Further the encouragement, or permitting, of a set-by-set alteration of algorithm and / or software results in a less-than-maximum achievement of the potential that automated essay content marking offers. Altering the marking algorithm or software will inevitably lead to poor reproducibility, poor consistency and poor accountability. This in turn would lead to a loss of credibility for the automated marking of essays for content.

For all the developmental and test essay sets used in this study the model essay answer was expected to obtain a mark of 100%. In the study each model essay answer did indeed produce a mark of 100%. Thus it can be inferred by the reader that the model answer, content data structure, and the content marking algorithm can work together to produce a valid mark albeit in the case of a model essay answer.

4.2.2 'Linear' content data structure

One of the author's early concerns over the content data structure was that as the model essay grew in size linearly, then the content data structure could grow 'explosively'. In practice the number of entities of the content data structure produced for the various essay sets used by this author is about 0.7 to 0.8 of the word count of the model essay answer. The Table 4.2.2 below is in ascending order of the number of entities and shows this information for all the author's essay sets.

Essay Set	Number of Words	Number of Entities	Ratio [Entities / Words]
CatMat	6	4	0.7
RG	25	16	0.6
BigBox	28	27	1.0
iNet	47	38	0.8
162Q3	80	58	0.7
162Q2	58	62	1.1 *
162Q5	77	63	0.8
162Q6	309	83	0.3 *
OUQ7	119	92	0.8
162Q1	118	97	0.8
162Q4	128	126	1.0
RAB1-5	264	154	0.6
3202Q6	207	189	0.9
3202Q5	350	239	0.7
OUQ12	398	331	0.8
SE1-2	525	289	0.6

Table 4.2.2: Relationship between size of model essay and content data structure

Nearly all the ratio of number of entities in the content data structure divided by number of words in the model essay answer appears to range from around 0.6 to 1.0. Only two model essay answers (which are asterisked in the table above) have ratios that are clearly outside this range. Essay set 162Q6 (ratio 0.3) has 4 marks (out of 10 possible) based on a series of calculations rather than being a true essay, while essay set 162Q2 (ratio of 1.1) is more factual in nature in comparison with other model essay answers.

It must be noted that this is not strictly a 'linear' relationship in the sense of the pure mathematical linear relationship. However, by being given or creating the model essay answer then an approximation of the size of the content data structure may be made. As a general rule, by doubling the model essay answer (in words) it is expected that a doubling of the associated content data structure (in entities) also occurs. A ten-fold increase in the size of the model essay answer then would result in about a ten-fold increase in the size of the content data structure.

4.3.1 Developmental essay sets

These essay sets were used to design, develop and code the algorithms for style and content. As it has been stated many times elsewhere in this thesis, the style part of this research is in suspension until style marked essay sets become available to this author. In a research project of this nature there has of necessity to be iteration between the design, the development and the coding phases using the development essays as the common factor linking these phases together. It must be noted that all the development essay sets are not known to be part of any real assessment, and certainly there is an absence of assessment instrument as far as this author is aware. Where marks have been assigned to any of the developmental essay set(s) then these marks were assigned by the author.

The shorter pieces of text were first used to try out newer designs or alternative designs. Using these shorter pieces allowed more bounded experimentation to occur. Once the content algorithm, content data structure and software were shown to operate satisfactorily on the shorter pieces of text then progressively longer pieces were tested. This scaling up facilitated larger-scale experimentation. This approach was likewise was applied in relation to the size of the essay set where smaller essay sets were tested before larger essay sets.

4.3.1.1 test

This essay set was used as a general essay set and as 'work-in-progress'. As the need arose essay(s) and or schema(s) were tried out using this essay set. Likewise new algorithm(s) and new software were tried out using this essay set. A similar situation exists for software development as there was the use of a 'work-in-progress' or WIP directory and a directory called 'SearSoft' for released software. This is the essay set used for timing throughput for the various parts for SEAR software.

4.3.1.2 CatMat: ~ 4 entities in the content schema

This was the first essay set used for content marking. The content data structure consisted of four entities, one relationship (a SimplexAnd) and three fact entities one each for cat, sat and mat.

4.3.1.3 RG: ~ 16 entities in the content schema

This is a much shorter version of the developmental essays devised by Dr. Butler. This essay set is limited to the years of Robert Gordon's birth, retirement and death together with place of birth and how much Robert Gordon inherited. A larger schema based on the same subject is shown on the author's personal web site – http://www.jkp.christie.btinternet.co.uk and repeated in Appendix D, to illustrate the expected format for the content schema to be produced by an examiner. The content data structure produced from this larger schema is in the order of 60 entities in size.

4.3.1.4 BigBox: 5 essays ~ 27 entities in the content schema

This is based on the box of objects mentioned near the beginning of section 4.1 and is a non-sensible piece of text. Its existence is to provide a basis for flexibility of content to aid in the development of the content data structure. The author is free to alter the contents of the BigBox essay set at will without distorting real situations such as the personal history of Robert Gordon.

4.3.1.5 iNet: 7 essays ~ 39 entities in the content schema

This essay set is based on the Robert Gordon University's Intranet welcome message. This was the first essay set used with essays that were deliberately created to represent a range of extent of essay content. Three essays were created based on each sentence of the model essay answer. The purpose of these essays was to check that the Content Report and the Schema Report were correct. A further three essays were based on different pairs of sentences taken from the model essay answer. There was an additional essay, which was based on the original text of the welcome message. Further the correct allocation of marks, the correct reporting of 'Usage' and 'Coverage' was also first tested using this essay set.

4.3.1.6 RAB 1-5: 5 essays ~ 154 entities in the content schema

This author was particularly grateful and fortunate to receive these five essays. These five essays were produced by Dr Butler and are based on a short history of Robert Gordon, the founder of Robert Gordon's College and its off-shoot The Robert Gordon University that is available from that university's own web site.
(<http://www.rgu.ac.uk/about/theuni/page.cfm?pge=733> Accessed 4th September 2002).

Dr Butler intended these five essays to be very similar in content, yet significantly different. Three of these essays purport to represent "honest student" work, one represents a poor essay (RAB 5) and two represent a good essay content (RAB 1 and RAB 2). The remaining two essays purport to be representative of "less honest student" work. One essay (RAB 3) is a double sized version of (RAB 2) by having its original text repeated, while the other is a "bad faith" essay (RAB 4). Appendix F contains a copy of a on-topic essay (RAB 1) and the bad faith essay (RAB 4) to facilitate the reader in determining what a "bad faith" essay is.

These two "less honest" essays were created by Dr Butler in order to test the robustness of any software this author created. Any automated essay marking software must deal with any deliberate attempt at duping. Hence RAB 3 is a crude attempt at gaining more marks by using duplication of the essay text. RAB 4 is a sophisticated attempt to use key facts but without the correct relationship(s) between these key facts and without the correct context surrounding these facts. In other words Dr Butler deliberately created a "bad faith" essay.

This author has never received a "bad faith" essay, and indeed never expects to receive such an essay. Equally this author is not aware of any academic who has, or has admitted to having, received such an essay. This author is at a loss to explain why an essayist would deliberately construct a "bad faith" essay, especially for a high stakes assessment.

4.3.1.7 SE 1-2: 2 essays ~ 289 entities in content schema

This is a brace of staff development essays on economic and social history, namely the introduction of the first commercial railway traffic. The difference between the essays is that one is written using the style and vocabulary of a upper A/B social class (SE 1), while the other essay is written in the style and vocabulary of a middle to lower social class (SE 2). These essays were used in marker development exercises to show them that marking the style of essays may be susceptible to subjective effects arising from the essayist's vocabulary and other identifiers such as essayist's name.

The author is given to understand that the exercise based on the use of these two essays has been conducted at a number of Scottish and English universities.

4.3.2 Marked test essay sets

In this section the specific questions that were used as the assessment instrument are repeated here for the reader's benefit. The marks allocated for each question, part and sub-part thereof are also given in square brackets.

Table 4.3.2 below gives the name of both the second markers. Both are experienced lecturers. The author approached these two individuals on the recommendations that were jointly made by the author himself, and his PhD supervisory team. The author offered both the individuals the choice of the essay sets to second mark. Both elected to mark those essay sets that they were comfortable in marking.

Second Marker	Organisation
Ms Caroline Norton	Aberdeen College
Dr Mohamed Sharif	Napier University

Table 4.3.2: Table of second markers

The first six essay sets are taken from December 2000 Data Communications for the Level 1 Postgraduate Diploma / MSc Degree in Information Systems. There were six questions in this paper. The author did not use complete questions but only used selected parts of each question. The criterion for selection was to use only those parts of questions where a textual answer was required.

4.3.2.1 162-Q1: Part C i, ii and iii 68 essays ~ 97 entities in the content schema

"For a file of data to be transferred between two computers:

(i) what must exist between the two computers? [3]

(ii) what tasks must be performed? [6]

(iii) what is used to ensure co-operation between the two computers? [1]"

4.3.2.2 162-Q2: Part C 45 essays ~ 62 entities in the content schema

"The ISO OSI 7-layer model is used to describe communications protocols. In general terms, without discussing the specific role of each layer, what are the features of a layer? [10]"

4.3.2.3 162-Q3: Part A, B i and ii 62 essays ~ 58 entities in the content schema

"(a) A telephone quality channel nominally has a bandwidth of approximately 3kHz (300Hz to 3.4kHz). Hi-Fi music requires a bandwidth of at least 15kHz.

Explain briefly how, subjectively, a listener receiving music over a telephone channel would be aware the reproduction was not 'hi-fi'. [3]"

"(b) The human hearing frequency response covers frequencies up to approximately 20kHz. State, with reasoning, whether any advantage or disadvantage would be obtained by:

(i) extending hi-fi bandwidth to include frequencies up to 25kHz.

(ii) reducing hi-fi bandwidth to 10kHz. [2]"

4.3.2.4 162-Q4: Part A, B and C 50 essays ~ 126 entities in the content schema

"(a) Data is sent along a channel as NRZ pulses with a data rate of 75bits per second. Give an indication, with reasoning, of the required channel bandwidth to ensure 100% readability of the data without using excess bandwidth. [6]"

"(b) The above system is modified to produce a new system where the data is carried by On-Off-Keying (OOK). Quantify the change this effect would have on the bandwidth. [4]"

"(c) Outline how data communications can be accomplished using asynchronous serial transmission. [10]"

4.3.2.5 162-Q5: Part A and B 37 essays ~ 63 entities in the content schema

"(a) Discuss briefly what is meant by the term "Ionosphere" when applied to the Earth's atmosphere. Explain briefly why it changes from day to night and from summer to winter. [6]"

"(b) Interaction between the Ionosphere and a radio wave is dependent upon the frequency of the radio wave in question. Describe what is meant by the terms 'critical frequency' and 'maximum useable frequency'. [4]"

4.3.2.6 162-Q6: Part B and C 23 essays ~ 83 entities in the content schema

"(b) Explain briefly why the time delay calculated above might cause problems / confusion if two people talk to each other over a telephone channel. Assume that both people are on the ground, but that their telephone conversation is routed via the above satellite. [4]

Explain what is meant by a radio-wave being 'vertically polarised'. [2]"

"(c) By describing its operation, explain why the shift register and digital logic implementation of a Cyclic Redundancy Check (CRC) will result in all 0's in the receiver shift register if there are no errors in transmission. [10]"

The following two essay sets were provided by the Open University. These essay sets were examination scripts and were provided as word documents. The actual questions were not provided at the same time as this author received the essays.

4.3.2.7 OU-Q7: Part A and B 18 essays ~ 92 entities in the content schema

"(a) Assuming that all records held in a file are the same (known) size, briefly compare the direct retrieval of data from disk storage when space is allocated using

(i) contiguous allocation,

(ii) block oriented file mapping. [3]

(b) What is the major problem caused by the use of contiguous allocation of space on disk. [1]"

4.3.2.8 OU-Q12: Part A, B, C and D 20 essays ~ 331 entities in content schema

"(a) Explain the meaning of the term pipelining and show how a four stage pipeline could increase running speed by up to four times. State the criteria that must be satisfied for the successful operation of the pipeline. [6]

(b) From a simple analysis it might appear that the greater the number of stages in a pipeline, the greater the speed increase. Explain why this observation is not correct in practice. [2]

(c) Compare the philosophies behind CISC and RISC and explain why RISC architectures are more suitable for pipelined systems. [8]

(d) In fact pipelining is possible with CISC architectures. Suggest how this could be achieved and assess the speed increase with a four stage pipeline. [2]"

These last two essay sets are taken from December 2000 Telecommunications Transmission Systems for the Stage 3 / 4 in these three programmes MEng in Electronic & Electrical Engineering, BEng (Honours) in Electronic & Communications Engineering, and the BSc (Eng) (Honours) in Electronic & Electrical Engineering.

4.3.2.9 3202-Q5: Part A and B 27 essays ~ 239 entities in the content schema

"(a) Describe the processes involved in the transmission of a speech signal using Pulse Code Modulation. Illustrate your answer with reference to a PCM system using 3 bit samples. [14]"

"(b) Why do practical PCM systems require timing and synchronisation information to be sent along with the samples from the PCM coder to the PCM decoder. [6]"

4.3.2.10 3202-Q6: Part Ai - Av and B 32 essays ~ 189 entities in the content schema

"(a) 'Digital communications systems are really analogue systems'. Describe the significance of this statement making reference to:

(i) The bandwidth of a pulse train. [2]
(ii) The bandwidth of a communications link. [2]
(iii) The attenuation of a communications link. [2]
(iv) The use of decision thresholds in a digital communications receiver. [2]
(v) Noise in a digital communications receiver. [2]"

"(b) Why do errors occur in the transmission of information over digital communications links? [3]"

4.3.3 Range of marks awarded

There is a total of 126 marks that were awarded in total from all the marked essay sets. In total there are 28 parts to these essay sets. The distribution of both the question marks and part marks are given in the two following tables 4.3.3.b and 4.3.3.c.

Question Value	Frequency
4	1
5	1
10	3
13	1
16	1
18	1
20	2

Table 4.3.3.a: Distribution of Question Marks

Part Value	Frequency
1	2
2	9
3	4
4	3
6	5
8	1
10	3
14	1

Table 4.3.3.b: Distribution of Part Marks

Across all the essay sets there is a range in the marks assigned to the question and a range of marks assigned to the parts within the questions. These ranges should permit the investigation of the effect of actual manual marks awarded with the computer assigned marks. There is thus the opportunity to examine the performance of automated essay content marking in a variety of marking situations.

The range of marks awarded do not appear to be too different from that which would be expected to be awarded under normal assessment circumstances. The range goes from the awarding of a single mark or two marks that invites the essayist to write small parts of text, to major awarding of marks of 10 and 14 which requires the essayist to construct lengthy text responses. The range of marks available in this research project should minimise any doubts a reader may have in that a special set of questions were taken, nor that a certain range of marks awarded was specifically selected for this research project.

4.4 Statistical analysis of data

When the first marking was conducted there was no intention of using the essays for this research project. In other words the first markers were not aware that their work was going to be, albeit retrospectively, used in this research project. This must not be taken as a comment on the quality of the first markers nor as a comment on the quality of their marking. Both the first markers and the two known second markers appear to be insensitive to spelling errors and grammar errors. That is both these types of markers would overlook these errors when marking content and giving the essayists the benefit of the doubt. The first and second markers appear to be also insensitive to the use of any shorthand terms used by the essayist, for example the use of the shorthand term "O/H" in place of the term "Overhead", would be equally accepted by human markers.

However the current state of the automated marking software is sensitive to errors of spelling and grammar. Equally the automated marking software is sensitive to the presence of any shorthand expressions used by essayists. By being sensitive to these spelling errors, grammar errors and shorthand terms means that the software will therefore not award marks, thereby resulting in lower marks awarded than those awarded by the human markers.

Furthermore it must be noted that there are major differences between the first and second marking that will affect the statistical analysis of the data.

The first marking was done on the original hand-written submissions, with all the corrections, scorings out and the existence of non-textual elements. It is extremely important to note that these submissions contained a variety of these non-textual elements, such as graphs, tables, equations and diagrams. The first markers were, of course, absolutely correct in using non-textual elements in the marking of these essays. Therefore there will be essays whose first mark is very different from the computer mark [and any second marking]. By the use of specific codes the author has indicated in the computer mediated essays where non-textual elements had originally been present.

The second marking was done using paper copies of the essays, as they would be processed by the computer – typed essays with no non-textual elements present, although these versions were annotated to represent where diagrams and equations were present in the original submissions. All the spelling errors and grammatical errors were retained when the computer versions of the essays were prepared. Had the second marking been conducted on the original submissions then it would be nigh impossible to ensure that the first marker's essay annotations would not influence the second marker or second marking.

This results in several points to note when considering the experimental results.

1] First and second marking are not directly comparable. In fact the second marking and computer marking are more equitable in format.

2] Any third, and subsequent, marking would be conducted on the same basis as the second marking. That is it would be based on the computer version of the essays.

3] No essay set was marked by both the second markers. This author had hoped that one essay set could have been triple marked, that is marked by the original marker and both the second markers, as this would have provided an opportunity to study inter-marker correlation between a larger pool of markers.

Should this author have the opportunity to conduct a multi-marker study again then the following would be core aspects of the experiment.

1] Sufficient copies of each essay submission would be taken to meet the number of markers to be involved in the experiment. Further copies would also be taken should the experiment include a mark-remark component. Using a sufficient number of copies should ensure that each marker, for first marking or for remarking, would be using a clean essay copy as a baseline.

2] Under half the markers would only mark the copy of the original version, while under the other half would only mark the version prepared for the computer marking. The remaining markers would mark both versions of the essay. This tri-fold approach in the use of the markers would then facilitate comparison of the marking the original versus the computer versions, and that comparison would cover inter-marker and intra-marker comparisons.

3] Human markers would be encouraged to fully annotate the submissions that they mark. Perhaps the written annotations should be augmented by recording the visual and the verbal in order to obtain a full profile of human marking. Care must be taken in any annotation(s) being made and being recorded since the Data Protection Act 2002 imposes limitations on what annotation markers may make. A secondary effect that arises from this act is that many educational organisations are advising their staff to avoid making annotation(s) on any essay or on any other alternative media.

The recording of the annotation(s) would serve a <u>two-fold purpose</u>. <u>Firstly</u> to gain insight into how the human markers assign their marks. <u>Secondly</u> to generate a databank of feedback comments that could in time be incorporated into the author's software.

Chapter 5: Discussion of findings [for content only]

5.1 Interpretation of findings

This chapter is organised into a number of parts, each of which specifically examines one aspect of how automated marking performs against human marking.

The reader's attention is drawn to the following appendices used in this chapter:

- Appendix G: Content Reports
- Appendix H: Inter-marker Summary
- Appendix I: SPSS output
- Appendix J: Scatter plots

Appendix G contains all the Content Reports as produced by the SEAR software for each of the test essay sets. It is from these outputs that the reader knows exactly what information the software provides to the examiner. Each report contains information on the date and time of the content marking, the number of essays marked during that processing run together with the identity of the content data structure used in the marking process. For each essay the number of words found, the number of sentences found together with percentages of Usage and Coverage, the mark awarded per part and the overall mark expressed both as an integer and as a percentage are generated. Any later essay marking, such as caused by a late submission from the essayist, is appended to the end of the initial report.

Appendix H contains the raw data obtained from the first human marker, the second human marker where present and the SEAR software. The purpose of this particular appendix is to draw the reader's attention to the raw marks awarded to the essays and shows the spread of marks awarded by each marking method to each essay. It also acts as a source for the statistical processing used in this chapter.

Appendix I contains the outputs generated from the SPSS statistical package. The reader's attention is drawn to the statistics contained in this appendix as these statistics are used in different sections of this chapter. The reader will note that a system of star ratings is displayed beside some of the data. A "**" indicates that the correlation is significant at the 0.01 level of significance, whereas "*" indicates that the correlation is significant at the 0.05 level of significance. For each essay set there is given a simple set of descriptive statistics together with Pearson's correlation coefficient and Spearman's RHO correlation coefficient.

Appendix J provides the reader with a visual representation of the data present each of the ten test essay sets. Each scatter plot has the human generated marks on the X-axis and the computer generated marks on the Y-axis. Where there is second human marks available then they are also plotted on the Y-axis.

5.1.1 Analysis by Version of Content Data Structure

There are four versions of each content data structure for each of the essay sets used. These versions were created to show that the content marking algorithm did in fact respond to or pick up on the case of the word and the Roget's classification number. Performance, as in the ability to mark correctly, was expected to be different for each of the four versions. The "Original" version defined as that version obtained from the examiner's answer schema, was intended to become the baseline performance version, and the other three versions were expected to display a varying decrease in performance. The differences between the four versions are outlined in descriptions of these versions as shown in Table 5.1.1.a below.

Version	Description
Original	As produced from the Answer Schema
NoCap	Original, but with all words in lower case
NoNum	Original, but with all Roget's classification numbers set to 0000 and all Parts-of-Speech set to *
NoNumCap	Combined effect of the NoCap and NoNum

Table 5.1.1.a: Content Data Structure Versions

The results of testing proved the original version to be the worst performing and equally worst to the "NoCap" version. The best performing version was the "NoNumCap", and it was often equally best with the "NoNum" version. Table 5.1.1.b below shows the effect on the mark awarded on the model answer from the four different content data structure versions.

NoCap Version

The author proposed that making the words in the essay and the words in the content data structure all the same case would circumvent all the matching, and hence marking, problems due to differences of text case between the essay words and the words in the content data structure. Therefore it was surmised that there should be no performance improvement where every essay contained words which exactly matched in case; but it was surmised that there should be a detectable improvement in those essays where previously un-matched content data structure entities are now being matched.

Hence essays in this latter category would obtain enhanced marks. Testing proved this supposition to be true.

The author elected to make all text lower case, rather than all uppercase, as it is easy on the eye when reading. The results would be expected to remain the same had all the text been converted to uppercase.

NoNum Version

The author proposed that there would be a serious deterioration in marking performance when all the Roget classification numbers were set to 0000 and all the Parts-of-Speech were set to *. The reason is that all Roget equivalents would not be used. Therefore only exact match between essay words and words in the content data structure would be the basis for awarding essay marks. Of the four possible versions of the content data structure this version was expected to produce the poorest marking performance.

In the event the marking performance for this version of the content data structure was counter-intuitive as the performance was better than both the Original version and the NoCap version. The author will undertake further experimentation, post PhD, to resolve either why this version is counter-intuitive or to identify what improvements are required to the SEAR system in order to realise the initial proposal.

NoNumCap Version

The author proposed that this version would perform better than the NoNum version, as the requirement for the exact case matching of the words would be relaxed. The author also proposed that this version would exhibit a performance that was worse than the original for exactly the same reasons given for the NoNum version discussed above. In the event this version proved to be counter-intuitive since this version was the best performing version of all the four used by the author. The performance increase resulting from the relaxation of the exact matching of case was seen in the NoCap version versus Original versions, so there was no surprise here. The surprise in using this version is fully attributable to the NoNum counter-intuitive effect as seen above.

Solving, or resolving, the counter-intuitive performance for the NoNum version should solve, or resolve, the same problems that exist with the NoNumcap version.

In this research project the NoNumCap version was adopted as being the version of the content data structure to use throughout. When the resolution has been determined for the counter-intuitive problem then all the essays sets, both the development essays and test essays, will be reprocessed.

Essay Set	Original	NoCap	NoNum	NoNumCap
EE162 Q1	40	40	100	100
EE162 Q2	100	100	100	100
EE162 Q3	100	100	100	100
EE162 Q4	63	69	100	100
EE162 Q5	20	20	100	100
EE162 Q6	33	33	100	100
OU Q7	17	33	100	100
OU Q12	81	86	100	100
EE3203 Q5	30	100	30	100
EE3202 Q6	69	69	100	100

Table 5.1.1.b: Model essay mark [%]
– effect of version of Content Data Structure

Since the NoNumCap version was the consistently best performing of all the four versions it was decided to use the results from using this version in the analysis throughout this chapter.

5.1.2 Analysis of each test essay set

The table and graph below shows both the types of correlation for the first marker and the marks generated by the software system.

Essay Set	Pearson	Spearman
EE162 Q1	0.594**	0.596**
EE162 Q2	0.404**	0.376**
EE162 Q3	0.008	-0.068
EE162 Q4	0.238*	0.336**
EE162 Q5	0.302*	0.394**
EE162 Q6	0.187	0.207
OU Q7	0.263	0.374
OU Q12	-0.348	-0.097
EE3202 Q5	0.128	0.005
EE3202 Q6	0.157	-0.004

Table 5.1.2:

Correlation between Human and Software Marking

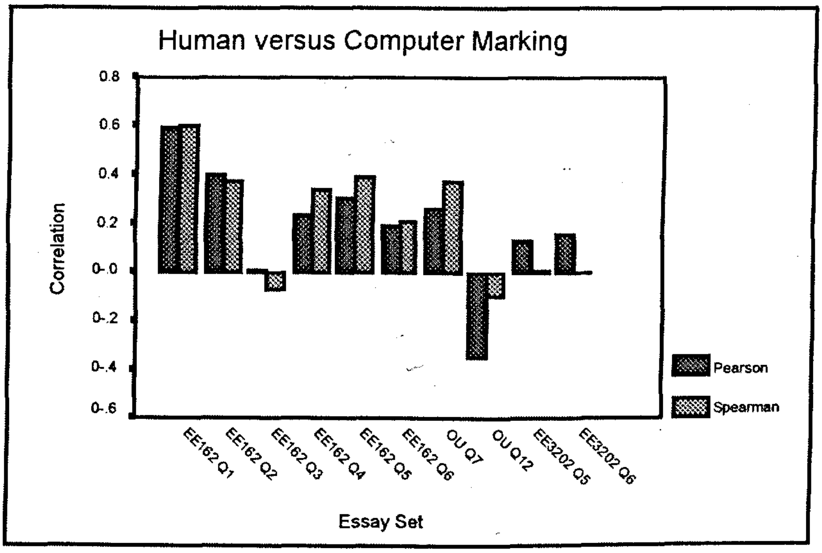


Figure 5.1.2:

Graph of Correlation Human versus Computer Marking

Initial results did yield even better correlation coefficients of about 0.7 but only in one or two essay sets. At the same time the remaining essay sets yielded somewhat poorer results than those reported in the above table. Further development of the software system during the course of this research has resulted in a less extreme situation.
The author does not regard the initial extreme results to be useful. It is better to have a software system that does not perform in the extreme.

At best the current state of the software system appears to produce results no better than the results expected from a pair of mediocre human markers. At worst the software system appears to have a performance comparable to a pair of (very) poor human markers. However the reader should note that nearly all the essay sets had some type of non-textual element present in the answer schema. These elements are easy for the essayist to create and easy for the marker to award marks, but it is not possible for the software to mark these. When viewed in this light then, perhaps the software system did not perform too badly after all.

The software system tended to produce more marks either at either extremes (zero or maximum) of the possible marks for each essay set. This may be obtained for two reasons:

Firstly the threshold for awarding marks may been set too high.
Secondly the algorithm may lack sensitivity.

If the threshold had been set too high then this would lead to a higher than expected number of zero marks being awarded. A high threshold effectively sets an artificially high standard, almost in the style of performance criteria – either one passes or one fails, there being no middle value.

A lack of sensitivity in the algorithm leads to a bistate or a Gaian situation in which the software system will surpress marks to zero until a particular value has been reached at which point it then awards full marks.

5.1.3 Analysis by mark awarded

The table below shows the effect, if any, on the volume of mark awarded by the examiner on the automated marking.

Essay Set	Mark Awarded	Pearson	Spearman
EE162 Q1	10	0.594**	0.596**
EE162 Q2	10	0.404**	0.376**
EE162 Q3	5	0.008	-0.068
EE162 Q4	16	0.238*	0.336**
EE162 Q5	10	0.302*	0.394**
EE162 Q6	6	0.187	0.207
OU Q7	6	0.263	0.374
OU Q12	21	-0.348	-0.097
EE3202 Q5	20	0.128	0.005
EE3202 Q6	13	0.157	-0.004

Table 5.1.3: Analysis by Mark Awarded

It would appear that essays attracting a mark of 10 give a good [*] to excellent [**] inter-marker correlation when marked by the SEAR software. Where there is a large allocation of marks associated with a schema then the performance of automated marking is reduced. The worst automated performance is with a 21-mark assessment. This fall in performance may be due to the answer schema or the software or the marking algorithm not being able to cope with the larger size of essay required to cover that volume of marks. Perhaps the software has to be enabled to cover the more complex relationships that may arise in such a larger answer schema. A possible cure is to avoid having large mark allocations being present.

Where an examiner assigns 6 or so marks to an answer then automated assessment performs better than for the 20 or 21 mark answer, but not quite good enough to be classed as statistically significant at the 0.05 level of significance. Perhaps the fall in performance is due to the answer schema being very specific requiring the essayist to be "spot on" with their answer. Alternatively the answer schema may sufficiently echo the assessment instrument's stem thereby giving the essayist an opportunity to gain marks merely by repeating the question stem. More selectivity in the construction of the answer schema may cure this problem.

5.1.4 Analysis by type of student

Three types of essayist were involved in this research project. These were third and fourth year full-time undergraduate students, full-time post-graduate conversion students and open / distance learning undergraduate students. The table below indicates the changes in automated marking performance with type of student.

Essay Set	Student Type	Pearson	Spearman
EE162 Q1	Undergraduate	0.594**	0.596**
EE162 Q2	Undergraduate	0.404**	0.376**
EE162 Q3	Undergraduate	0.008	-0.068
EE162 Q4	Undergraduate	0.238*	0.336**
EE162 Q5	Undergraduate	0.302*	0.394**
EE162 Q6	Undergraduate	0.187	0.207
OU Q7	Open / Distance	0.263	0.374
OU Q12	Open / Distance	-0.348	-0.097
EE3202 Q5	Postgraduate	0.128	0.005
EE3202 Q6	Postgraduate	0.157	-0.004

Table 5.1.4: Analysis by Student Type

The automated assessment of the undergraduate essay sets appear to offer the best performance, with the postgraduate next and the open / distance essay sets bracketing the postgraduate essay sets. It may be that the undergraduate assessment instruments require the essayists to be specific thereby the answer schema is specific and this results in a highly specific content data structure. The other two types of essay sets are from essayists who may be expected to generate more articulate essays or to operate at a lower level of specificity. The vagueness in answer schema then cascades through to the automated marking algorithm.

However the author wishes to raise a concern here. Although there are ten essays sets that were made available to the author, it is highly dangerous for any reader to make much of the above table. To correctly determine if there is an effect of arising from different types of students then considerably more essay sets from each identifiable student type would have to be analysed and to have the appropriate statistical tests applied.

5.1.5 Analysis by size of answer schema

The following table indicates how the size of the content data structure, measured in terms of the number of entities making up that content data structure, affects the performance of the automated marking.

Essay Set	Number of Entities in Content Data Structure	Pearson	Spearman
EE162 Q1	97	0.594**	0.596**
EE162 Q2	62	0.404**	0.376**
EE162 Q3	58	0.008	-0.068
EE162 Q4	126	0.238*	0.336**
EE162 Q5	63	0.302*	0.394**
EE162 Q6	83	0.187	0.207
OU Q7	92	0.263	0.374
OU Q12	331	-0.348	-0.097
EE3202 Q5	239	0.128	0.005
EE3202 Q6	189	0.157	-0.004

Table 5.1.5: Effect of Size of Answer Schema

The table clearly shows a peaking in performance around the 100-entity mark in a range of 60 to 120 entities, with a fall off on either side of this range. For a smaller number of entities then there may be insufficient entities to cover the expected range of essays, leading to the suggestion that the answer schema or algorithm may require a certain threshold of complexity before it is useful for automated marking. On the other hand a large number of entities may lead to problems of an essay only being sparsely matched with the content data structure thereby giving the illusion that the essay poorly matched when in fact the essay may be acceptable to a human marker.

The effect of content data structure on the performance suggests that these may be an optimal range of answer schema that examiner seeking to deploy automated marking should try to achieve. Too small and the answer schema becomes too restrictive. Too large and the answer schema becomes too difficult to effectively match.

5.2 Analysis of human marking performance

It could be considered un-scientific to have only human versus computer marking. There is a need to establish which, if any, of the marking methodologies is awarding the correct marks.

The range of marks awarded, the average and standard deviation by the different human markers provide an insight into the process of marking an essay for content.

5.2.1 First and second human marking

In this study there is not a true comparison between the first and the second human markers as outlined in chapter 4 section 4. For proper, or true, inter-marker comparison both sets of markers would have to mark exactly the same material, as would be the case under normal academic circumstances. When viewing first and second marker performance the reader must remember that the second marking was not done on exactly the same materials as the first marker. The first marker marked the hand-written submission with all its rough work, changes, tables, diagrams etc, whereas the second markers were only privy to the printouts of the pure essay text.

Essay Set	First v Second Marker		Second Marker
	Pearson	Spearman	
EE162 Q1	0.704**	0.700**	X
EE162 Q2	0.810**	0.740**	X
EE162 Q3			
EE162 Q4			
EE162 Q5	0.164	0.277	X
EE162 Q6			
OU Q7			
OU Q12			
EE3202 Q5	0.644**	0.708**	Y
EE3202 Q6	0.749**	0.749**	Y

Table 5.2.1: Comparison of First and Second Markers

From the above table it is noted that the two essay sets marked by Y produce correlation indicating that as a second marker, Y, is producing marks that are comparable with the first marker. In general comparison with the various second marking data shown in Chapter 2 this occurrence of second marking is in the upper end of the range of inter-marker correlation.

Marker X displays a mixture of very good inter-marker correlation with the first marker. This is especially true in the essay set EE 162 Q2 with a correlation of 0.810, yet with essay set EE 162 Q5 produces an inter-marker correlation of 0.164, which is towards the lower range of acceptability as seen in the various inter-marker correlation tabled in Chapter 2. Of course this does not, and should not, imply that for this particular essay set one marker is more correct than the other. The realisation that the marks awarded are different is the extent to any possible claims on marker quality may be made.

In instances where both second markers exhibit a high correlation with the first marker there is no guarantee that both the markers are correct. Given that there is such a format difference in the actual submissions marked, this author is surprised by just how good the actual inter-marker comparisons in this research project were produced as shown in the graph below.

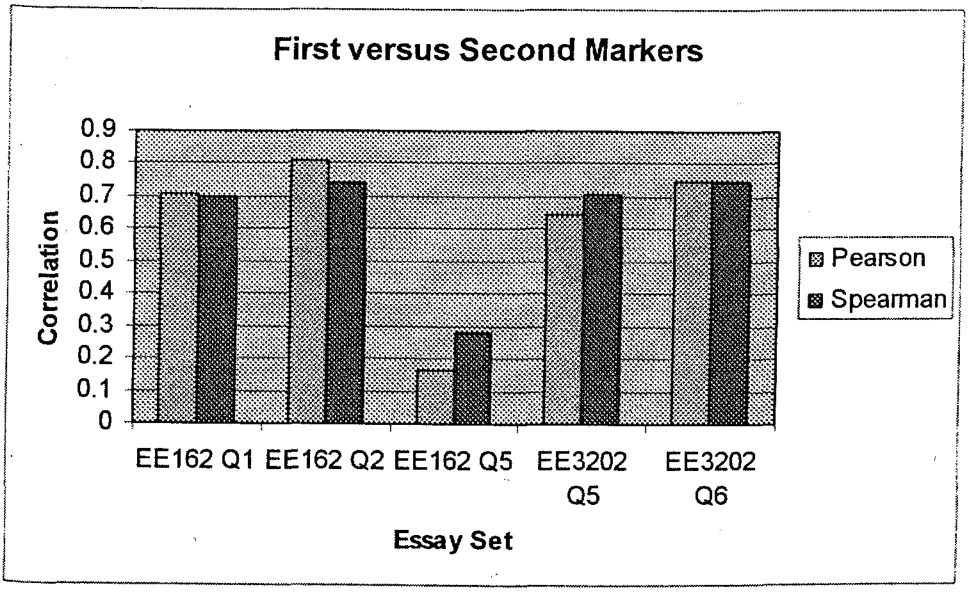


Figure 5.2.1: Graph of First versus Second Markers

5.3 Examination of poor automated marking performance

Regardless of the current performance of automated marking there is always the tantalising thought that the performance could be improved. The two routes that this author envisages for improving the performance are firstly to improve the algorithm itself [Chapter 6 section 6.4 outlines the myriad of potential routes that this author has already identified] and secondly to overcome problems sourced from within the essays themselves. The next three sub-sections outline these problems and suggest possible approaches to minimise their effects or to overcome them completely.

5.3.1 Errors of spelling

The easiest possible solution to this problem is to identify those essay words that appear to have been misspelt by the essayist. All accidental misspellings sourced from the examiner's answer schema would be removed. All deliberate misspellings sourced from the examiner would, of course, have to be retained. However this is not as easy as would be suggested by first considerations. English spelling rules are not always rigidly enforced so applying spelling rules with known exceptions (to these rules) do not guarantee perfect operations every time.

Using a dictionary implies that there exists a complete and comprehensive dictionary, which would have to be maintained as new words are added. It is doubtful if a dictionary would in itself contain all the various technical terms across any range of subjects, let alone contain all the English words in existence. If such a dictionary were obtainable then it would require an enormous disk storage space and necessitate lengthy processing times.

Use of Soundex type spell checkers would require the generation and storage of the Soundex code with each word and would require the author to obtain a large enough wordbank, then to generate unique Soundex codes for every word in this wordbank. Most word-processor users accept that the software based spell checkers that are available today do not operate correctly. However the performance this type of software ameliorates with each successive generation released. Therefore the ideal of the perfect elimination of all accidental spelling is unlikely to be reached.

5.3.2 Errors of grammar

The easiest possible solution to this problem is to use some sort of automated grammar checking to identify and correct grammar errors. Grammar rules are less well defined than spelling rules. The grammar checking software in use today is somewhat suspect in what is identified as being a grammar error. Echoing spell checking software, the ideal of the perfect grammar checking software is also unlikely to be reached.

5.3.3 Use of abbreviations, shorthand and other similar devices

The easiest possible solution to this problem is in the hands of the examiner. Should there be alternatives in the form of recognisable abbreviations and other shorthand devices then it is an obligation laid on the examiner to list the alternatives so that they may be included in the answer schema. The current algorithm will allow alternatives to be accepted into the content data structure provided, of course, these alternatives are declared beforehand in the answer schema. It should not be the responsibility of this author to devise and maintain sets of appropriate alternatives, nor would it be feasible for this author to do so.

There is often a list of standard abbreviations found at the back of better quality (and often older) dictionaries. There may be made a case for the compilation of a super-list of all "common" abbreviations. Such a list would undoubtedly be large and there would have to be consideration made for the abbreviations used in plain format (for example UN) as well as "dotted format" (for example U.N.). The effect of the use of such a list on marking essays would be to reduce marking throughput. This reduction in throughput would not necessarily be accompanied by any improvement in the marking performance in terms of accuracy.

The effectiveness of employing such a list of abbreviations in the marking of essay content is an open question left for the reader to ponder.

5.3.4 Other potential sources of poor automated performance

There are other possible reasons for poor results in the automated content marking of essays. Most of these reasons are due to the fact that humans communicate in rational, yet seemingly non-logical, forms of expressions. Examples of this are the use of idioms, the use of metonymies and the use of slang. In the author's experience of marking essays there has been a low level (but unrecorded) of such usage. The author's marking experience has been solely in the field of information technology. Markers in other fields may have different experiences and different outlooks.

5.3.4.1 Idioms

An idiom is an expression the meaning of which is not conveyed by the actual words used. For example "Charles Darwin ploughed a lonely furrow when he developed his greatest ideas." where the phrase ploughed a lonely furrow is used in place of the phrase was alone.

The range of idiomatic expressions that could be used is probably extensive and is just as probably subject to local effects arising from culture, education and so on as well as local in the geographical sense. Human markers are largely able to detect and correctly respond to any idioms used by the essayist, as human markers have the intelligence to make the transition from the idiomatic expression to the underlying meaning therein represented. The problem with automated content marking is the initial detection of the idiomatic expression from which to establish the underlying meaning. The use of an encyclopaedia of idioms would be useful here in the initial detection of idiomatic expression, but the problem of the correct establishment of the underlying meaning may still remain. To check for idioms would generate a large negative impact on the throughput of marking essays for content.

The author has not yet established the extent of idiomatic expressions used in essays. It is possible that the use idiomatic expressions by essayists may represent a non-problem when marking essay content by computer.

5.3.4.2 Metonymies

A metonymy is a substitution of an attribute or an adjunct for the actual word or words that would otherwise be used. For example "There were a lot of faces when the city's Xmas lights were switched on." where the word faces is used in place of the word people.

The use of metonymies by the essayist may create a problem for a human marker if the marker fails to identify such an expression. However, for computer based marking the use of metonymies is a major problem, as the computer has to make the connection of the attribute to the item being referenced. In the example given, a human marker could make the connection between "faces" and "people", as the marker would know that each person has a face. For computer marking the connection will have to be made explicitly. Therefore the answer schema provided by the examiner would have to be augmented by a, possibly large, listing of all the potential attributes for each candidate word present in the answer schema. The overhead for creating such a list could be large and, again, there would be a reduction in content essay marking throughput.

The author has no information on the extent that metonymies are used by essayists. Until the use of metonymies is known then, again, this may represent a non-problem in content essay marking.

5.3.4.3 Slang

Here the author is not referring to, nor condoning, the use of abusive language. Here the term "slang" refers to expression of words, or phrases, used by restricted groups of people. The use of slang by the essayist may echo the problems of using idiomatic and metonymies in that detection and establishment of the underlying meaning is easy for human markers but not so easy in an automated environment. Until the extent of the use of slang by essayists is established then this may be yet another non-problem in content essay marking.

5.4 Marking throughput

The ten test essay sets used were used to measure throughput. This proved useful as there is a range of number of essays per set and there is a range of essay size within each essay set. Table 5.4 below shows the number of words, number of essays and the time in seconds. The laptop computer used for this measurement is a Digital HiNote VP Pentium with 24MB RAM and 730MB HDD using Windows 75 ™® as the operating system.

Essay Set	Total Number of Words	Total Number of Essays	Time [seconds]
EE162 Q1	10,406	69	600
EE162 Q2	4,885	46	262
EE162 Q3	8,210	63	375
EE162 Q4	12,167	52	589
EE162 Q5	6,574	38	324
EE162 Q6	4,584	24	222
OU Q7	2,288	19	115
OU Q12	9,353	21	487
EE3202 Q5	6,954	28	405
EE3202 Q6	6,796	33	362
TOTAL	**72,217**	**393**	**3,741**

Table 5.4: Marking throughput

It can be seen from this table that the SEAR software processes on average about 185 words per second. An alternative view is that it takes on average about 10 seconds per essay. Each measure shows that SEAR operates considerably faster than human marking in terms of pure throughput does. Further the use of the SEAR software potentially offers the examiner the capability to undertake other operations such as the possible identification of academic misconduct, and considering that when marking content there is a possibility of providing style marking as a bonus.

In Chapter 5 and Chapter 6 this author has suggested several technical improvements that could improve the performance of the marking of essays for content. Each of these suggested improvements and any other emerging suggestions for improvements that may be identified as this project further develops will always lead to a reduction in throughput. This arises from the realisation that each improvement necessitates an increase in complexity in design and in the resulting software. Therefore each improvement adopted will reduce throughput in marking. However by starting a high throughput there is considerable leeway for adding improvements before SEAR's performance becomes comparable with human marking.

Chapter 6: Conclusions and Recommendations

6.1 Review of significance of work

In its ultimate form automated essay marking would negate, remove or substantively reduce all the subjective problems as listed and outlined in Chapter 2 of this thesis. Perhaps it is appropriate here to summarise each source of subjectivity in marking and to state how automated marking will either remove / negate or reduce such subjectivity.

6.1.1 **Absence of marking schema** – subjective effect is <u>removed</u>

The absence of a marking schema will bring automated essay marking to a complete stop. With no marking schema there is no basis for awarding any marks to any essay. This is especially true for content marking. As the marking for style is dependant on a sample set of essays being marked by at least one human marker to calibrate the marking software then the absence of such calibration in the role of being a marking schema would make the automated style marking completely impossible.

6.1.2 **Position of essay in essay set** – subjective effect is <u>removed</u>

Each essay in turn is examined. At the commencement of the marking any essay all the appropriate settings should be reset to the same initial values that were applied to the first essay marked with the exception of the essay set batch data and the running values being left as persistent data. This would have the effect of each essay being marked as if it were the first and last of the batch. In other words the batch size is effectively set to one. Thus this eliminates the effect of a human marker becoming tired and the resulting effect this tiredness would have on the awarding of marks to any essay whilst the human marker is in this particular state.

6.1.3 **Influence of the previous essays marked** – subjective effect <u>removed</u>

This resetting to the known initial value for each new essay to be marked also removes the subjective effect of the quality of the previous essay or essays in the essay set. This then removes any memory effect(s) of the previous essay(s) marked and consequentially removes memory effects from influencing the marking of the current essay. The only persistent data would be logged in a result file.

6.1.4 Vocabulary, voice and errors of spelling and of grammar

– subjective effect is reduced

Automated essay marking may be useful in the separation of vocabulary and voice from other errors. Suitable programming would ensure that the software would be made insensitive to the effects of both vocabulary and voice. It is thus possible for these particular subjective effects to be removed.

It is unanimously accepted that automated spell checkers and grammar checkers are not completely accurate. However should the software make temporary or "on-the-fly" corrections to the essay as it is being marked then that would be acceptable if its purpose was solely to systematically remove these types of errors in order for an unbiased marking to take place. So the ability to make temporary correction(s) for these types of errors could negate adverse marking of content that would otherwise be deemed valid by a human marker. This then provides the opportunity to avoid systematically marking an essay to the detriment of the overall grade, especially for content marking, on pure language errors. This in itself is a good thing to do; however there is the ever-present danger that automated correction might actually be detrimental to the mark awarded to the essay, this is through the introduction of false errors. Thus the essay is marked down simply due to the software, with its over zealous automated correction.

The permanent correction of spelling errors and or grammar errors to any essay would not be acceptable. In the academic environment any tampering with a student's own work for whatever reason has been and will always be held as not acceptable. In any case permanent correction or alteration is probably illegal.

6.1.5 Introductions and Conclusions – subjective effect is removed

It is important to realise that in the context of automated marking, essays have to be marked 'as is'. Software should mark the essay as a whole not necessarily over- or under-weighting the structural elements of any essay. This is especially true for content marking.

6.1.6 **Marker's mood** – subjective effect is removed

Only in the realms of science fiction do computers display what could be categorised as intelligence. Currently software and hardware systems have at present no 'real' intelligence, even though many computer scientists claim the computer-based systems have some 'artificial' intelligence. It therefore follows that computer systems will not have mood. One needs intelligence before the possibility of mood can exist and only then is it possible for the subjective effects and discrimination effects to be developed. Thus there is no affect on the essay being marked.

6.1.7 **Submission media: handwriting** – subjective effect is negated

The mode of automated marking envisaged in this research project is computer oriented. Consequentially in this research all essays must be word-processed before this type of automated marking is possible. Of course hand-written essays or word-processed essays may be scanned and stored by computer equipment. Nevertheless the scanned versions of hand-written essays would have to be converted into a word-processed format; thereby the effects of both good and bad handwriting are negated.

6.1.8 **Submission media: document file format** – subjective effect is reduced

Limiting the choice of media to computer mediated files will be advantageous to the marking process. The research underpinning this research is predicated on various versions of Microsoft Word ™® the word-processing package that is very widely available in the beginning of the 21st century. However there are indeed several different word-processing packages available. Any word-processed essays that have been created by using word-processing package(s) other than Microsoft Word ™® would therefore all have to have their text extracted by an analogous software process to what this author has already developed for Microsoft Word ™®. The other approach would be first to convert the document into a Microsoft Word ™® formatted document for the subsequent extraction of text. There are, of course, many file conversion utilities for the converting of word-processed files from various alternative formats into the Microsoft Word ™® format. However a problem arises in that these conversion utilities are not completely perfect, so extraneous errors may be introduced in the process of file conversion that may impact on the final mark awarded to an essay. The impact is more likely to be negative rather than to be positive.

6.1.9 Subject of essay – subjective effect is reduced

For automated objective summative marking there appears to be little effect stemming from the subject being examined. There is indeed very little information on effect of subject when performing automated essay marking. Because the evidence is minimal or just non-existent this suggests that the effect on marks stemming from the subject is also minimal or non-existent. Further research, as stated in section 6.3.2.1 below, is required to finally make such a conclusion in settling this question.

6.1.10 Name Stereotyping – subjective effect is removed

It is important that software should not be constructed so as capable of recognising any specific essayist's name, or word-processed file name. It would be grossly unprofessional to develop software otherwise.

6.1.11 The race of the author and reverse discrimination – subjective effect is removed

Software used for automated marking must be written in such a way so that it avoids embedding any bias or discrimination. It would be against the standard principles of the computer profession to deliberately seek to include routines for these particular subjective effects.

6.1.12 Gender – subjective effect is removed

It is possible that there is indication of gender of the essayist in the computer document file name. Should there be any such indication, then the marking software is not, and should not, be programmed to act on this gender information. It would be unprofessional for the programmers to do otherwise.

6.1.13 Marker's expectations – subjective effect is removed

Each essay will be marked as is it is presented. There should be no facility in the software for any individual teacher expectations to be associated with any essay of any set of essays.

6.1.14 Essayist's appearance – subjective effect is removed

Only the essay has to be marked by the software, and only the essay has to be considered. Other than the marking schema no other input is required or expected. This means that the appearance of the essayist is irrelevant in the marking process.

6.1.15 Size of the essay set – subjective effect is removed

As only one essay is marked at any one point of time, and each time an essay is to be marked the appropriate settings or variables are reset to their initial values. Size of the essay set is irrelevant, due to the fact that the size of the essay set is effectively one. Therefore the order in which essays are marked in the essay set or the size of the essay set or the size of each essay should have zero effect on performance.

Of course the more essays in the essay set, the bigger each essay is, and the more complex the content marking schema is will each have an effect, but this effect will exhibit itself in reducing the throughput rate. Throughput rate, that is the volume of work being completed in unit time, and marking performance is not linked but are completely independent aspects. Throughput rate will reflect the marking demands placed on the automated marking system.

For any computer, time and date are mere parameters, and as such may be used as a trigger to commence an offline or an unattended session of automated marking. Offline is a mode of computer operation that is often used to process cheques, print statements, and so on, usually overnight, so that (in the morning) staff have the results to hand without having to wait. The computer has no concept of becoming tired; which means the computer performs its task without degradation of performance. As computers do not tire like humans, automated offline marking may be conducted at any time. When an essay is marked by computer or indeed by a human marker, the time and date are irrelevant to the marking process and best used for time stamping process logs.

Process logs are, of course, necessary for the level of record keeping normally associated with any data processing environment in general and in particular an academic environment. Process logs are likely to contain information such as start / finish date and time, essay set identifier, number of essays marked and specific essay information – essay identifier, marks, feedback, and so on. These logs would be for reference use only, not for the process of marking.

6.1.16 "Own Tutee" effect – subjective effect is removed

This effect can only be removed, as the software should not be permitted to recognise the identity any essayist. All essays must be, and will be, treated equally.

6.1.17.1 Essay length 1 – subjective effect is <u>removed</u> for **content**

The only effect essay length should have is that a longer essay would necessarily take a longer time to process. This is the only effect that would be permitted. As the author's software package, SEAR, operates at the sentence level for the marking of **content** the overall length of the essay is never considered.

6.1.17.2 Essay length 2 – subjective effect is <u>reduced</u> for **style**, negated for **content**

For marking style, SEAR does have candidate metrics for the total number of words and the number of different words, and therefore essay length does affect the mark for **style**. However these metrics are objective in nature. It is legitimate to use both these two metrics in the automated marking for **style**.

For marking content SEAR only counts the words to give an impression of the essay length. SEAR does not use essay length for any part of the marking of **content**.

6.1.18 Inter-marker – subjective effect is <u>reduced</u>

This aspect will be both reduced in effect and shifted in time. It is reduced in effect because the examiner has to develop a highly objective formatted marking schema. The more detailed the marking schema then the less the opportunity presents itself for differences to occur between markers. This is true for both manual and automated marking procedures, and for many other procedures in everyday use across the globe.

Minimising variation and sources of variation or conflict is one of the mantras of improving quality in systems. The shifting in time should occur because examiners and second markers should remove conflict(s) in the marking schema before the first essay is marked, rather than, as at present, waiting until the whole set of essays has been marked by the examiner and then to have the second marking started. Any change in the essay set environment such as alterations to the content marking schema must necessitate a complete rerun of the marking process.

6.1.19 Intra-marker – subjective effect is removed

The same essay in the same essay set must be awarded the same marks every time the automated marking software is run. In the specific task of marking for content the marking schema would have to remain the same; while in the specific task of marking for style the same sample essays used for developing the linear model would have to remain the same. If the sampling of essays for style marking were to be flawed in some way, then using a different sample would result in a different linear model being produced.

Clearly linear model differences are solely due to the sampling, not the marking software. Note that should the sampling be flawed in any way then this will have a disastrous effect on whatever activity the sampling is required for – not just for the special case of automated essay marking. To enable a repeat marking process to be facilitated there may, of course, be the requirement to reset log files, reset system parameters and all the other similar data.

6.2.1 Achieving the potential of automated marking

In section 1.6 a series of key questions were posed; now it is time to determine how well these questions have been answered. For ease of reference, these questions are repeated below:

What, if any, are the problems of manually marking essays?

Most of the literature reviewed in Chapter 2 describes the many multi-facetted problems that beset manual essay marking. The author expected that **some** problems would be found; however the scale and range of the problems that were identified surprised even him. Marking essays manually proved to be **very** problematic.

What, if any, is the potential of using computers to mark essays?

The research work of this author, allied with the work of other researchers, does confirm that computer software is potentially capable of marking word-processed essays. That capability is, perhaps, small at present, but the software is slowly being improved and refined in order to maximise the potential of automated essay marking.

Is there any existing software that will mark essays?
If not, could such software be developed?
How effective could this software be in assessing essays?

Taking these questions together, it is clear that software now exists that is **potentially** capable of marking word-processed essays. The author's software can mark essays with an effectiveness that is not dissimilar to that displayed by human markers. This effectiveness is increasing along a paradigm that may prove similar to that of the development of powered flight.

How far along Bloom's Taxonomy would automated content marking be possible?

In Chapter 1, the author pondered how far along this taxonomy automated essay marking could traverse. It would appear that the author's software copes effectively with the lowest cognitive level [knowledge] but is not yet capable of handling higher cognitive levels. However, still using the analogy of flight, who is prepared to speculate what future software will be capable of? The collection of a wider subject range of essays, together with a wider coverage of Bloom's taxonomy, will accelerate further development of the SEAR software. Each stage of the development will improve software robustness and raise the taxonomic range covered.

6.2.2 Against the acceptance criteria of automated marking

This section revisits the criteria that Kaplan (Kaplan, 1992; 1998) proposed for automated (essay) marking. The author introduced the reader to Kaplan's set of criteria in Chapter 2 of this thesis.

SEAR

When this author has completed his PhD research work, and when he has also completed any required technical software enhancements then he will actively pursue the fullest field testing of his software system called SEAR.

The field-testing will not solely concentrate on just the technical marking essays but on the acceptability of using the SEAR package itself. The field-testing is envisaged to include the opinions of the examiners, the markers and of course the essayists. Likewise the acceptability of having essays marked by SEAR or by any other appropriate software marking packages will be canvassed. Having the opportunity of using any software, no matter how perfectly it performs, does not in itself confer acceptability in its use.

One particular aspect of field-testing is to determine what type or types of essay are better for the application of automated essay marking. Since the SEAR software may be applied for style marking or content marking or both, an additional complication is created. There may arise the situation that acceptable performance may be obtained for style marking certain essay type(s) in some specific subject(s) at certain level(s). However a different set of subject and level combination(s) may be better suited for content marking. Hopefully there may be level(s) within subject(s) that are suitable for both style and content marking together.

Through human history there have been many cases where the possession of some technology, or another, has not been acceptable.

In the field-testing questions will be raised on marking performance with an equal emphasis on acceptability of the procedure, especially from the essayists. This author does not see how any benefit(s) may arise if the examiners and the markers apply any software to the marking of essays only to have the essayists rebelling against, or appealing against, the marks awarded by such software. All the time and operational benefits would be lost in having to deal with all the extra operational workload imposed by the essayists' legitimate concerns.

The set of criteria to be discussed below is (fully) comprehensive from the organisational and the examiners' cum the markers' points-of-view. Unfortunately the same set of criteria are woefully incomplete from the point-of-view of the essayists. There is only one of the criteria (see 6.2.2.6) that focuses on the acceptance of the automated marking, but that is limited to acceptability to human markers. There is no mention of acceptability to essayists, examining bodies, awarding bodies, employers and so on. This is the closest the criteria get to considering the essayist's views.

Direct consideration of other essay marking software will be included, but only when the relevant facts are known for these packages. The reader will not be surprised to learn that most of the developers of automated marking software packages are rather silent on the criteria they use to show how good their software is, other than the accuracy of their software vis-à-vis manual marking.

This author now seeks to review his own software package, SEAR, against the set of the criteria of acceptance. Since SEAR deals with both the style marking and the content marking of word-processed essays this may expose any differences occurring between marking style and marking content. Where these differences occur in the following discussion of any of the criteria, these differences will be highlighted.

The reader is reminded that in this author's research project the style marking of essays is not as well developed as it the content marking. This is solely due to the failure of the author to obtain marked essay sets for style.

6.2.2.1 Ease of creating a scoring schema

For <u>style</u> the answer is either "it is" or "its not" depending on the possibility of the identification of a common set for metrics. The existence, and hence the deployment, of a common set of metrics that may be applied to all essay topics and at all levels of academic studies would greatly promote the use of software to produce marks for essay style.

However if such a set of common metrics are in fact determined then there will be less need for training; some sample essays would have to be checked by a human marker before the whole essay set is marked by the software.

If there is no identifiable set of common metrics then the marker(s) have to produce a marked sample set of essays that have been taken from the essay set that is to be marked.

The more metrics used in the software then the bigger the size of the sample set essays has to become. To be statistically correct the size of the sample essay set has to contain at least twice the number of marked essays than the number of metrics to be used by the software.

For content SEAR requires very minimal input form the examiner. All that is required is a detailed marking scheme (should be present anyway) and a model essay answer for checking that the content marking software is operating correctly. In essence a (little) more effort by the examiner to begin with. SEAR requires the detailed marking schema to be created by a simple text editor (any ASCII editor would do) and to be laid out following a small set of simple rules.

Other software packages require considerable set-up activities. For these packages several thousand words, on the specific essay topic to be assessed, have to be processed in order to create the required word associations, co-locations and so on before any essays may be marked. Even then, some tuning or calibration may be required to check that the marking procedure is operating correctly.

6.2.2.2 Ability to score on various mark regimes

All SEAR reports are exportable to standard spreadsheet software. This export facilitates the expression of the results in whatever format the reporting of the marks is required. In the case of, for example, the need to express the result as a band, then it is easy for a column of the spreadsheet to be created for this conversion to be accomplished. Technically all that is required is a "LOOKUP table" function or a suitable alternative function. In the case of performance criteria, the same approach as for a conversion to a band may be used for both style and content. Further, for content marked by SEAR, the model may be set-up in the schema to mark performance criteria directly.

6.2.2.3 Ease of identification of non-scoring elements

For style those words, clauses, sentences and paragraphs that adversely affect the marks awarded could be flagged, as could spelling errors and grammar errors. This flagging, or annotation, should be placed in the essay where needed with or without a summary at the end of the essay or as a separate item. This would be better done on a copy of the essay, as the original essay should be kept as is for a variety of reasons. These reasons for keeping the original essay as it was submitted range from legal to the need to remark. The legal requirement stems from the need to avoid tampering with the essayist's own work.

The need to remark, regardless of how or why that requirement to remark is created, has to be performed on the same basis as the original marking, otherwise the remarking may not be as creditable nor be directly comparable with the original marking.

In the case, for example, of sentences being non-relevant vis-à-vis the expected content, or being the mere repetition of previous content, then these sentences could be annotated in a manner that is similar to the annotation for style outlined above. Likewise there could be annotation for those sentences for which any matching against the content schema proves fruitless.

For <u>content</u> it is easy to identify those elements of the marking schema that are not scored by each essay; thus it is easy to show where the marks were not awarded. From the standpoint of the essay and essayist it is a good bit harder to show the non-scoring elements.

It should be easy for any content marking software to list those non-scoring elements to the essayist. To indicate to the essayist where those non-scoring elements should be positioned in their essay would possibly prove to be difficult.

6.2.2.4 Ease of modification should scoring errors occur

For <u>style</u>, in all the software packages scoring errors are likely to be sourced in (a) an invalid sample that was picked to develop the mathematical model, or (b) if the human marking of this sample is invalid. Both sources of errors may be present.

The correction of these scoring errors, however caused, requires the taking of a second sample from the essay set. This second sample would have to be marked by human marker(s), then the mathematical model to be generated for a second time and so on. This is not a quick or easy procedure to undertake lightly.

For the other software packages the scoring errors of <u>content</u> may be caused by the choice of the texts used to create the various relationships between the words that are contained in the texts. In this case the correction action would be horrendous to achieve, certainly it would not be easy to accomplish.

However for SEAR, with its examiner generated simple answer schema for <u>content</u>, then any changes that are required should be straightforward to complete. Re-processing the essays from the revised content schema should take no more effort than did the first content marking processing.

Should the generation of the content data structure from the answer schema be automated then it would become even easier to correct any scoring errors.

6.2.2.5 Consistent, reproducible scoring

All the software marking packages must mark to the highest standards of consistency and reproducibility. One of the main reasons for the widespread use of computers in today's modern world is the fact that all computers operate in a consistent reproducible manner, and there is the highest expectation among computer users that this remains so.

Further, one of the many axioms for using computer software to mark essays is to remove subjective effects of manual marking, where the presence of subjective factors in manual marking lead to inconsistency and irreproducible operation.

6.2.2.6 Acceptability of scores or results to human markers

It is clearly axiomatic that the human markers who use automated marking must find the marks so produced acceptable. There would be little point in markers performing automated marking in the full knowledge that the output from automated marking is to be rejected out-of-hand.

However there is a hierarchy of acceptability that had to be explored. To be completely acceptable automated (essay) marking has to be acceptable at each and every level of the hierarchy. Each level of the hierarchy automated marking has different criteria to satisfy for acceptance. When all the criteria for acceptance in all the levels of the hierarchy are satisfied then automated marking will be deemed as being wholly acceptable.

The lowest level (that is an operational level), of the hierarchy is where Kaplan's criteria is used.

The next level in the hierarchy is tactical - tactical in the sense of whether automated essay marking is acceptable as a methodology. Automated objective marking has reached a market size, or presence, where there is sufficient deployment practice in many academic organisations to establish tactical acceptance of such marking. However for automated essay marking there is not the same practice presence; in fact the presence is small. Although the presence is small today there seems to be a growing interest in automated marking of free text responses and its bigger relation essays. Who may accurately predict what the practice presence will be in five to ten years from today?

Still at the tactical level this author has met with several examiners who steadfastly reject the notion that any software can mark an essay because they think that any non-trivial essay is just too complex an entity for any software to safely mark. These examiners are possibly prepared to accept automated essay marking, once it can be clearly demonstrated that the marks generated by automated essay marking software are valid. Equally this author has met a few examiners who find the idea of even considering using software to mark any essay abhorrent and the reader will not be surprised that these examiners will reject outright all automated essay marks per se.

Thus at the tactical level of the hierarchy of acceptance, automated essay marking may be stymied depending on which of the body of examiners the balance of opinion makers lie - those for, those awaiting to decide, and those decidedly against automated essay marking.

The highest level in the hierarchy is strategic. Assuming that the lower hierarchy levels of operational and tactical are satisfied then what acceptance criteria is left to be satisfied. At the strategic level the criteria for acceptability are analogous to that currently facing automated objective marking, namely which subject or subjects coupled at which level or levels within these subjects should automated objective marking be deployed. Likewise for automated essay marking the criteria is founded on selecting what level(s) in which subject(s) should automated essay marking be employed.

However there are two additional criteria that this author foresees as having to be satisfied namely that of essay type and the purpose of marking. These two additional are outlined in the following pair of paragraphs.

One part of the field-testing mentioned in section 6.2 above will be concerned with the establishment of the type(s) of essays that are suitable for only style marking, only content marking and those suitable for both style and content marking at the same time.

Another part of the field-testing will seek to determine the best purpose of marking. Should automated essay marking be limited to style only or content only? Or would automated essay marking be better applied to both style and content to the maximum appropriateness? When essays are being marked for content it may prove beneficial to provide some feedback on the essay style to the essayists even if the examiner deliberately makes no allocation of marks for essay style?

Also at the strategic level those in charge of academic organisations must consider the impact on their stakeholders of the introduction of automated marking methodologies in general, yet alone automated essay marking. For any academic organisation the internal stakeholders are their students and their staff, while the external stakeholders are the various funding bodies, awarding bodies and a myriad of professional bodies.

Generally professional bodies tend to be very strict in how the course(s) operate that they, the professional bodies, award their recognition to. This author would expect that any changes in course operation that could give rise to the alienation of professional bodies would not be lightly undertaken by any academic strategist.

6.2.2.7 Defensibility

Chapter 2 clearly shows that manual marking is very susceptible to subjective effects acting on the human markers. The inverse view of this is that with large subjectivity in the marking process then the marks awarded may "not be defensible". "Not defensible" in the sense that the marker may not be able to fully explain why he or she awarded that mark at that time to a specific piece of (essayist's) submitted work.

For automated essay marking all the marks awarded have to be defendable. This criterion is not too difficult to satisfy. First general level of subjectivity is very much reduced when automated marking is deployed. All the marks so awarded must be consistent and reproducible and as computer systems are being used then there is high expectancy that the marks awarded will be consistent and reproducible. Second the software can be programmed so as to show how the marks were awarded to those with a legitimate requirement to have that information.

6.2.2.8 Accuracy

It is axiomatic that any marking software has to be accurate. Without at least comparable accuracy vis-à-vis manual marking then the software is not acceptable with any further comment needless.

This criterion complements a pair of criteria covered above in the section on consistent, reproducible scoring (6.2.2.5) and the section on the acceptability of scores or results to human markers (6.2.2.6).

6.2.2.9 Coachability

There must be no opportunity for the essayist to be coached into creating an essay that will receive better than deserved mark awarded. Of course nearly every essayist has the expectation that when presented with the essay topic then some, more or less limited, information on how to approach the expected solution. This information may take many forms, but regardless of the format of the information it will in all likelihood be given on a cohort wide basis.

All students (essayists) are given the same information at the same time. This represents coaching that is acceptable. Coaching where one essayist is in receipt from support from one tutor in the deliberate attempt to gain a better than deserved mark by the expression of the essay, or its format, is not acceptable.

For style, in the case of the main software packages for marking style, for effective coaching to occur, the essayist must be able to reproduce the mathematical model to be used in such marking. The same model development environment would need to be replicated. Therefore the same texts for building the model would need to be found and then processed in exactly the same manner. Under such exacting circumstances surely the essayist would find it easier to better their essay writing style than to "beat the system"?

For SEAR, and assuming that a common set of metrics has been identified, then coaching may not be that possible. There are two features of SEAR that minimise coaching opportunity. First there is no need to let the essayist know what metrics are being used to mark style (the same for other software style markers). Even if a common set of metrics is known to the essayist and the would-be coaching tutor there is still the second feature of SEAR to minimise coaching, that of the manual mark sample set. The examiner marks a sample of the essays from the essay set in order to calibrate, or train, the software. For coaching in style marks the essayist must be able to have access to the essays that have been selected to be in the sample set for calibration and the marks manually awarded to these essays. Since the selection and the marking both occur after the submission deadline, the essayist will then have to travel back in time to submit their coached essay on, or before, the submission deadline. Travelling backwards in time is a little beyond today's technology.

For content the examiner may, or may not, issue the marking schema. If the marking schema is issued to the essayists then that is for the benefit of the essayist.

For SEAR the detailed answer schema must not be revealed to the essayist, thereby eliminating coaching. If coaching for content were to happen then for that one essay the coverage of the content data structure would be high, as would the usage. Reporting both the coverage and the usage as being high then the examiner may wish to explore this essay manually in the light of the examiner's knowledge of the essayist. The author imagines that an essay so constructed to completely match the content data structure used for marking the content would probably show a very poor essay style.

6.2.2.10 Cost

For style SEAR requires a marked sample set of essays drawn from the essay set. The number of essays in the sample must be at least twice the number of metrics in the SEAR marking algorithm for style. Thus the examiner is left with a decision to use SEAR or not. If the essay set is not much bigger than twice the number of metrics used then there is no real benefit to the examiner in using SEAR. However, for example, for a cohort of 500 students then manually marking 50 sample essays and then using SEAR to mark all the 500 essays clearly shows a benefit for the cost involved. This is especially true when the 50 used as a sample would have to be manually marked anyway as would be the other 450 essays in the essay set.

The other style marking software also requires calibration in order to build the underpinning mathematical model. Generally these packages require a considerable number of marked sample essays to function correctly. The larger the sample becomes the higher the break-even point for the examiners becomes.

The other content marking software do require considerable set-up costs in that sample texts have to be identified by the examiner. These texts then have to be processed to create the necessary word relationships that underpin the use of these software packages.

For content SEAR requires very little cost to the examiner. Good examiners will readily produce excellent answer schemas for their essay topics. In using SEAR the cost to the examiner is in the construction of the answer schema in compliance with the few simple rules of construction. There is, however, one additional cost namely that of the production of a model essay answer. The model essay answer is used by SEAR to confirm that the content data structure has been correctly constructed from the examiner's answer schema. Rarely do examiners supply any model essay answers, especially in prose format.

6.2.3 Review acceptance criteria of automated marking

In this sub-section this author combines marking essay style and marking essay content into the one term of essay marking. This author contends that in terms of the criteria already discussed in this chapter and the ones yet to be discussed there is not substantive reason to keep style marking and content marking separate.

Considering these ten criteria discussed in this chapter they appear to be far ranging in scope and, where necessary, quite severe in nature. However, there are three gaps in the range. These gaps are essayists, external examiners and professional bodies.

Essayists

Is automated essay marking acceptable to the student body at large? There has been some research into finding out if students find automated objective testing acceptable. In fact students do accept automated objective testing. Students do not find lengthy delays in the return of their courseworks acceptable. Given that automated essay marking is only really cost effective for large scale or large sized classes and that such classes are predominantly first year or stage 1 operating at the lower end of Bloom's Taxonomy then it may be acceptable to use this marking methodology. For students in the later years or the later stages of their degree then this author expects that their acceptance of automated essay marking will lessen and possibly rightly so. If the mass of essayists are against automated essay marking then this author foresees that trouble between academics and essayists could arise. Nevertheless it is the academics that set the submission requirements and also set the marking methodology. The reader must bear in mind that students are not required to like the assessment strategy and that some students hate essays, other students hate examinations. How often do academics ask their students if they like the assessment strategy, assessment instruments, and the marking methodology?

External examiners

Many academic organisations have already large-scale deployment of automated objective testing. Therefore many external examiners must be deploying such a methodology. When sufficient numbers of academics are satisfied with automated objective testing then the time may have come that automated essay marking becomes accepted. This author suggests that going from manual to automated objective marking may at the present time prove to be too great a step for the external examiners, yet alone adopting automated essay marking.

Provided the proper development and pilot studies are conducted to show that the comparison with manual testing with automated essay marking then the external examiners may come to accept automated essay marking. This acceptability by the external examiners may be, on a course-by-course basis or a school-or-school basis.

However external examiners are more concerned with award bearing decisions of the later years or the later stages of the course and they tend to have only passing interest in the years or stages when automated essay is most likely to be deployed. Therefore the approval of external examiners may not be a heavy weight criterion to satisfy.

Professional bodies

Like external examiners, professional bodies are concerned with the awards that the students receive. Professional bodies rely on the quality and integrity of the academics and external examiners to maintain award standards. Often professional bodies are populated by academics who act as external examiners, and who themselves may be using automated marking methodologies.

In closing this sub-section this author suggests that automated essay marking is possible, and is apparently acceptable against the criteria covered in this sub-section. There is not a single criterion failure. This author is not aware of any public failure(s) in automated essay marking. There has been far too many public sandal(s) and public failure(s) of manual marking. So why are most essays being marked manually? Why the raised eyebrows, rolling eyes and other (strong) signs of disbelief and contempt when this author mentions automated essay marking?

- Could there be a lack of awareness of the methodology?
- Could there be resistance to change in methodology?
- Are markers fully aware of the subjective factors in marking?
- Could there be an attitude against automated essay marking?

Perhaps the answer to the last question of the four is the key to the future adoption of automated essay marking.

6.3 Possible future research work

There are several disparate potential routes to further research emanating from the current state of research in this PhD. The first two are natural, or not un-expected, research development routes. The remaining research routes constitute new major branches from the algorithms at their current stage of development. All these research routes are now detailed below.

The most promising way to accomplish these developments may be to create a new research group or create a sub-group within an established research programme.

In Chapter 2, the work of two research pair's (Borja and Spader (1985) and McDonald and Samson (1979)) in the area of providing structured objective feedback was outlined. For style marking and for content marking the development of a generic list of feedback codes may prove to be an important next step. This generic list may be based on Borja and Spader like codes augmented with McDonald and Samson like attachments. Borja and Spader / McDonald and Samson pairings both worked within the confines of a paper-based medium. However with the computer-based medium used in automated essay marking it could be desirable to directly enter the feedback directly into a copy of the essay, and hence return the copy essay to the essayist. This way of providing feedback combines the best of both the Borja and Spader / McDonald and Samson approaches; when coupled with the power of the computer it could raise the provision of feedback to unheard of (possibly even previously unthought of) levels of performance.

6.3.1 Style algorithm

Development of the STYLE algorithm was halted due to the unavailability of style marked essay set(s). When such marked essay set(s) become available then development may be restarted as a two-route (or as, possibly, a two-stage) development.

6.3.1.1 Complete as is

This first route is to continue to develop this algorithm with potential plain text metrics that the author has already identified, augmented by further plain text metrics as they are discovered by the author. The author has confined the current research to plain text metrics as this facilitates comparison with the work of other style researchers. Furthermore this approach facilitates a very broad range of potential metrics.

6.3.1.2 Extend with word-processing sourced metrics

This second route is to extend the set of plain text metrics with metrics sourced from use of word-processors. This author has no evidence of other researchers using metrics arising from the use of word-processors. Word-processor sourced metrics are limited in range, namely ~

- text enhancement (use of bold, italics, underlining and combinations thereof);
- fonts (type, size, colour and again combinations thereof);
- textual style; and
- combinations of all of the above.

6.3.2 Content Algorithm

This algorithm appears to work well for some types of essay and is designed for only textual elements of an essay. There are therefore two routes to extending this algorithm.

6.3.2.1 Increase range of essays for content algorithm

The first route is to improve the algorithm to cover more essay types ~ always provided that a content schema can be produced!

The range of subjects that may be subjected to automated essay marking would need to be determined. Knowing which subject area(s) are ideal and which are to be avoided would mean that the application of an in-appropriate methodology is not performed.

6.3.2.2 Extend content algorithm with non-textual elements

The second route, and this may prove to be hard to pursue, is to include non-textual elements in the content algorithm. Non-textual elements are considered to be tables, graphs, figures, and so on. This route is a desire, or perhaps a drive, to increase the utility of the algorithm by including these non-textual elements as modern word-processors facilitate them. Although this route may be potentially difficult to achieve, the rewards of success therefrom are expected to be very fruitful.

6.3.3 Extend application to use non-English languages

This author developed both his algorithms using English, as this is his native language. However the author made the decision to keep the algorithms non-language dependent as far as possible. Expanding into non-English language(s) would definitely require support from language experts. The specific nature of what non-English language support is required is obviously outwith the author's control at this stage and could be determined by both research and commercial imperatives.

Non-English ASCII based languages are candidates for automated essay marking, surely as must be essays written by native French essayists being style marked by French academics, and native German essayists being content marked by German academics, and so on. Automated style marking in languages other than English superficially appears possible without much alteration to the SEAR style marking algorithm. However for automated content marking there appears to be a major obstacle when using languages other than English. This obstacle is that the use of Roget Thesaurus is a major component of the SEAR content marking algorithm and it is only available for the English language. This author is not aware of any non-English language(s) equivalent of Roget Thesaurus.

6.3.4 Develop anti-plagiarism mechanisms

Sadly noting the rising tide of plagiarism sweeping academia, the author has deliberately developed his algorithms with a requirement that should both the algorithms be successful, then anti-plagiarism measures may be developed without major re-development of any of the algorithm. There is a statistic, called the Cosine Angle, that may be applied to, for example the Schema Report, to identify close pairings between essays. The Cosine Angle statistic is used determine the closeness between two, or more, vectors. In this situation the concept of a vector maps unto an essay. Other plagiarism detection measures, such as the Horn statistic (Dollard and Mowrer, 1947), may be deployed.

6.3.5 Develop measurement of specific essay traits

This author would concentrate on specific essay traits by first replicating the research work of two researchers who worked on the measurement of specific essay traits. The two researchers' areas are Dollard and Mowrer (1947) on measuring tension, and Hiller, Marcotte and Martin (1969) on measuring opinionation, vagueness and specificity-distinctions. Having replicated these researchers work, albeit possibly augmented by the work of other researchers, then the exploration of what ever additional traits that the expert linguistics have identified.

6.4 Research based technical improvements

There are a few areas where some technical improvements are possible, nay may even be required, and that require some research before these improvements may be realised. Most, if not all, of these improvements require the input or advice from properly qualified linguistic experts. The author of this research work is not a (computational) linguistic expert.

6.4.1 Wordlist(s)

Obtaining more comprehensive wordlist(s) would help both the style and content marking, by inter alia providing larger and / or more appropriate wordlist(s). An example of this is the comparison of the Roget's Thesaurus and each of the two different formats of the Oxford and Collins Thesauri. This comparison would not be on the basis of the number of words present, but that the comparison will be made on the structure of how these words are classified within the three different thesauri (Roget original, Oxford A-Z and Collins number).

Lists of phrases based on Roget's Thesaurus have been prepared – researching the incorporation of these lists into the content marking process may lead to some technical improvement.

6.4.2 Foreign Language

For content, one of the main factors in the marking algorithm is the use of Roget classified wordlists. The author is not aware of any similar non-English language reference work. Therefore before operating in non-English language(s) there is a requirement of obtaining, either directly or by creation, of various Roget comparable wordlists.

6.4.3 Control of Tense

One problem that has occurred that needs, quite possibly considerable, linguistic technical research is that of controlling or analysing sentences that have verbs not in the present tense. Even the casual examination of the collection of wordlists used in this research project show that these wordlists are essentially of one tense – the present tense. For example "sit", "stand" and "walk". The problem is with the words generated by the different tenses. For example past tense "sat", "stood" and "walked". The English language in heavily populated with irregular verbs. Proper linguistical advice is required in this particular matter. The development and use of a robust stemming algorithm is a highly probable system requirement.

6.4.4 Control of Plurals

Taking the most casual examination of the collected wordlists used in this research project there very few plurals to be found. To be confident of handling plurals correctly some technical linguist research will need to be completed into the handling of regular and irregular plurals. For example the plural of seat, man and mouse are seats (regular), men (semi-regular) and mice (irregular – for the animal) or mouses (regular – for the computer peripheral). Again, proper linguistical advice should be sought here.

6.4.5 Passive Voice versus Active Voice

Allied with the paragraph above on control of tense there is a need to develop better means of handling the two different voices possible in text – that of passive and active voices. For example consider the following two sentences – "The cat sat on the mat", a sentence expressed in the active voice, and "The mat was sat on by the cat", which is a sentence expressed in the passive voice. Clearly these two very simple sentences are equivalent in content and meaning. Equally clear is that the change of word order coupled with the addition of the word "was" in the correct place converted the active voice sentence into the passive voice sentence. However such clarity cannot always to be present, therefore a few very simple rules will not suffice to convert passive voice sentences into active voice sentences. As the algorithm for content marking, as it currently developed, appears to operate better with sentences and essays expressed in the active voice than the passive voice, there is a requirement to operate effectively for both voices.

Therefore some technical research will need to be completed to maximise the analysis of sentences which are expressed in either active or passive voices. Yet again, proper linguistical advice is required here.

6.4. 6 Pronoun Use

The English language makes heavy use of pronouns, where a pronoun is used in place of the previously named person or item. For example, consider the following set of texts:

A] James donated a book to the library. The library was pleased that James donated a book to the library.

B] James donated a book to the library. They were pleased that he did so.

C] The library was pleased to receive his donation. James had donated a book to them.

These texts are, or should be, equivalent in meaning. Yet the second and third versions should appear to most readers as more natural expressions, whereas the first version by being more explicit in expression it becomes more a turgid text.

When marked by computer, then the first expression would be easier to mark than the other two expressions. Pronoun use therefore firstly requires some memory, or persistence, of earlier nouns to be retained [second version] and secondly also requires flagging for future use [third version].

6.4.7 Handling spelling errors and grammar errors

Grammar errors affect the content marking of essays as incorrect tense, or incorrect phrasing may confuse the content marking algorithm. Should the amount of affected text be large enough then the marks awarded to an essay may be falsely suppressed or falsely inflated. The number of grammar errors and the profile of these grammar errors could be the source of a small series of metrics to be used in the marking of essay style.

Spelling errors will affect the content marking of essays as those misspelt words may not be matched, in fact **will not** be matched, with any elements of the content schema. In this case there will be a suppression of the marks awarded to the essay. There is a very slim chance, *a priori,* that any misspellings would lead to any inflation of marks awarded. The number of spelling errors and the profile of these spelling errors could be the source of a small series of metrics to be used in the marking of essay style.

6.5 Concluding remarks

It can be seen from the above that the author's work has not merely broken completely new ground in the field of automated assessment, but has also shown how the work can be extended in a number of potentially fruitful directions. He himself intends to explore some of these, and he hopes that other workers will also feel that it is worthwhile becoming involved in this highly important field.

References

Allot, N; Fazackerly, P; Halstead, P

A knowledge driven aid to the automated assessment of free text

AISB Quarterly: 1st Biennial workshop, 1994a, 88, 19-24

Allot, N; Fazackerly, P; Halstead, P

Automated assessment: evaluating a knowledge architecture for natural language processing.

Applications and Innovations in Expert Systems, 1994b, 2, 319-334

Billing, D

Chapter 25 Criteria for Essay Marking by McDonald, R; Sansom, D

Indictators of Performance ed Billing, D; 1979, 168-169

Bishop, R L

Computer analysis of natural language text for style and content in the context of instruction of writing

Computers in the Undergraduate Curricula Uni IOWA Conference, 1970,

June, 1.7 - 1.11

Blok, H

Estimating the Reliability, Validity, and Invalidity of Essay Ratings

Journal of Educational Measurement, 1985, 22(1), 41 - 52

Borja, F; Spader, P H

AWK: Codes in grading essays

College Teaching, 1985, 33(3), 113-116

British Standards Institute

BS 7988 A code of practice for the use of information technology in the delivery of assessments

British Standards Institute, 2001, 18 October, 1 - 40

Brown, S; Rust, C; Gibbs, G

Strategies for Diversifying Assessment in Higher Education

The Oxford Centre for Staff Development, 1994, 1 - 51

Burstein Kukich Wolff Lu Chodorow Barden-Harder Harris

Automated Scoring Using A Hybrid Feature Identification Technique

http://www.ets.org/research/dload/acl99rev.pdf, 1998, 1 - 5

Burstein, J; Kaplan, R; Wolff, S; Lu, C
Using Lexical Semantic Techniques to classify free-responses
http://www.ets.org/research/dload/siglex.pdf, 1996, 1 - 12

Burstein, J; Leacock, C; Swartz, R
Automated Evaluation of Essays and Short Answers
5th International Computer Assisted Assessment Conference, 2001a, 2-3 July, 42-54

Byrne, C
Tutor marked assignments at the Open University: a question of reliability
Assessment in Higher Education, 1980, 5(2), 150-167

Carter, RS
How invalid are marks assigned by teachers?
Journal Educational Psychology, 1952, 43, 218 - 228

Cast, B M D
The Efficiency of Different Methods of Marking English Composition: Part 1
British Journal of Education, 1939, 9, 257-269

Cast, B M D
The Efficiency of Different Methods of Marking English Composition: Part II
British Journal of Education, 1940, 10, 49-60

Catterall, J
What is a University Essay?
http://www.macarthur.uws.edu.au/ssd/ldc/Essay.html, 1994, 1-4

Chapman, R L
Roget's International Thesaurus
HarperCollins, 1996

Chase, C I
Essay test scoring: interaction of relevant variables
Journal of Educational Measurement, 1986, 23 (1), 33-41

Christie, J R

Computer-assisted Assessment of Essays

2nd Computer Assisted Assessment Conference, 1998, 17-18 June, 85-89

Christie, J R

Automated Essay Marking - for both Style and Content

3rd Computer Assisted Assessment Conference, 1999, 16-17 June, 39 - 48

CLASS Centre for Learning and Assessment

Chapter 6: Essays & Reports

Student's Guide to Effective Study, 2000, 17 - 21

Cockburn, B; Ross, A

Essays

Teaching Higher Education Series: 8 University of Lancaster, 1978

Coffman, W E

On the validity of essay tests of achievement

Journal of Educational Measurement, 1966, 3(2) summer, 151-156

Coffman, W E; Kurfman, D

A comparison of Two Methods of Reading Essay Examinations

American Educational Research Journal, 1968, 5(1) Jan, 99-107

Cohen, J

A coefficient of agreement for nominal scales

Educational and Psychological Measurement, 1960, 20 (1), 37 - 46

Cohen, J

Weighted KAPPA: Nominal scale agreement with provision for scaled disagreement or partial credit

Psychological Bulletin, 1968, (70) 4, 213-220

Crain, C

The Bard's Fingerprints, article by Caleb Crain

Lingua Franca, 1998, JulyAugust, 26-39

DeLoughry, T J

Duke Professor pushes concept for grading essays by computer

The Chronicle of Higher Education, 1995, October 20, A24 - A25

Dollard, J; Mowrer, O H

A method of measuring tension in written documents

Journal abnormal and social psychology, 1947, 42, 3-32

Donley, M

Marking Advanced Essays

English Language Teaching Journal, 1978, 32(2), 115-118

Dutch, R A

Roget's Thesaurus

Penguin Reference Books, 1966

Edwards, R P A; Gibbon, V

Words your children use

Burke Books, 1973

Erwin, PG; Calev, A

The Influence of Christian Name Stereotypes on the Marking of Children's Essays

Journal of Educational Psychology, 1984, 54, 223-227

Eysenck, H J

The validity of Judgements as a Function of the Number of Judges

Journal of Experimental Psychiatry, 1939, 25, 650-654

Fajardo, DM

Author Race, Essay Quality, and Reverse Discrimination

Journal of Applied Social Psychology, 1985, 15(3), 255-268

Finlayson, D S

The reliability of the marking of essays

British Journal of Educational Psychology, 1951, 21, 126-134

Fleiss, J L

Measuring nominal scale agreement among many raters

Psychological Bulletin, 1971, (76) 5, 378-382

Foltz, P
Demonstration of automatic essay grading and feedback with LSA
Peter Foltz, 1998, 1 - 2

Foltz, P W
Latent semantic analysis for text-based research
Behaviour Research Methods, Instruments & Computers, 1996, 28(2), 197-202

Foltz, P W; Laham, D; Landauer, T K
The Intelligent Essay Assessor: Applications to Educational Technology
http://imej.wfu.edu/articles/1999/2/04/index.asp, 2001, 1 - 9

Foltz, P; Laham, D; Landauer, T K
Automated essay scoring: Applications to educational technology
Peter Foltz, 1999, 1 - 6

Gajar, A H
A Computer Analysis of Written Language Variables and a Comparison of Compositions Written by University
Journal of Learning Difficulties, 1989, 22(2) Feb, 125-130

Gunning, R
The technique of Clear Writing: Appendix B (revised edition)
McCraw-Hill Book Company, 1968

Hales, LW; Tokar, E
The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay
Journal of Educational Measurement, 1975, 12(2), 115 - 117

Hall, C G W; Daglish, N D
Length and Quality: an exploratory study of inter-marker reliability
Assessment and Evaluation in Higher Education, 1982, 7(2), 186-191

Hamel, T
Grading by Computers: The Programs Don't Get an A, but They're Worth Considering
The Chronicle of Higher Education, 1988, October 12, B2

Hararp, H; McDavid, JW
Name Stereotypes and Teachers' Expectations
Journal of Educational Psychology, 1973, 65(2), 222-225

Harway, N I; Iker, H P
Computer Analysis of Content in Psychotherapy
Psychological Reports, 1964, 14, 720-722

Hearst, M A
The debate on automated essay grading.
IEEE Intelligent Systems, 2000, sept/oct, 22-37

Holmes, D
The evolution of stylometry in Humanities scholarship
Literary and Linguistic Computing, 1998, 13, 111 - 117

Holmes, R
Marks my words: Student essays can now be graded by machine
New Scientist, 1998, 12

Hughes, D C; Keeling, B
The use of Model Essays to Reduce Context Effects in Essay Scoring
Journal of Educational Measurement, 1984, 21(3), 277-281

Iker, H P; Harway, N I
A computer approach towards the analysis of content
Computers in Behavioral Science, 1965, 10, 173-183

Iker, H P; Harway, N I
A Computer Systems Approach towards the Recognition and Analysis of Content
Computer Studies in the Humanities and Verbal Behavior, 1968, 1, 134-154

Jacoby, H
Note on the marking system in the astronomical course at Columbia College, 1909 - 10
Science, 1910, May 27, 819

Johnson, V E

On Bayesian Analysis of Multirater Ordinal Data: An Application to Automatic Essay Grading

Journal of the American Statistical Association, 1996, March, 43-51

Kaplan RM & BA, Wolff S E, Burstein JC, Lu C, Rock DA

Scoring Essays Automatically Using Surface Features

GRE Research, 1998, August, 1-13

Kaplan, R M

Using a Trainable Pattern-Directed Computer Program to Score Natural Lanaguage Item Responses

GRE Research Report, 1992, April, 1-43

Kirkpatrick, B

The Concise Oxford Thesaurus, in clear A-Z from

Oxford University Press, 1997

Kniveton, B H

A correlational analysis of multiple-choice and essay assessment measures

Research in Education, 1996, 56 Nov, 73-84

Landis, J R; Koch, G G

The Measurement of Observer Agreement for Categorical Data

Biometrics, 1977, 33, 159 - 174

Landy, D; Sigall, H

Beauty is Talent: Task evaluation as a function of the performer's physical attractivness

Journal of Personality and Social Psychology, 1974, 29(3), 299 - 304

Larkey, L S

Automatic Essay Grading using Text Categorisation Techniques

SIGIR '98, Melbourne, Australia, 1998, August 24-8, 90-95

Markham, LR

Influences of Handwriting Quality on Teacher Evaluation of Written Work

American Educational Research Journal, 1976, 13(4), 277-283

Marshall, JC

Composition Errors and Essay Examination Grades Re-examined

American Educational Research Journal, 1967, 4(4), 375-385

Mason, O; Grove-Stephenson, I
Automated free text marking with Paperless School
6th International Computer Assisted Assessment, 2002, July, 213 - 219

McDonald, R; Sansom, D
Use of assignment attachments in Assessments
Assessment in Higher Education, 1979, 5 (1), 45-55

Michell, T; Russel, T; Broomhead, P; Aldridge, N
Towards Robust Computerised Marking of Free-Text Responses
6th International Computer Assisted Assessment, 2002, July, 233 - 249

Newstead, S; Dennis, I
The reliability of exam marking in psychology
The Psychologist, 1994, 7(5), 216-9

Nyberg, V R
Reliability of the Alberta Essay Scales
Alberta Journal of Educational Research, 1980, 26(1), 64-67

Oxford University Press
The Oxford Thesaurus, in A-Z form
Focus Multimedia

Page, E B
Grading Essays by Computer: Progress Report
Invitional conference on testing problems, New York City, 1966, 87-100

Page, E B
The Imminence of Grading Essays by Computer
Phi Delta Kappen, 1966, 47, 238-243

Page, E B
The use of the computer in analyzing student essays
International Review of Education, 1968, 14, 210-224

Page, E B
Computer Grading of Student Prose, Using Modern Concepts and Software
Journal of Experimental Education, 1994, 62(2), 127-142

Page, E B
Grading Essays by Computer: Why the Controversy ?
NCME Symposium, New York, 1996, April 11

Page, E B
The Second Blind Test with ETS: PEG predicts the Graduate Record Exams
AERA/NCME Symposium Grading Essays by Computer, Chicago, 1997, March 27

Page, E B
The Classroom Challenge and Write America !
AERA/NCME Symposium Grading Essays by Computer, Chicago, 1997, March 27

Page, E B; Keith, T Z; Lavoie, M J
Computer Grading of Essays Traits in Student Writing
Annual meeting National Council Measurements in Education, NY, 1996, April 11

Page, E B; Lavoie, M J; Keith, T Z
Construct Validity in the Computer Grading of Essays
Annual meeting American Psychological, New York City, 1995, August 13

Page, E B; Petersen, N S
The Computer Moves into Essay Grading updating the Ancient Test
Phi Delta Kappan, 1995, 76(7), 561-566

Peacock, M
Handwriting versus wordprocessed print: an investigation into teachers' grading of English Language and
Journal of Computer Assisted Learning, 1988, 4, 162-172

Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K
Comparing the Validity of Automated and Human Essay Scoring
GRE Research, 2000, June, 1-23

Powers, DE;Burstein, JC;Chodorow, M;Fowles, ME;Kukich,K
Stumping e-rater: challenging the validity of automated essay scoring
Computers in Human Behaviour, 2002, 18, 103 - 134

Roget, P M; Roget, J L; Roget S R
Roget's Thesaurus
Roydon Publishing Co. Ltd., London, 1972

Starch, D; Elliott, EC
Reliability of the grading of high-school work in English
School Review, 1912, 20, 442 - 457

Starch, D; Elliott, EC
Reliability of grading work in Mathematics
School Review, 1913a, 254 - 259

Starch, D; Elliott, EC
Reliability of grading work in History
School Review, 1913b, 21, 676 - 681

Thorndike
Educational Measurement
Educational Measurement, Chapter 10, 271

Tollefson, N; Tracy, D B
Test Length and Quality in the Grading of Essay Responses
Education, 1980, 101(1), 63-67

Townsend, M A R; Kek, L Y; Tuck, B F
The effect of mood on the reliability of essay assessment
British Journal Educational Psychology, 1989, 59, 232-240

Townsend; Hicks; Thompson; Wilton; Tuck; Moore
Effects of Introductions and Conclusions in Assessment of Students Essays
Journal of Educational Psychology, 1993, 85 (4), 670-678

Weinstein, B
Software designed to grade essays
Boston Globe Online, 1998, 21 June, 1 - 4

Wiseman, S
The marking of English composition in Grammar School Selection
British Journal Educational Psychology, 1949, 19, 200-209

Wresch, W
The Imminence of Grading Essays by Computer - 25 Years Later
Computers and Composition, 1993, 10 April, 45-58

Bibliography

Aborn, M; Rubenstein, H; Sterling, T D

Sources of contextual constraint upon words in sentences

Journal of Experimental Psychology, 1959, 57 – 3, 171 - 180

Adman, P; Warren, L

Frames of Mind

TLTP Workshop papers on Assessment in Learning in HE, 1993, 41-46

Adman, P; Warren, L

A strategy for educational technology in higher education

Journal of Computer Assisted Learning, 1994, 10, 50-54

Al-jarrah, M M; Torsun, I S

An empirical analysis of COBOL programs

Software - Practice and Experience, 1979, 9, 341-359

Allot, N; Fazackerly, P; Halstead, P

A knowledge driven aid to the automated assessment of free text

AISB Quarterly: 1st Biennial workshop, 1994a, 88, 19-24

Allot, N; Fazackerly, P; Halstead, P

Automated assessment: evaluating a knowledge architecture for natural language processing.

Applications and Innovations in Expert Systems, 1994b, 2, 319-334

Angeesing, J

Open-ended computer-marked tests

Teaching Earth Sciences, 1989, 14(1), 17-19

anonymous

Program to grade essays stirs debate

www.sjmercury.com/breaking/docs/005051.htm, 1998, 1 - 3

Applebee, A N

Microteaching, Component Skills and the Training of Teachers: an Evaluation of a Research and Development

British Journal of Educational Technology, 1976, 7(2) May, 35-44

Balla, J; Boyle, P

Assessment of Student Performance: A framework for improving practice

Assessment & Evaluation in Higher Education, 1994, 19:1, 17-28

Beckwith, R; Miller, G A; Tengi, R

Design and Implementation of the WordNet Lexical Database and Searching Software

WordNet, 0, 62-77

Benford, S; Burke, E; Foxley, E

Consciousness raising throuhg automated assessment

Learning Technology Research CS dept Uni Nottingham, 1997, 1 -5

Berry, R E; Meekings, B A E

A style analysis of C programs

Computing Practices, 1985, 28(1) Jan, 80-88

Beveridge, A

Automating Word from Delphi

A Beveridge, 1999, -, 1

Beynon, A L

An analysis of the Strategy Used in the development of a CAL Program in the History of Education

British Journal of Educational Technology, 1981, 12(1) Jan, 70-79

Biggs, JB

Study Behaviour and Performance in Objective and Essay Formats

The Australian Journal of Education, 1973, 17(2), 157 - 167

Billing, D

Chapter 25 Criteria for Essay Marking by McDonald, R; Sansom, D

Indictators of Performance ed Billing, D, 1979, book, 168-169

Bishop, R L

Computer analysis of natural language text for style and content in the context of instruction of writing

Computers in the Undergraduate Curricula Uni IOWA Conference, 1970, jun, 1.7 - 1.11

Bliss, J; Monk, M; Ogborn, J, 070990698

various extracts

Qualitative Data Analysis for Educational Research, 1983, various

Blok, H

Estimating the Reliability, Validity, and Invalidity of Essay Ratings

Journal of Educational Measurement, 1985, 22(1), 41 - 52

Boothroyd, D

Getting the message

New Electronics on Campus, 1998, Autimn, 6 - 8

Borja, F; Spader, P H

AWK: Codes in grading essays

College Teaching, 1985, 33(3), 113-116

Born, G, 1-85032-11

The file formats handbook

Thompson Computer Press, 1995

Borrow, D G; Winograd, T

An overview of KRL, a knowledge representation language

Cognitive Science, 1977, 1, 3-46

Bowers, R

Using html for online editing

http://www.vcol.net/swerner/rbowers/demo.html, 1999, 1 - 13

Bowers, R

Marking essays via html frames

http://cwis/waisp/www-data/lists/acw-l/9610, 1999, 1 - 6

British Standards Institute

BS 7988 A code of practice for the use of information technology in the delivery of assessments

British Standards Institute, 2001, 18 October, 1 - 40

Brown, S; Rust, C; Gibbs, G, 1-873576-2

Strategies for Diversifying Assessment in Higher Education

The Oxford Centre for Staff Development, 1994, 1 - 51

Bull, J

Using technology to assess student learning

Using technology to assess student learning, 0, 41-49

Burrows, G

Using computers to teach and test quantitative methods for business

British Journal of Educational Technology, 1996, 27(2), 144-146

Burstein Kukich Wolff Lu Chodorow Barden-Harder Harris

Automated Scoring Using A Hybrid Feature Identification Technique

http://www.ets.org/research/dload/acl99rev.pdf, 1998, 1 - 5

Burstein, J; Chodorow, M

Automated Essay Scoring for Nonnative English Speakers

http://www.ets.org/research/dload/acl99rev.pdf, 1999, 1 - 8

Burstein, J; Kaplan, R; Wolff, S; Lu, C

Using Lexical Semantic Techniques to classify free-responses

http://www.ets.org/research/dload/siglex.pdf, 1996, 1 - 12

Burstein, J; Kukich, K; Wolff, S; Lu, C; Chodorow, M

Computer Analysis of Essays

http://www.ets.org/research/dload/ncmefinal.pdf, 1998, 1 - 13

Burstein, J; Kukich, K; Wolff, S; Lu, C; Chodorow, M

Enriching Automated Essay Scoring Using Discourse marking

http://www.ets.org/research/dload/dscrfinal.pdf, 2001, 1 - 8

Burstein, J; Leacock, C; Swartz, R, 0-9539572-

Automated Evaluation of Essays and Short Answers

5th International Computer Assisted Assessment Conference, 2001a, 2 - 3 July, 42-54

Burstein, J; Marcu, D

Toward using Text Summarization for Essay-Based Feedback

TALN 2000 / http://www.ets.org/research/dload/talnsum.pdf, 2000, 16-18 Oct, 1 - 9

Burstein, J; Marcu, D

Benefits of Modularity in an Automated Essay Scoring System

http://www.ets.org/research/dload/colinga4.pdf, 2000, 1 - 6

Burstein, J; Marcu, D; Andreyev, S; Chodorow, M

Towards Automatic Classification of Discourse Elements in Essays

http://www.ets.org/research/dload/Burstein.pdf, 2001, 1 - 8

Burstein; Kukich; Wolff; Lu; Chodorow

Automated Scoring Using Discourse Marking

ETS, 2001b, 1 - 6

Byrne, C J

Tutor marked assignments at the Open University: a question of reliability

Assessment in Higher Education, 1980, 5(2), 150-167

Byrne, C J

Computerized Question Banking Systems: I - The State of the Art

British Journal of Educational Technology, 1976, 7(2) May, 44-64

Carter, RS

How invalid are marks assigned by teachers?

Journal Educational Psychology, 1952, 43, 218 - 228

Cartwright, G F; Derevensky, J L

The Development of Computer-Assisted Testing as an Adjunct to Traditional Instructional Processes in

British Journal of Educational Technology, 1978, 3(9) Oct, 166-169

Cast, B M D

The Efficiency of Different Methods of Marking English Composition: Part 1

British Journal of Education, 1939, 9, 257-269

Cast, B M D

The Efficiency of Different Methods of Marking English Composition: Part II

British Journal of Education, 1940, 10, 49-60

Catterall, J

What is a University Essay?

http://www.macarthur.uws.edu.au/ssd/ldc/Essay.html, 1994, 4

Chapman, R L, 0-00-47071

Roget's International Thesaurus

HarperCollins, 1996

Chase, C I

Essay test scoring: interaction of relevant variables

Journal of Educational Measurement, 1986, 23 (1), 33-41

Christie, J R

Computer-assisted Assessment of Essays

2nd Computer Assisted Assessment Conference, 1998, 17-18 June, 85-89

Christie, J R, 0-9533210-

Automated Essay Marking - for both Style and Content

3rd Computer Assisted Assessment Conference, 1999, 16-17 June, 39 - 48

Clark, N J

'Why wasn't I taught this stuff years ago?' Using a computer to assess and improve study skills

Journal of Computer Assisted Learning, 1990, 6, 174-189

CLASS Centre for Learning and Assessment, 1-901085-2

Chapter 6: Essays & Reports

Student's Guide to Effective Study, 2000, 17 - 21

Cockburn, B; Ross, A, 0-9016990-

Essays

Teaching Higher Education Series: 8 University of Lancaster, 1978

Coffman, W E

On the validity of essay tests of achievement

Journal of Educational Measurement, 1966, 3(2)summer, 151-156

Coffman, W E; Kurfman, D

A comparison of Two Methods of Reading Essay Examinations

American Educational Research Journal, 1968, 5(1) Jan, 99-107

Cohen, J

A coefficient of agreement for nominal scales

Educational and Psychological Measurement, 1960, 20 (1), 37 - 46

Cohen, J

Weighted KAPPA: Nominal scale agreement with provision for scaled disagreement or partial credit

Psychological Bulletin, 1968, (70) 4, 213-220

Cole N S

Why computerize assessment?

http://www.ets.org/vol1_1.html, 1997, 5

Cole, S A C

Using required departmental grading profiles

Annual meeting National Council of Teachers of English, 1987, November, 1-9

Collins, A M; Loftus, E F

A spreading-activation thoery of semantic processing

Psychological review, 1975, 82(6), 407-428

Conger, J L, 187873915

Windows API Bible

1992

Cooke, D; Craven, A H; Clarke, G M, 0-7131-344

Basic Statistical Computing

Edward Arnold, 1982

Crain, C

The Bard's Fingerprints, article by Caleb Crain

Lingua Franca, 1998, JulyAugust, 26-39

Daigon, A

Computer Grading of English Composition

English Journal, 1966, 55, 46-52

Dawson, J L

Suffix removal and word conflation

ALLC Bulletin, 1974, Michaelmas, 33-46

Deadman, G

An analysis of Pupils' reflective writing within a hyper-media framework

Journal of Computer Assisted Learning, 1997, 13, 16-25

DeLoughry, T J

Duke Professor pushes concept for grading essays by computer

The Chronicle of Higher Education, 1995, October 20, A24 - A25

Dempster, J A

Question Mark Designer for Windows

Active Learning, 1994, 1 Dec, 47-50

DeRemer, ML

Writing Assessment: Raters' Elaboration of the Rating Task

Assessing Writing, 1998, 5(1), 7 - 29

Dewhurst, F; Long, B A

Computer Aided Assessment of Animal Handling Skills in Applied Biology

Computer Education, 1988, Feb, 31-32

Diederich, P B

Cooperative Preparation and Rating of Essay Tests

English Journal, 1967, 56, 573-584 nc

Doell, D F

Gunning Fog Index

http://www.pima.edu/¬ddoell/tw/ghiex.html, 2000, 1-3

Dollard, J; Mowrer, O H

A method of measuring tension in written documents

Journal abnormal and social psychology, 1947, 42, 3-32

Donley, M

Marking Advanced Essays

English Language Teaching Journal, 1978, 32(2), 115-118

Dufrene, D D; Nelson, B H

A simplified system for grading WP assignments

Business Education Forum, 1988, April, 26-28

Duncan, C

A pragmatic approach to courseware developments in WIMP environments

British Journal of Educational Technology, 1990, 21(1), 31-40

Dutch, R A
Roget's Thesaurus
Penguin Reference Books, 1966

Edouard, L J; Harris, F T C
A Computer Card for Marking by the Examiner
British Journal of Educational Technology, 1976, 7(3), 37-40

Edwards, A L, 0-7167108
Multiple Regression and the Analysis of Variance and Covariance
W H Freeman and Company, 1979

Edwards, R P A; Gibbon, V, 0-222-0122
Words your children use
Burke Books, 1973

Eller, B F; Kaufman, A S; McLean, J E
Computer-based assessment of cognitive abilities: current status / future directions
J. Educational Technology Systems, 1986, 15(2), 137-147

Ellington, H
Returning to University from the World of Work - a Study Guide for Mature Postgraduate Students
CLASS, 2000, 1 -25

Ellington, H; Earl, S
Assessing Lower-Cognitive Skills
CLASS, 2001, May, 1-26

Ellington, H; Earl, S
Assessing Higher-Cognitive Skills
CLASS, 2001, May, 1-21

Ellington, H; Earl, S
Assessing Higher-Cognitive Skills
CLASS, 1998, 1-21

Ellington, H; Earl, S
Assessing Lower-Cognitive Skills
CLASS, 1998, 1-26

Epstein, J; Klinkenberg, WD
From Eliza to Internet: a brief history of computerised assessment
Computers in Human Behaviour, 2001, 17, 295 - 314

Erwin, PG; Calev, A

The Influence of Christian Name Stereotypes on the Marking of Children's Essays

Journal of Educational Psychology, 1984, 54, 223-227

ETS

Reinventing Assessment: Speculations on the future of large-scale educational testing

ETS Research, 1998, 1

Eysenck, H J

The validity of Judgements as a Function of the Number of Judges

Journal of Experimental Psychiatry, 1939, 25, 650-654

Fajardo, DM

Author Race, Essay Quality, and Reverse Discrimination

Journal of Applied Social Psychology, 1985, 15(3), 255-268

Fellbaum, C

English Verbs as a semantic Net.

WordNet, 0, 40-61

Fellbaum, C; Gross, D; Miller, K

Adjectives in WordNet

WordNet, 1993, August, 26-39

Finlayson, D S

The reliability of the marking of essays

British Journal of Educational Psychology, 1951, 21, 126-134

Fleiss, J L

Measuring nominal scale agreement among many raters

Psychological Bulletin, 1971, (76) 5, 378-382

Follman, J; Wong, M; Miller, W

Piles, number of grade categories, and theme grading

Child Study Journal, 1975, 5(1), 37-44

Foltz, P

Demonstration of automatic essay grading and feedback with LSA

Peter Foltz, 1998, 1 - 2

Foltz, P W

Latent semantic analysis for text-based research

Behaviour Research Methods, Instruments & Computers, 1996, 28(2), 197-202

Foltz, P W; Laham, D; Landauer, T K
The Intelligent Essay Assessor: Applications to Educational Technology
http://imej.wfu.edu/articles/1999/2/04/index.asp, 2001, 1 - 9

Foltz, P; Laham, D; Landauer, T K
Automated essay scoring: Applications to educational technology
Peter Foltz, 1999, 1 - 6

Forsyth, R
Bristol Stylemetry Research Unit
Computer Studies and Mathematics, Uni West England, 1998

Forsythe, G E; Wirth, N
Automatic Grading Programs
Communications of the ACM, 1965, 8(5), 275-278

Foster, D
The man who can read your writing
The Guardian, 1998, 13-14

Foubister, S P; Michaelson, G J; Tomes, N
Automatic assessment of elementary standard ML programs using Ceilidh
Journal of Computer Assisted Learning, 1997, 13, 99 - 108

Freeborn, D, 0-333-4056
A course book in English grammar
MacMillan, 1987

Gajar, A H
A Computer Analysis of Written Language Variables and a Comparison of Compositions Written by University
Journal of Learning Difficulties, 1989, 22(2) Feb, 125-130

Gannon, P, 0-7131-643
Assessing Writing: Principles of marking written English
Edward Arnold, 1985

Gardner, J; Morrison, H; Jarman, R
The impact of high access to computers on learning
Journal of Computer Assisted Learning, 1993, 9, 2-16

Gathy, P; Denef, J-F; Haumont, S
Computer-assisted self-assessment (CASA) in histology
Computers Educ., 1991, 17(2), 109-116

GRE

Research Reports from the GRE Board 1999-2000

GRE, 2000, 1-15

Green, R P

Towards Solving the Essay Dilemma

High School Journal, 1979, 62(7), 293-297

Gunning, R

The technique of Clear Writing: Appendix B (revised edition)

McCraw-Hill Book Company, 1968

Hales, LW; Tokar, E

The effect of the quality of preceding responses on the grades assigned to subsequent responses to an essay

Journal of Educational Measurement, 1975, 12(2), 115 - 117

Hall, C G W; Daglish, N D

Length and Quality: an exploratory study of inter-marker reliability

Assessment and Evaluation in Higher Education, 1982, 7(2), 186-191

Hamel, T

Grading by Computers: The Programs Don't Get an A, but They're Worth Considering

The Chronicle of Higher Education, 1988, October 12, B2

Hararp, H; McDavid, JW

Name Stereotypes and Teachers' Expectations

Journal of Educational Psychology, 1973, 65(2), 222-225

Harden, R M; Smyth, J J

Computer-based study guides II: educational components and advantages

Medical Teacher, 1994, 16(4), 315-321

Harway, N I; Iker, H P

Computer Analysis of Content in Psychotherapy

Psychological Reports, 1964, 14, 720-722

Haswell, RH

Grading Student Writing

Assessing Writing, 1999, 6(1), 133 - 138

Hawkridge, D

Problems in Implementing Computer-managed Learning

British Journal of Educational Technology, 1974, 5(1) Jan, 31-43

Hearst, M A
The debate on automated essay grading.
IEEE Intelligent Systems, 2000, sept/oct, 22-37

Hiller, J H; Marcotte, D R; Martin, T
Opinionation, Vagueness, and Specificity-Distinctions: Essay traits measured by computer
American Educational Research Journal, 1969, 6 (2), 271-286

Hofland, K; Johansson, S, 82-7283-02
Lancaster-Oslo-Bergen Corpus
Word Frequencies in British and American English, 1982

Hofland, K; Johansson, S
Word frequencies in British and American English
1982

Hollingsworth, J
Automatic Graders for programming classes
Communications of the ACM, 1960, 3, 528-529

Holmes, D
The evolution of stylometry in Humanities scholarship
Literary and Linguistic Computing, 1998, 13, 111 - 117

Holmes, R
Marks my words: Student essays can now be graded by machine
New Scientist, 1998, 12

Hughes, D C; Keeling, B
The use of Model Essays to Reduce Context Effects in Essay Scoring
Journal of Educational Measurement, 1984, 21(3), 277-281

Humanities Computing Unit
Humanities Computing Unit
University of Oxford, 0, 1 - 2

Humphreys, G, 067910667
Teach yourself English grammar
Hodder and Stoughton Limited, 1979

Hutchinson, B
Who's testing what for whom ?
British Journal of Educational Technology, 1994, 25(3), 220-221

ICE-GB

Download the ICE-GB sample Corpus

ICE-GB, 1999, 1 - 4

Iker, H P; Harway, N I

A computer approach towards the analysis of content

Computers in Behavioral Science, 1965, 10, 173-183

Iker, H P; Harway, N I

A Computer Systems Approach towards the Recognition and Analysis of Content

Computer Studies in the Humanities and Verbal Behavior, 1968, 1, 134-154

Issacson, P C; Scott, T A

Automating the Execution of Student Programs

SIGCSE Bulletin, 1989, 21(1) Jun, 15-22

Jackson, D

Using software tools to automate the assessment of students programs

Computers Education, 1991, 17(2), 133-143

Jackson, M

Making the grade: The formative evaluation of essays

http://ultibase.edu.au/develop/Articles/jacks1.html, 1994, 2

Jacoby, H

Note on the marking system in the astronomical course at Columbia College, 1909 - 10

Science, 1910, May 27, 819

Johnson, S; Maher, B

Monitoring Science Performance Using a Computerized Question Banking System

British Journal of Educational Technology, 1982, 13(2) May, 97-106

Johnson, V E

On Bayesian Analysis of Multirater Ordinal Data: An Application to Automatic Essay Grading

Journal of the American Statistical Association, 1996, March, 43-51

Kaplan RM & BA, Wolff S E, Burstein JC, Lu C, Rock DA

Scoring Essays Automatically Using Surface Features

GRE Research, 1998, August, 1-13

Kaplan, R M

Using a Trainable Patern-Directed Computer Program to Score Natural Lanaguage Item Responses

GRE Research Report, 1992, April, 1-43

Kelly, E F; Stone, P J, 072046180
Computer Recognition of English Word Senses
North-Holland Publishing Company, 1973

Kennedy, G, 058223153
An introduction to corpus linguistics
Longman, 1988, various
Kintsch, W; Yarbrough, J C
Role of Rhetorical Structure in Text Comprehension
Journal of Educational Measurement, 1982, 74 (6), 828-834

Kirkpatrick, B, 0-19-86012
The Concise Oxford Thesaurus, in clear A-Z from
Oxford University Press, 1997

Kleeman, J
Now is the time to computerize pen and paper tests.
Question Mark Computing Ltd., 1998, Jan, 1-6

Klein, S; Simmons, RF
A Computational Approach to grammatical Coding of English Words
Journal of the Association for Computing Machines, 1963, 10(3), 334-347

Knight, K; Marcu, D
Statistics-Based Summarization - Step One: Sentence Compression
ETS, 0, 1 - 8

Kniveton, B H
A correlational analysis of multiple-choice and essay assessment measures
Research in Education, 1996, 56 Nov, 73-84

Lake, A; Cook, C
STYLE An Automated Program Style Analyser for Pascal
SIGCSE Bulletin, 1990, 22(3) Sep, 29-33

Landis, J R; Koch, G G
The Measurement of Observer Agreement for Categorical Data
Biometrics, 1977, 33, 159 - 174

Landy, D; Sigall, H
Beauty is Talent: Task evaluation as a function of the performer's physical attractivness
Journal of Personality and Social Psychology, 1974, 29(3), 299 - 304

Langendoen, D T, 003910116
The study of syntax
Holt, Rinehart & Winston, 1971

Larkey, L S, 1-58113-01
Automatic Essay Grading using Text Categorisation Techniques
SIGIR '98, Melbourne, Australia, 1998, August24-8, 90-95

Leiblum, M D; Coenen, A M L; van Luijtelarr, E L J M
A computer-aided self-testing system for biological psychology
Journal of Computer Assisted Learning, 1994, 10, 229-239

Leigh, D J; Halliday, E
Measuring commercial program complexity
Software Quality Management, 1994, July, 375-383

Levy, L B; Friitz, K V
Status report on the computer grading of essays
Counseling Center Reports, Uni Wisconsin, 1972, 5(10), 1-14

Lloyd, D; Martin, J G; McCaffery, K
The Introduction of Computer-based Testing on an Engineering Technology Course
Assessment & Evaluation in Higher Education, 1996, 21(1), 83-90

Lockwood, F
Different Formats for Summative Assessment Material - variations on a theme
British Journal of Educational Technology, 1981, 12(3) Oct, 235-242

Long, R B, 022649260
The sentence and its parts
The University of Chicago Press, 1980

MacLeod, A; Meyers, A; Grishman, R
COMLEX Syntax
http://cs.nyu.edu/cs/projects/proteus/comlex, 2001, 1-3

Malehorn, H
Ten measures better than grading
The Clearing House, 1994, Jul / Aug, 323-324

Markham, LR
Influences of Handwriting Quality on Teacher Evaluation of Written Work
American Educational Research Journal, 1976, 13(4), 277-283

Marshall, JC

Composition Errors and Essay Examination Grades Re-examined

American Educational Research Journal, 1967, 4(4), 375-385

Marshall, S

An intelligent marking assistant: an application of artifical intelligence in teaching

Higher Education Research and Development, 1986, 5 (2), 201-211

Mason, O; Grove-Stephenson, I, 0-9539572-

Automated free text marking with Paperless School

6th International Computer Assisted Assessment, 2002, July, 213 - 219

Maunder, P

Marking Time? - A Review of the Assessment of Essays in Economics at A-Level

Economics, 1991, 27(3) 115, 119-123

McCollum, K

How a computer program learns to grade essays

http://vci.cso.uiuc.edu/courses/.../review904585320.html, 1998, 1 - 4

McDonald, R; Sansom, D

Use of assignment attachments in Assessments

Assessment in Higher Education, 1979, 5 (1), 45-55

Micceri, T; Pritchard, W H; Barrett, A J

Must computer courseware evaluation be totally subjective ?

British Journal of Educational Technology, 1989, 20(2), 120-128

Michell, T; Russel, T; Broomhead, P; Aldridge, N, 0-9539572-

Towards Robust Computerised Marking of Free-Text Responses

6th International Computer Assisted Assessment, 2002, July, 233 - 249

Miller, G

Nouns in WordNet: A lexical inheritance system.

WordNet, 1993, 10-25

Miller, G A; Beckwith, R; Fellbaum,C; Gross,D; Miller,K

Introductionto WordNet: an on-lone lexical database

WordNet, 1993, august, 1-9

Miltsakaki, E
Locating Topics in Text Processing
ETS, 0, 1 - 12

Miltsakaki, E; Kukich, K
Automated Evaluation of Coherence in Student Essays
ETS, 0, 1 - 8

MoodWatch
Software to detect mood of e-mail user
Aberdeen Press & Journal, 2000, 18-9-2000, 16

Moore, R; Marshall, D
Automated Coursework Assessment over the Internet
Dept Computer Studies Cardiff University, 1999, 1 - 9

Mosenthal, P; Tamor, L; Walmsley, S A, 0-582-2830
Research on Writing
Longman, 1983

Murray, B
The latest techno tool: essay-grading computers
http://www.apa.org/monitor/aug98/grade.html, 2000, 1 - 3

Myers, A E; McConville, C B; Coffman, W E
Simplex structure in the grading of essay tests
Educational and Psychological Measurement, 1966, 26(1), 41-54

NCET, 1-85379-34
Using It for assessment - key issues
NCET, 1995, 8-9

Neill, N T
Computer-based Testing with Question Mark Professional
Computer Education, 1993, June, 23-26

Newbould, C A; Massey, A J
A computerized Item Banking System (CIBS)
British Journal of Educational Technology, 1977, 2 May, 114-123

Newstead, S; Dennis, I
The reliability of exam marking in psychology
The Psychologist, 1994, 7(5), 216-9

Nolan, J, R

An expert fuzzy classification system for supporting the grading of student writing samples

Expert Systems with Applications, 1998, 15, 59-68

Norton, L S

Essay-writing: what really counts?

Higher Education, 1990, 20(4), 411-442

Nyberg, V R

Reliability of the Alberta Essay Scales

Alberta Journal of Educational Research, 1980, 26(1), 64-67

Oxford University Press, 5-0313660

The Oxford Thesaurus, in A-Z form

Focus Multimedia

Page, E B

The Imminence of Grading Essays by Computer

Phi Delta Kappen, 1966, 47, 238-243

Page, E B

Grading Essays by Computer: Progress Report

Invitional conference on testing problems New York City, 1966, 87-100

Page, E B

The use of the computer in analyzing student essays

International Review of Education, 1968, 14, 210-224

Page, E B

Computer Grading of Student Prose, Using Modern Concepts and Software

Journal of Experimental Education, 1994, 62(2), 127-142

Page, E B

Grading Essays by Computer: Why the Controversy?

NCME Symposium New York, 1996, April 11

Page, E B

The Classroom Challenge and Write America !

AERA/NCME Symposium Grading Essays by Computer Chicago, 1997, March 27

Page, E B

The Second Blind Test with ETS: PEG predicts the Graduate Record Exams

AERA/NCME Symposium Grading Essays by Computer Chicago, 1997, March 27

Page, E B; Fisher, G A; Fisher, M A

Project Essay Grade: A FORTRAN program for Statistical Analysis of Prose

British Journal of Mathematical & Statistical Psychology, 1968, 21, 139

Page, E B; Keith, T Z; Lavoie, M J

Computer Grading of Essays Traits in Student Writing

Annual meeting National Council Measurements in Education NY, 1996, April 11, 8 off

Page, E B; Lavoie, M J; Keith, T Z

Construct Validity in the Computer Grading of Essays

Annual meeting American Psychological New York City, 1995, August 13, 3 off

Page, E B; Petersen, N S

The Computer Moves into Essay Grading updating the Ancient Test

Phi Delta Kappan, 1995, 76(7), 561-566

Parker, A; Hamblen, J O

Computer Algorithms for Plagiarism Detection

IEEE Transactions on Education, 1989, 32(2) May, 94-99

Partington, J, 1-85889-12

Introduction to Computer-aided Assessment in Higher Education

TLTP Project ALTER / UCoSDA, 1995, software

Paulus, D H; McManus, J; Page, E B

Some applications of natural language computing to computer-assisted instruction

Contemporary Education, 1969, 40, 280-285

Paxton Software

Grademaster

http://www.microserv.com/users/viking/Paxton-Software.html, 2000, 1 - 3

Peacock, M

Handwriting versus wordprocessed print: an investigation into teachers' grading of English Language

Journal of Computer Assisted Learning, 1988, 4, 162-172

Penny, J; Johnson, RL; Gordon, B

The effect of rating augmentation on inter-rater reliability: An empirical study of a holistic rubric

Assessing Writing, 2000, 7, 143 - 164

Peterson, N D

Concordance Analyser of COBOL and Fortran Source Programs

Software Age, 1970, Jan, 13-14

Pitt, M
The use of electronic mail in undergraduate teaching
British Journal of Educational Technology, 1996, 27(1), 45-50

Plagiarism Organisation
Plagiarism Organisation, 1999, 2

Plain English Campaign
Plain English Campaign
Plain English Campaign, 1999, 1 - 3

Plain English Campaign
A-Z guide of alternative words
Plain English Campaign, 1999, 1 - 16

Powers DE, Burstein JC, Chodorow M, Fowles ME, Kukich K
Comparing the Validity of Automated and Human Essay Scoring
GRE Research, 2000, June, 1-23

Powers, DE;Burstein, JC;Chodorow, M;Fowles, ME;Kukich,K
Stumping e-rater: challenging the validity of automated essay scoring
Computers in Human Behaviour, 2002, 18, 103 - 134

Pritchett, N; Zakrzewski, S
Interactive Computer Assessment of Large Groups: Student Responses
Innovations in Education and Training International, 1996, August, 242-247

Pritchett, N; Zakrzewski, S
Computerised Formal Assessment: A Pilot Study of Interactive Computer Examination of Undergraduates at the
British Journal of Educational Techology, 1995, 26(2), 152-153

Quirk, R; Greenbaum, S;Leech, G; Svartvik, J, 0-582-5244
A grammar of contemporary English
Longman, 1972

Race, P
Ten worries about assessment
British Journal of Educational Technology, 1992, 23(2) May, 141

Reek, K A
The TRY System - or - How to Avoid Testing Student Programs
SIGCSE Bulletin, 1989, 21(1), 112-116

Rees, M J

Automatic Assessment Aids for Pascal Programs

SIGPLAN, 1982, 17(10) Oct, 33-42

Reid, S; Findlay, G

Writer's workbench analysis of holistically scored essays

Computers and Composition http://corax.utexas.edu/cac/, 2002, 1 - 11

Robinson, M

Using Email and the Internet in Science Teaching

Journal of Information Technology for Teacher Education, 1994, 3(2), 229-238

Roebuck, M

Computer-Assisted Assessment

Programmed Learning and Educational Technology, 1972, 9, 283 - 291

Roget, P M; Roget, J L; Roget S R, 0-946674-9

Roget's Thesaurus

Roydon Publishing Co. Ltd., London, 1972

Romppel, M

Resources related to content analysis and text analysis

http://gwdu19.gwdg.de/~mromppe/contold.htm, 1998, 1 - 7

Rowley, J

All students in higher education should submit their assignments in word processed form. Should they?

British Journal of Educational Technology, 1994, 25(3), 225-227

Rubin, D

New program reads, critiques student essays

http://www.cavalierdaily.com/.Archives/1998/September/..., 1998, 1 - 3

Russell, M; Haney, W

Testing Writing on computers: An experiment comparing student performance on tests conducted via computer

http://epaa.asu.edu/epaa/v5n3.hmtl, 1997, 5 (3), 1 - 19

Rust, W B, 027331639

Objective Testing in Education and Training

Pitman, 1973, book, 27-34

Sager, N, 020106769
Natural language Information Processing
Addison-Wesley Publishing Company, Inc., 1981

Sansom, I
Wordlist 87n
The Guardian, 1998, 8th August, 8

Schwartz, M
Word convertress
http://user.cs.tu-berlin.de/~schwartz/pmh/elser/word6/format.html, 1997, 1 - 29

Sciarone, A G
A fully automatic homework checking system
IRAL, 1995, 33(1) Feb, 35-46

Script and Pattern Recognition Research Group
Script and Pattern Recognition Research Group
http://www.doc.nu.ac.uk/HAND/index.html, 1997, 1 - 9

Shavelson; Baxter; Pine; Yure; Goldman; Smith
Alternative Technologies for Large Scale Science Assessment: Instrument of Education Reform
School Effectiveness and School Inprovement, 1991, 2(2), 97-114

Sheppard, K
Two feedback types: Do they make a difference ?
Journal of Language Teaching and Research in Southeast Asia, 1992, 23(1), 103-110

Siemens, R
Practical Content Analysis Techniques for Text-Retrieval in Large, Un-tagged Text-bases
ACM Conference, 1993, 11th, 293-300

Simmons, R F
Answering English Questions by Computer: A Survey
Communications of the ACM, 1965, 8(1) Jan, 53-70

Slotnick, H B
Toward a theory of computer essay grading
Journal of Educational Measurement, 1972, 9(4), 253-263

Smith, O R
GENDEX: GENeral InDEXer of Words with Context. A Concordance Generator.
Computer Studies in the Humanities and Verbal Behaviour, 197072, 3, 50-53

Smith, R E

Examination by Computer

Computers in Behavourial Science, 1963, 8, 76-79

Smith, WL

Introduction to Special Issue

Assessing Writing, 1998, 5(1), 3 - 5

Starch, D; Elliott, EC

Reliability of the grading of high-school work in English

School Review, 1912, 20, 442 - 457

Starch, D; Elliott, EC

Reliability of grading work in Mathematics

School Review, 1913a, a, 254 - 259

Starch, D; Elliott, EC

Reliability of grading work in History

School Review, 1913b, 21, 676 - 681

Stephens, D

Using computer assisted assessment: time saver or sophisticated distraction ?

Active Learning, 1994, 1 Dec, 11-15

Story, R E

An explanation of the effectiveness of latent semantic indexing by means of a Bayesian Regression model

Information Processing & Management, 1996, 32(3), 329-344

Sweedler-Brown, C O

The effect of training on the appearance bias of Holistic Essay Graders

Journal of Research and Development in Education, 1992, 26(1), 24-29

Sydes, M; Hartley, J

A thorn in the Flesch: observations on the unreliability of computer-based readability formulae

British Journal of Educational Technology, 1997, 28(2), 143-145

The Oxford Text Archive

The Oxford Text Archive

University of Oxford http://ota.ahols.ac.uk

Thompson, C

New Word Order: The attack of the incredible grading machine

http://www.linguafranca.com/9907/nwo.html, 2001, 1 - 14

Thomson, N D
Literary Statistics = Part 1 to 6
ALLC Bulletin, 1973, Michaelmas, 1 - 36

Thorndike
Educational Measurement - Chapter 8, 10, & p 271
Educational Measurement

Thorndike, E L; Lorge, I
The teacher's word book of 30,000 words
1952

Thorne, M P
The Specification and Design of Educational Microcomputer Systems
British Journal of Educational Technology, 1980, 11(3), 178-184

TOEFL
Computer-Based TOEFL Test
TOEFL, 1999, 1 - 2

Tollefson, N; Tracy, D B
Test Length and Quality in the Grading of Essay Responses
Education, 1980, 101(1), 63-67

Townsend, M A R; Kek, L Y; Tuck, B F
The effect of mood on the reliability of essay assessment
British Journal Educational Psychology, 1989, 59, 232-240

Townsend; Hicks; Thompson; Wilton; Tuck; Moore
Effects of Introductions and Conclusions in Assessment of Students Essays
Journal of Educational Psychology, 1993, 85 (4), 670-678

Trathen, C; Sajeev, A S M
A Protocol for computer mediated education across the internet
British Journal of Educational Technology, 1996, 27(3), 204-213

Tucker, A B, 012702550
Text Processing Algorithms, Languages and Applications
Academic Pess, 1979

Underwood, T; Murphy, S
Interrater Reliability in a Califonia Middle School English / Language Arts Protfolio Assessment Program
Assessing Writing, 1998, 5(2), 201 - 230

UWE-BSRU

Bristol Stylometry Research Unit

http://www.csm.uwe.ac.uk/csm/stylometry/index.html, 1998, 1

Viadero, D

Making the Grade

Education Week, 1995, 31st May, 33-35

Ward, G

Various MOBY wordlists

http://www.dcs.shef.ac.uk/research/ilash/Moby/, 2000, 2 November, various

Watkins, J J; Calverley, G J; Bacon, R A

Implementation of a Complete Learning Environment

Innovations in Education and Training International, 1995, August, 239-244

Weinstein, B

Software designed to grade essays

Boston Globe Online, 1998, 21 June, 1 - 4

Weisberg, S, 0-471-8795

Applied Linear Regression 2ed

Wiley, 0

Wendel, F C; Anderson, K E

Grading and Marking Systems: What are the Practices, Standards?

NASSP Bulletin, 1994, 78 Jan, 79-84

Whale, G

Paperless Assignment Marking

ACSC-14, 1991, Kensington, 32:1-8

Whalen, T E

The analysis of essays by computer: A simulation of teacher ratings

Annual meeting American Educational Research Association, 1971, February, 1-26

Whalen, T E

A validation of the Smith test for measuring teacher judgement of written composition

Education, 1971, 93, 173 - 175

Williams, C B

A note on the statistical analysis of sentence-length as a criterion of literary style

Biometrika, 1940, 31, 356-61

Winer, B J

Statistical principles in experimental design

McGraw-Hill, 0

Wiseman, S

The marking of English composition in Grammar School Selection

British Journal Educational Psychology, 1949, 19, 200-209

Woolson, R F, 0-471-8061

Statistical Methods for the Analysis of Biomedical data, 257-260

Worrell, S R

Grademaster

http://www.microserv.com/users/viking/Paxton-Software.html, 2002

Wresch, W

The Imminence of Grading Essays by Computer - 25 Years Later

Computers and Composition, 1993, 10 Apr, 45-58

Yeh, SS

Validation of a Scheme for Assessing Argumentative Writing of Middle School Students

Assessing Writing, 1998, 5(1), 123 - 150

Yule, G U

On sentence-length as a statistical characteristic of style in prose

Biometrika, 1938, 30, 363-390

Appendix A: Table of Metrics for Style Marking
Metrics with Frequencies > 3

	Author	Daigon	Page	Bishop	Whalen	Slotnick	Christie	Gajar	Page	Johnson	Page	Frequency
	Year	1966	1966,68	1970	1971	1972	1988	1989	1994	1996	1996	AVERAGE
Metric	**No. of Metrics**	13	29	13	27	23	7	15	34 not clear	6	50 ex 290+	22
1	average sentence length	y	y	y	y	y	y	y		y		8
2	average word length	y	y		y	y	y			y		6
3	number of words		y	y		y	y	y		y		6
4	number of commas		y		y	y				y		4
5	number of prepositions	y	y		y					y		4
6	average paragraph length			y		y		y				3
7	number of colons		y		y	y						3
8	number of paragraphs		y		y	y						3
9	number of parentheses		y		y	y						3
10	number of questions		y		y			y				3
11	number of quotation marks		y		y	y						3
12	number of relative pronouns		y		y	y						3
13	number of semi-colons		y		y	y						3
14	SD sentence length		y		y	y						3
15	SD word length		y		y	y						3
16	spelling errors		y		y					y		3
17	word list(s)	y			y	y						3

Metric with Frequencies = 2

	Author	Daigon	Page	Bishop	Whalen	Slotnick	Christie	Gajar	Page	Johnson	Page	Frequency
	Year	1966	1966,68	1970	1971	1972	1988	1989	1994	1996	1996	AVERAGE
Metric	**No. of Metrics**	13	29	13	27	23	7	15	34 not clear	6	50 ex 290+	22
18	Dale list		y		y							2
19	declarative sentence frequency		y			y						2
20	FLESCH Formula	y		y								2
21	number of connective words		y		y							2
22	number of different words	y						y				2
23	number of exclamations		y					y				2
24	number of hard words	y					y					2
25	number of sentences					y	y					2
26	sentence length			y				y				2
27	subordinating conjunctions		y		y							2
28	type token				y			y				2

Metrics with Frequencies = 1 ...

	Author	Daigon	Page	Bishop	Whalen	Slotnick	Christie	Gajar	Page	Johnson	Page	Frequency
	Year	1966	1966,68	1970	1971	1972	1988	1989	1994	1996	1996	Average
Metric	**No. of Metrics**	13	29	13	27	23	7	15	34 not clear	6	50 ex 290+	22
29	% polysyllabic words			y								1
30	adverbs					y						1
31	AND				y							1
32	average segment token type(s)							y				1
33	average words / T unit							y				1
34	capitialisation errors				y							1
35	different words					y						1
36	FANG Formula			y								1
37	FOG Formula						y					1
38	gerunds					y						1
39	GUNNING Formula			y								1
40	imperative sentence frequency					y						1
41	interrogative sentence frequency					y						1
42	LORGE Formula	y										1
43	number of apostrophes		y									1
44	number of capitals				y							1
45	number of characters						y					1
46	number of dashes		y									1
47	number of hyphens		y									1
48	number of periods		y									1
49	number of slashes		y									1
50	number of syllables/100 words	Y										1
51	number of underlined words		y									1

... Metrics with Frequencies = 1

	Author	Daigon	Page	Bishop	Whalen	Slotnick	Christie	Gajar	Page	Johnson	Page	Frequency
	Year	1966	1966,68	1970	1971	1972	1988	1989	1994	1996	1996	Average
Metric	**No. of Metrics**	13	29	13	27	23	7	15	34 not clear	6	50 ex 290+	22
52	overuse adjectives				y							1
53	overuse articles			y								1
54	overuse passive verbs			y								1
55	past participles					y						1
56	prepositional phrases	y										1
57	punctuation			y								1
58	SD paragraph length					y						1
59	sentence structure			y								1
60	sentences ending punctuation		y									1
61	sentences: simple/75 sentences	y										1
62	SO				y							1
63	statement /composition							y				1
64	subject-verbs openings		y									1
65	syntactic maturity							y				1
66	T units							y				1
67	THEN				y							1
68	Thorndike list	y										1
69	title present		y									1
70	TO BE				y							1
71	usage errors				y							1
72	vocabulary/diversity							y				1
73	vocabulary/fluency							y				1
74	WHEN				y							1
75	WINNETKA Formula	y										1

Character counts [11]

~ commas
~ colons
~ semi-colons
~ parentheses: () {} []
~ quotation marks: " '
~ apostrophes
~ capitals
~ number of characters
~ dashes
~ hypens
~ slashes: / \

Syllable [0]

Word Counts [4]

~ total words
~ different words
~ hard words
~ specific words: AND, THEN, TO BE, SO, WHEN

Word List(s) [1 ... 5]

~ specified:
Dale,
LOB,
Thorndike,
Edward-Gibb,
Roget,
...

Word Types [12]

~ relative pronouns
~ prepositions
~ function / stop
~ connective words
~ adverbs
~ gerunds
~ nouns
~ verbs
~ adjectives
~ underlined [not tested in SEAR]
~ past participles
~ overuse: adjectives, articles, passive verbs

Phrase [4]

~ prepositional phrases
~ statement / composition
~ subject-verb openings
~ punctuation

Sentence [3]

~ number of sentences
~ sentence structure
~ ending punctation

Paragraph [4]

~ number of paragraphs
~ type token
~ T units
~ subordinating conjunctions

Title [1]

~ present

Errors [4]

~ spelling
~ grammar
~ capitalisation
~ usage errors

Summary [15]

~ average word length
~ average sentence length
~ average paragraph length
~ SD word length
~ SD sentence length
~ SD paragraph length
~ average words / T unit
~ average segment token type(s)
~ % polysyllabic words
~ syntactic maturity
~ vocabulary / diversity
~ vocabulary / fluency
~ sentence frequency: imperative, interrogative, declarative
~ simple sentences / 75 sentences
~ syllables / 100 words

Special [5]

~ FOG & Modified FOG // GUNNING
~ FANG
~ LORGE
~ WINNETKA
~ FLESCH

Frequencies [3 profiles]

~ word length: 1, 2, ... 25, 26/26+
~ sentence length: 1, 2, ... 100, 101/101+
~ paragraph length: 1, 2, ... 100, 101/101+
both by #sentences and #words

Summary of Metrics, by Type

Counts		
	Character Counts	11
	Syllable	0
	Word Counts	4
	Word Lists	1
	Word Types	12
	Phrases	4
	Sentences	3
	Paragraph	4
	Errors	4
	Title	1
Statistical		
	Summary	15
	Special	5
	Frequencies	3
	TOTAL	67

- Word length profile,
- Sentence length profile, in both linear and logarithmic scales,
- Frequencies of specific word types,
  - Frequencies of specific words, pairs of words, and so on.

- Hapax Legomena and Dis Legomena,

  that is words that appear only once or twice per essay,
- Distribution of the last five syllables of each sentence,
- Where each syllable is classed as either being "long" or "short",
- The probability of an author's new text containing new words,

  based on Fisher's statistical prediction of finding a new butterfly species,
- The use of certain "marker words",
- The frequency of short words, that is words of two and three letters long,
- The frequency of vowel words, that is words starting with a vowel,

#89 Norton: Types of sentences:
Introductory
Quotes – Sayings
Linking
Argument
Conclusions
Factual –descriptive infp
Quoted research
Described research
Metioned research
Textbook information

Carroll's token ratio

Herdan's K

Appendix C: Examples of Content Data Structure

Examples of data structure

Here the author cites some examples of the data structure to illustrate how the data structure holds the content schema.

1] A simple essay:

Content: Robert Gordon was born in 1668 in the Castlegate, Aberdeen. He retired in 1720, and died in 1731. His father name was Arthur Gordon, and his mother was called Isabella Menzies.

Abridged Data structure:

1	a	a	0	2	15	22	0	
2	a	a	1	3	6	9	12	
3	a	a	2	4	5	0	0	
4	f	born	3	0	0	0	0	359 v
5	f	1668	3	0	0	0	0	
6	a	a	2	7	8	0	0	
7	f	castlegate	6	0	0	0	0	
8	f	aberdeen	6	0	0	0	0	
9	a	a	2	10	11	0	0	
10	f	retired	9	0	0	0	0	283 v
11	f	1720	9	0	0	0	0	
12	a	a	2	13	14	0	0	
13	f	died	12	0	0	0	0	360 v
14	f	1731	12	0	0	0	0	
15	a	a	1	16	17	0	0	
16	f	father	15	0	0	0	0	166 n
17	s	s	15	18	21	0	0	
18	o	o	17	19	20	0	0	
19	f	arthur	18	0	0	0	0	
20	f	a	18	0	0	0	0	
21	f	gordon	17	0	0	0	0	
22	a	a	1	23	24	0	0	
23	f	mother	22	0	0	0	0	166 n
24	s	s	22	25	28	0	0	
25	o	o	24	26	27	0	0	
26	f	isabella	25	0	0	0	0	
27	f	i	25	0	0	0	0	
28	f	menzies	24	0	0	0	0	

2] Holding a quote:

Content:

Newton's First Law is that "An object will remain at rest or continue to travel at constant speed in a straight line unless acted on by an unbalanced force".

Abridged Data Structure:

1	s	s	0	2	23	0	0
2	s	s	1	3	8	13	18
3	s	s	2	4	5	6	7
4	f	an	3	0	0	0	0
5	f	object	3	0	0	0	0
6	f	will	3	0	0	0	0
7	f	remain	3	0	0	0	0
8	s	s	2	9	10	11	12
9	f	at	8	0	0	0	0
10	f	rest	8	0	0	0	0
11	f	or	8	0	0	0	0
12	f	continue	8	0	0	0	0
13	s	s	2	14	15	16	17
14	f	to	13	0	0	0	0
15	f	travel	13	0	0	0	0
16	f	at	13	0	0	0	0
17	f	constant	13	0	0	0	0
18	s	s	2	19	20	21	22
19	f	speed	18	0	0	0	0
20	f	in	18	0	0	0	0
21	f	a	18	0	0	0	0
22	f	straight	18	0	0	0	0
23	s	s	1	24	29	0	0
24	s	s	23	25	26	27	28
25	f	line	24	0	0	0	0
26	f	unless	24	0	0	0	0
27	f	acted	24	0	0	0	0
28	f	on	24	0	0	0	0
29	s	s	23	30	31	32	33
30	f	by	29	0	0	0	0
31	f	an	29	0	0	0	0
32	f	unbalanced	29	0	0	0	0

33	f	force	29	0	0	0	0

Appendix D: Example of Content Schema

Sample of a [short] Schema for Robert Gordon, founder of the Robert Gordon University.

Essay is marked out of 25.

```
Mark  Item
 0    Robert Gordon
 3            born 1668 Castlegate Aberdeen
 3            inherited 1680 £1,000
 5            graduated 1689 Marischal College Aberdeen
 3            retired 1720 Aberdeen
 1            died 1731
[max:15]

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

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

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

 1            Trader
 2                    Danzig Baltic
[max:3]

[max:25]
```

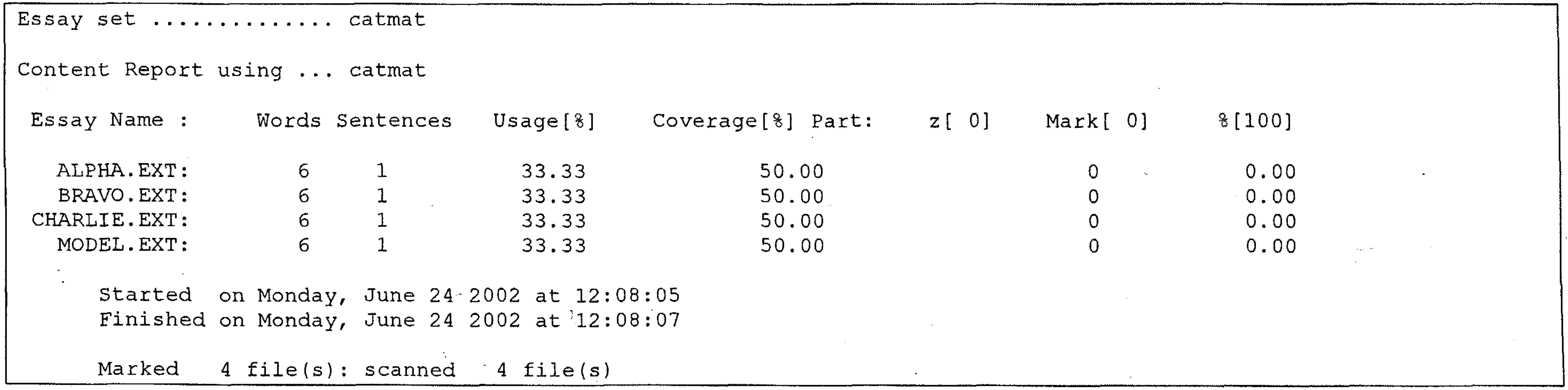
The above content schema is generated using the free software package called Notepad. Notepad is provided free with Microsoft Windows ™®.

The rows represent facts and the relationship between the facts.
Tabbing of the rows indicate the dependence between the rows.

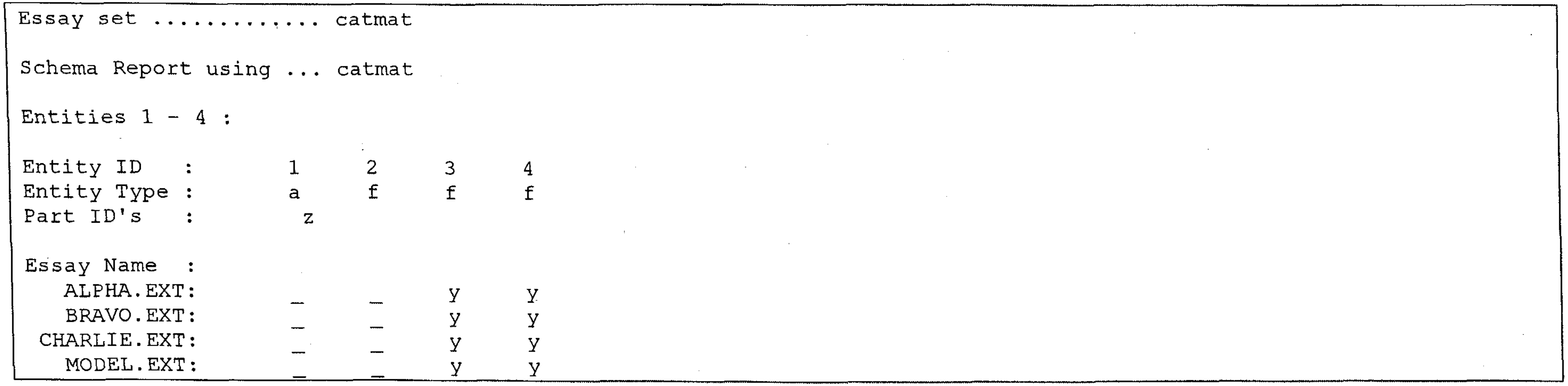
By settling on this particular layout, the automated generation of the data structure is possible.

Appendix E: Sample Content and Schema Reports

Below are the two reports, called Content Report and Schema Report, that have been generated from a very simple example of "The cat sat on the mat".



```
Essay set ............. catmat

Content Report using ... catmat

 Essay Name :     Words Sentences   Usage[%]    Coverage[%] Part:    z[ 0]    Mark[ 0]     %[100]

   ALPHA.EXT:        6      1         33.33            50.00                      0           0.00
   BRAVO.EXT:        6      1         33.33            50.00                      0           0.00
 CHARLIE.EXT:        6      1         33.33            50.00                      0           0.00
   MODEL.EXT:        6      1         33.33            50.00                      0           0.00

      Started  on Monday, June 24 2002 at 12:08:05
      Finished on Monday, June 24 2002 at 12:08:07

      Marked   4 file(s): scanned   4 file(s)
```



```
Essay set ............. catmat

Schema Report using ... catmat

Entities 1 - 4 :

Entity ID   :       1     2     3     4
Entity Type :       a     f     f     f
Part ID's   :        z

Essay Name  :
   ALPHA.EXT:       _     _     y     y
   BRAVO.EXT:       _     _     y     y
 CHARLIE.EXT:       _     _     y     y
   MODEL.EXT:       _     _     y     y
```

Appendix F: Example of "bad faith" essay

Dr R A Butler has created these two essays for the author to use in the development of the research project. In particular a "bad faith" essay which is shown below an acceptable essay.

The "acceptable" essay:

Robert Gordon was born in the 17th century in Aberdeen, he was the son of A Gordon, who worked in Edinburgh as a solicitor. His father lived in Straloch and made maps. He was also called Robert Gordon. His wife was Isabell Menzies. Her family were famous in the North East, many had been Provosts of Aberdeen.

When his father died in 1680, he left the 12 year old Robert a considerable amount of money, over £1000. Four years later he began to trade in the city because he had been elected a Burgess of Guild. He lived in the Castlegate and went to the from Marischal College. After graduating in 1689 he travelled extensively on the continent before establishing a business in Danzig in the Baltic as a merchant trader. At that time over 40,000 Scots were also working in Poland.

During the reign of Queen Anne Archives at Aberdeen University, dated 1692, show a generous donation towards the appeal fund for the rebuilding of part of Marischal College. Little else is known about his life abroad. It seems from documents from 1699 onwards that he was lending money on good terms as working capital for Aberdeenshire lairds who needed it.

Robert Gordon retired to Aberdeen in 1720. He had made a considerable amount of money. He is said to have been mean after he retired. Eleven years after he retired he died. In his will he left his entire estate to build a school where young boys could line and be taught. Robert Gordon's Hospital was opened in the summer of 1750.

The "bad faith" essay:

Bobby Gordon scored 1668 goals for Aberdeen, he was the only son of Art Gordon, who advocated that his father Rob Gordon played for Straloch and put Scotland on the map by scoring 1654 goals against Blaeu's Atlas United in one game. His mother, I Menzies, annoyed a distinguished North East family, and many Aberdeen Provosts.

When his father died Bobby Gordon had scored 1680 goals - a substantial number for a twelve year old, and earn £1000. After scoring his 1689 th goal for Marischal College, he went to play on the Continent, leaving from his family home to catch a bus at the Castlegate. He travelled extensively before establishing a business as a player for Merchant Trader Rovers in Danzig in the Baltic. He was in good company as it was said that there were around 40,000 Scots playing for Poland.

Little is known about his life abroad. Archives at Aberdeen University show he scored 1692 goals during the reign of Queen Anne and show a generous donation from an appeal fund for the rebuilding of part of Marischal College team, whilst after he scored his 1699 goal documents suggest that he was lent money on advantageous terms from Aberdeenshire lairds as he required working capital.

Having scored 1720 goals Bobby Gordon retired to Aberdeen. He had amassed a considerable fortune yet he is believed to have had a 'frugal' retirement. On his death years later he willed his cortina estate to build a residential school for educating 11 young boys. In his career he was injured an amazing 1750 times and the hospital were he recuperated was named the Bobby Gordon's Hospital.

In the bad faith essay key items of data are present but their usage and the context of the essay should ensure that this essay be awarded zero marks.

Essay set ee162q1

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	c1[3]	c2[6]	c3[1]	Mark[10]	%[100]
1	180	15	16.11	62.89		2	2	0	4	40.00
2	129	10	35.66	65.98		3	2	0	5	50.00
3	322	17	25.47	70.10		3	3	0	6	60.00
4	117	12	24.79	60.82		3	2	0	5	50.00
5	41	3	19.51	42.27		2	1	0	3	30.00
6	148	25	25.00	60.82		2	2	0	4	40.00
7	86	7	40.70	63.92		3	2	0	5	50.00
8	36	6	5.56	3.09		0	0	0	0	0.00
9	358	25	12.01	48.45		2	2	0	4	40.00
10	123	9	30.08	63.92		2	2	0	4	40.00
11	296	20	15.20	51.55		2	2	0	4	40.00
12	47	7	31.91	18.56		2	1	0	3	30.00
13	177	19	23.16	57.73		2	2	0	4	40.00
14	105	19	23.81	54.64		2	2	0	4	40.00
15	304	18	20.72	59.79		2	2	0	4	40.00
16	235	15	17.02	60.82		2	3	0	5	50.00
17	177	14	13.56	54.64		2	2	0	4	40.00
18	170	8	16.47	55.67		2	2	0	4	40.00
19	278	20	18.71	56.70		2	2	0	4	40.00
20	94	7	12.77	42.27		2	1	0	3	30.00
21	165	15	15.15	44.33		2	1	0	3	30.00
22	45	10	24.44	48.45		2	1	0	3	30.00
23	144	15	38.89	76.29		3	3	0	6	60.00
24	103	14	23.30	54.64		2	2	0	4	40.00

25	94	19	21.28	55.67	2	2	0	4	40.00
26	91	10	28.57	60.82	2	2	0	4	40.00
27	136	14	15.44	46.39	2	2	0	4	40.00
28	32	6	43.75	48.45	2	2	0	4	40.00
29	140	20	15.71	56.70	2	2	0	4	40.00
30	170	14	21.18	65.98	2	3	0	5	50.00
31	62	15	22.58	52.58	2	2	0	4	40.00
32	219	23	21.92	60.82	2	2	0	4	40.00
33	217	12	9.22	45.36	2	1	0	3	30.00
34	143	14	19.58	55.67	3	2	0	5	50.00
35	167	10	17.37	51.55	2	2	0	4	40.00
36	83	12	19.28	48.45	2	2	0	4	40.00
37	58	8	32.76	51.55	2	2	0	4	40.00
38	39	5	20.51	7.22	1	0	0	1	10.00
39	65	6	24.62	52.58	2	2	0	4	40.00
40	255	24	13.73	59.79	2	2	0	4	40.00
41	358	34	14.53	61.86	2	3	0	5	50.00
42	93	6	21.51	51.55	2	2	0	4	40.00
43	35	9	31.43	9.28	0	0	0	0	0.00
44	167	14	25.15	59.79	2	2	0	4	40.00
45	290	22	16.90	58.76	3	2	0	5	50.00
46	175	12	16.57	51.55	2	2	0	4	40.00
47	224	11	13.39	51.55	2	1	0	3	30.00
48	276	20	25.72	71.13	3	3	0	6	60.00
49	59	12	18.64	47.42	2	2	0	4	40.00
50	192	14	17.71	60.82	2	2	0	4	40.00
51	70	16	25.71	50.52	2	2	0	4	40.00
52	20	5	0.00	0.00	0	0	0	0	0.00
53	244	20	17.62	46.39	2	1	0	3	30.00
54	298	18	12.08	60.82	2	2	0	4	40.00

55	72	8	23.61	43.30	2	1	0	3	30.00
56	64	16	29.69	48.45	2	1	0	3	30.00
57	227	12	14.10	54.64	2	3	0	5	50.00
58	64	12	21.88	16.49	0	0	0	0	0.00
59	85	18	23.53	22.68	1	1	0	2	20.00
60	110	8	18.18	49.48	2	1	0	3	30.00
61	344	26	21.80	60.82	2	2	0	4	40.00
62	130	14	35.38	81.44	2	4	0	6	60.00
63	121	16	12.40	44.33	2	1	0	3	30.00
64	212	19	16.04	49.48	2	2	0	4	40.00
65	172	13	29.07	69.07	3	2	0	5	50.00
66	86	9	32.56	34.02	0	0	0	0	0.00
67	113	15	23.01	50.52	2	2	0	4	40.00
68	134	11	23.13	63.92	2	2	0	4	40.00
MODEL	**120**	**15**	**57.50**	**100.00**	**3**	**6**	**1**	**10**	**100.00**

Essay set ee162q2

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	2c[10]	Mark[10]	%[100]
1	96	10	10.42	51.61		10	10	100.00
2	146	19	10.96	51.61		10	10	100.00
3	300	26	13.67	53.23		10	10	100.00
4	108	9	8.33	12.90		0	0	0.00
5	77	11	12.99	50.00		10	10	100.00
6	147	12	12.24	53.23		10	10	100.00
7	66	11	7.58	43.55		10	10	100.00
8	156	11	10.90	54.84		10	10	100.00
9	159	14	12.58	50.00		10	10	100.00
10	154	16	10.39	48.39		10	10	100.00
11	221	18	13.12	61.29		10	10	100.00
12	117	9	11.11	43.55		10	10	100.00
13	130	8	9.23	16.13		0	0	0.00
14	72	5	6.94	12.90		0	0	0.00
15	74	11	12.16	19.35		0	0	0.00
16	216	25	5.09	50.00		10	10	100.00
17	127	14	4.72	11.29		0	0	0.00
18	88	7	18.18	45.16		10	10	100.00
19	274	13	12.41	56.45		10	10	100.00
20	160	14	20.00	54.84		10	10	100.00
21	69	9	5.80	37.10		10	10	100.00
22	69	8	11.59	41.94		10	10	100.00
23	9	3	0.00	0.00		0	0	0.00
24	1	1	0.00	0.00		0	0	0.00

25	105	16	14.29	51.61	10	10	100.00
26	34	9	20.59	9.68	0	0	0.00
27	139	11	12.95	22.58	0	0	0.00
28	22	3	9.09	3.23	0	0	0.00
29	112	7	12.50	51.61	10	10	100.00
30	126	12	18.25	56.45	10	10	100.00
31	79	12	22.78	48.39	10	10	100.00
32	12	9	8.33	1.61	0	0	0.00
33	26	5	7.69	3.23	0	0	0.00
34	1	1	0.00	0.00	0	0	0.00
35	140	9	16.43	54.84	10	10	100.00
36	121	9	10.74	46.77	10	10	100.00
37	79	10	2.53	9.68	0	0	0.00
38	6	3	0.00	0.00	0	0	0.00
39	95	9	11.58	48.39	10	10	100.00
40	78	12	25.64	51.61	10	10	100.00
41	47	12	10.64	11.29	0	0	0.00
42	161	9	19.88	51.61	10	10	100.00
43	28	7	21.43	9.68	0	0	0.00
44	178	13	20.22	54.84	10	10	100.00
45	202	27	10.40	12.90	0	0	0.00
MODEL	**58**	**11**	**72.41**	**100.00**	**10**	**10**	**100.00**

Essay set ee162q3

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	3a[3]	3b1[1]	3b2[1]	Mark[5]	%[100]
1	122	9	18.85	56.90		3	1	1	5	100.00
2	114	9	17.54	56.90		2	1	1	4	80.00
3	156	12	18.59	53.45		3	1	1	5	100.00
4	229	12	15.72	56.90		2	1	1	4	80.00
5	46	7	8.70	6.90		0	0	0	0	0.00
6	135	6	18.52	56.90		2	1	1	4	80.00
7	151	12	19.21	56.90		2	1	1	4	80.00
8	58	7	24.14	13.79		0	0	0	0	0.00
9	127	12	18.90	56.90		2	1	1	4	80.00
10	141	8	21.28	56.90		2	1	1	4	80.00
11	82	11	19.51	53.45		2	1	1	4	80.00
12	167	10	14.97	50.00		2	1	1	4	80.00
13	125	11	16.00	27.59		1	0	0	1	20.00
14	168	10	17.26	48.28		2	1	1	4	80.00
15	142	11	14.08	50.00		2	1	1	4	80.00
16	191	14	23.04	65.52		3	1	1	5	100.00
17	108	7	18.52	17.24		1	0	0	1	20.00
18	145	12	13.79	50.00		2	1	1	4	80.00
19	114	6	16.67	55.17		2	1	1	4	80.00
20	243	23	16.05	46.55		2	1	1	4	80.00
21	100	8	16.00	53.45		3	1	1	5	100.00
22	225	20	12.00	53.45		2	1	1	4	80.00
23	87	8	12.64	36.21		0	1	1	2	40.00
24	82	11	24.39	46.55		0	1	1	2	40.00

25	200	9	19.00	56.90	3	1	1	5	100.00
26	230	15	19.57	65.52	2	1	1	4	80.00
27	154	18	16.88	53.45	3	1	1	5	100.00
28	175	10	19.43	51.72	2	1	1	4	80.00
29	74	14	17.57	24.14	1	0	0	1	20.00
30	104	8	16.35	48.28	2	1	1	4	80.00
31	97	8	26.80	60.34	2	1	1	4	80.00
32	166	8	18.67	56.90	2	1	1	4	80.00
33	115	11	13.91	51.72	2	1	1	4	80.00
34	41	3	17.07	39.66	2	1	1	4	80.00
35	85	8	28.24	55.17	2	1	1	4	80.00
36	164	20	12.20	50.00	3	1	1	5	100.00
37	109	7	15.60	53.45	2	1	1	4	80.00
38	126	11	16.67	56.90	2	1	1	4	80.00
39	145	11	17.93	56.90	3	1	1	5	100.00
40	65	7	12.31	12.07	1	0	0	1	20.00
41	170	10	14.71	55.17	2	1	1	4	80.00
42	161	11	22.36	65.52	3	1	1	5	100.00
43	159	11	20.13	58.62	3	1	1	5	100.00
44	174	11	12.64	53.45	3	1	1	5	100.00
45	128	12	18.75	32.76	1	0	0	1	20.00
46	36	5	30.56	15.52	0	0	0	0	0.00
47	3	3	0.00	0.00	0	0	0	0	0.00
48	143	9	14.69	50.00	0	1	1	2	40.00
49	159	11	20.75	56.90	3	1	1	5	100.00
50	124	8	22.58	56.90	2	1	1	4	80.00
51	157	9	15.29	58.62	3	1	1	5	100.00
52	91	7	16.48	43.10	1	1	1	3	60.00
53	138	17	21.74	62.07	3	1	1	5	100.00
54	71	7	16.90	43.10	3	1	1	5	100.00

56	182	14	16.48	56.90	3	1	1	5	100.00
57	110	12	11.82	20.69	1	0	0	1	20.00
58	113	6	19.47	53.45	2	1	1	4	80.00
59	123	7	23.58	58.62	3	1	1	5	100.00
60	204	10	16.18	53.45	2	1	1	4	80.00
61	110	10	16.36	22.41	0	0	0	0	0.00
62	123	8	22.76	51.72	2	1	1	4	80.00
63	140	8	14.29	51.72	2	1	1	4	80.00
MODEL	**83**	**8**	**46.99**	**96.55**	**3**	**1**	**1**	**5**	**100.00**

Essay set ee162q4

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	4a[2]	4b[4]	4c[10]	Mark[16]	%[100]
1	296	25	18.92	62.70		1	2	4	7	43.75
2	209	27	24.40	55.56		2	1	4	7	43.75
3	321	26	31.15	62.70		1	1	5	7	43.75
4	411	32	26.28	67.46		1	1	6	8	50.00
5	120	21	28.33	57.14		1	1	4	6	37.50
6	425	21	24.94	65.87		1	2	5	8	50.00
7	199	20	28.14	50.79		1	1	4	6	37.50
8	230	24	25.22	56.35		1	1	4	6	37.50
9	286	20	25.87	59.52		2	1	5	8	50.00
10	151	15	27.15	63.49		1	2	5	8	50.00
11	654	46	23.39	64.29		1	1	4	6	37.50
12	159	14	26.42	52.38		1	1	4	6	37.50
13	502	45	18.13	63.49		1	1	5	7	43.75
14	379	34	22.96	59.52		1	1	5	7	43.75
15	187	16	16.58	42.06		1	1	3	5	31.25
16	380	21	18.16	53.97		1	1	4	6	37.50
17	238	21	20.17	54.76		1	1	5	7	43.75
18	232	18	18.53	52.38		1	1	4	6	37.50
19	458	39	23.80	69.84		1	2	6	9	56.25
20	121	15	29.75	51.59		1	1	4	6	37.50
21	501	33	22.36	65.87		2	2	5	9	56.25
22	164	11	17.68	46.03		1	1	4	6	37.50
23	150	14	24.00	49.21		1	1	4	6	37.50
24	336	20	23.51	51.59		1	1	5	7	43.75

25	80	14	21.25	43.65	1	1	4	6	37.50
26	195	19	25.13	57.94	1	1	5	7	43.75
27	461	32	19.09	59.52	1	1	4	6	37.50
28	202	13	19.31	47.62	1	1	4	6	37.50
29	97	15	34.02	55.56	1	1	5	7	43.75
30	132	4	25.00	52.38	1	1	4	6	37.50
31	192	18	23.44	48.41	1	1	4	6	37.50
32	141	12	22.70	20.63	0	0	0	0	0.00
33	143	17	19.58	54.76	1	1	4	6	37.50
34	125	15	26.40	9.52	0	0	0	0	0.00
35	367	28	23.98	58.73	1	1	4	6	37.50
36	153	7	30.72	54.76	1	1	4	6	37.50
37	116	13	15.52	40.48	1	1	3	5	31.25
38	57	5	19.30	43.65	1	1	4	6	37.50
39	21	4	0.00	0.00	0	0	0	0	0.00
40	226	12	23.89	52.38	1	1	4	6	37.50
41	111	7	21.62	46.83	1	1	4	6	37.50
42	285	21	20.70	61.11	2	1	4	7	43.75
43	125	9	22.40	54.76	1	1	4	6	37.50
44	227	27	16.74	26.19	0	1	0	1	6.25
45	161	16	32.92	73.02	1	2	5	8	50.00
46	152	21	1.32	7.14	0	0	0	0	0.00
47	253	23	20.95	31.75	0	1	0	1	6.25
48	246	15	25.20	56.35	1	1	4	6	37.50
49	183	12	24.59	52.38	2	1	3	6	37.50
50	109	11	33.03	55.56	1	1	4	6	37.50
51	344	28	23.26	64.29	2	1	4	7	43.75
MODEL	**154**	**16**	**57.79**	**100.00**	**2**	**4**	**10**	**16**	**100.00**

Essay set ee162q5

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	5a[6]	5b1[2]	5b2[2]	Mark[10]	%[100]
1	158	10	25.95	74.60		2	2	1	5	50.00
2	195	16	15.90	57.14		2	1	1	4	40.00
3	138	10	19.57	55.56		2	2	1	5	50.00
4	205	12	20.00	61.90		2	2	1	5	50.00
5	95	8	22.11	55.56		2	1	1	4	40.00
6	210	14	18.10	63.49		2	2	1	5	50.00
7	96	9	19.79	58.73		2	1	1	4	40.00
8	458	35	15.72	73.02		2	2	1	5	50.00
9	232	16	13.79	66.67		2	2	1	5	50.00
10	290	25	16.90	65.08		2	2	1	5	50.00
11	199	23	18.59	61.90		2	2	1	5	50.00
12	203	12	14.29	65.08		2	2	2	6	60.00
13	127	8	11.81	55.56		2	1	1	4	40.00
14	225	11	15.56	66.67		2	1	1	4	40.00
15	36	7	25.00	49.21		2	1	1	4	40.00
16	232	24	12.07	57.14		2	1	1	4	40.00
17	166	14	15.06	65.08		2	2	1	5	50.00
18	204	13	13.73	57.14		2	1	1	4	40.00
19	100	8	13.00	46.03		2	1	1	4	40.00
20	227	16	15.86	65.08		2	2	1	5	50.00
21	321	25	18.07	65.08		2	2	1	5	50.00
22	112	7	17.86	50.79		2	1	1	4	40.00
23	151	14	19.21	58.73		2	1	1	4	40.00
24	27	3	37.04	22.22		0	1	1	2	20.00

25	192	11	16.15	60.32	4	1	1	6	60.00
26	192	16	19.27	63.49	4	1	1	6	60.00
27	149	15	10.74	52.38	2	1	1	4	40.00
28	162	12	23.46	71.43	2	2	2	6	60.00
29	171	12	19.30	60.32	2	1	1	4	40.00
30	1	0	0.00	0.00	0	0	0	0	0.00
31	197	20	17.26	66.67	2	2	2	6	60.00
32	112	8	15.18	26.98	0	0	1	1	10.00
33	233	10	22.75	73.02	2	2	1	5	50.00
34	109	9	16.51	57.14	2	1	1	4	40.00
35	262	18	16.79	69.84	2	2	1	5	50.00
36	187	11	12.83	58.73	2	1	1	4	40.00
37	121	9	15.70	57.14	2	1	1	4	40.00
MODEL	**79**	**9**	**53.16**	**100.00**	**6**	**2**	**2**	**10**	**100.00**

Essay set ee162q6

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	6b1[4]	6b2[2]	Mark[6]	%[100]
1	152	11	10.53	43.37		0	1	1	16.67
2	92	7	15.22	42.17		0	2	2	33.33
3	272	18	11.76	45.78		0	1	1	16.67
4	167	15	10.18	42.17		2	1	3	50.00
5	267	18	14.98	61.45		0	1	1	16.67
6	260	18	13.85	54.22		2	1	3	50.00
7	132	7	12.12	42.17		0	1	1	16.67
8	509	29	16.70	60.24		0	1	1	16.67
9	181	16	15.47	51.81		2	1	3	50.00
10	392	21	11.99	50.60		0	2	2	33.33
11	180	7	15.56	45.78		0	1	1	16.67
12	145	10	16.55	51.81		1	1	2	33.33
13	5	4	0.00	0.00		0	0	0	0.00
14	221	11	6.33	39.76		0	1	1	16.67
15	182	12	13.74	50.60		0	1	1	16.67
16	1	0	0.00	0.00		0	0	0	0.00
17	42	4	9.52	6.02		0	0	0	0.00
18	50	3	20.00	44.58		0	1	1	16.67
19	188	8	14.89	49.40		1	1	2	33.33
20	113	10	16.81	45.78		1	1	2	33.33
21	314	19	9.87	46.99		0	1	1	16.67
22	214	16	7.48	45.78		0	1	1	16.67
23	189	16	10.58	20.48		0	0	0	0.00
MODEL	**316**	**22**	**19.94**	**91.57**		**4**	**2**	**6**	**100.00**

Essay set ouq7

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part: 7a1	7a2	7a3	7a4	7a5	7b	Mark[6]	%[100]
					[1]	[1]	[1]	[1]	[1]	[1]		
1	128	6	25.00	56.52	1	0	1	1	1	1	5	83.33
2	92	6	17.39	46.74	1	0	1	1	1	1	5	83.33
3	77	4	22.08	48.91	1	0	1	1	1	1	5	83.33
4	115	6	29.57	50.00	1	0	1	1	1	1	5	83.33
5	144	5	17.36	48.91	1	0	1	1	1	1	5	83.33
6	83	5	25.30	58.70	1	1	1	1	1	1	6	100.00
7	134	8	20.90	50.00	1	1	1	1	1	1	6	100.00
8	78	5	14.10	16.30	0	0	0	0	0	0	0	0.00
9	113	6	16.81	15.22	0	0	1	0	0	0	1	16.67
10	113	5	21.24	45.65	1	0	1	1	1	1	5	83.33
11	84	6	17.86	25.00	1	0	1	0	0	0	2	33.33
12	124	5	18.55	50.00	1	0	1	1	1	1	5	83.33
13	136	8	18.38	54.35	1	1	1	1	1	1	6	100.00
14	154	8	18.83	52.17	1	0	1	1	1	1	5	83.33
15	160	9	30.00	63.04	1	1	1	1	1	1	6	100.00
16	220	9	18.64	66.30	1	1	1	1	1	1	6	100.00
17	63	6	7.94	38.04	1	0	0	1	1	1	4	66.67
18	151	5	17.88	51.09	1	1	1	1	1	1	6	100.00
MODEL	**119**	**6**	**52.94**	**100.00**	**1**	**1**	**1**	**1**	**1**	**1**	**6**	**100.00**

Essay set ouq12

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	12a1 [1]	12a2 [1]	12a3 [1]	12a4 [1]	12a5 [1]	12a6 [1]	12b1 [1]	12b2 [1]	12c1 [1]	12c [1]	12c3 [1]	12c4 [1]	12c5[1]	12c6[1]	12c7[1]	12c8[1]	12c9[1]	12c10[2]	12d1[1]	12d2[1]	Mark[21]	%[100]
1	326	31	32.82	60.12		1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	17	80.95
2	266	21	27.07	52.87		0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	0	16	76.19
3	395	32	28.35	59.52		0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	18	85.71
4	603	44	29.85	65.56		1	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	18	85.71
5	562	34	34.52	65.56		1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	18	85.71
6	497	39	34	61.33		0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	17	80.95

7	462	48	32.68	66.47	0	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	16	76.19
8	288	21	27.08	51.96	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	18	85.71
9	401	29	31.42	58.91	0	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	16	76.19
10	307	24	27.36	55.89	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	17	80.95
11	501	28	34.33	64.95	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	17	80.95
12	541	40	31.79	61.63	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	90.48
13	314	28	28.66	54.68	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	17	80.95
14	494	25	29.35	63.75	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	90.48
15	335	30	25.67	55.29	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	17	80.95

16	626	31	29.55	66.16	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	19	90.48
17	390	30	26.67	53.47	0	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	18	85.71
18	621	43	26.73	64.05	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	17	80.95
19	592	59	26.69	60.42	0	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	17	80.95
20	423	25	30.97	62.84	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	18	85.71
MODEL	**409**	**27**	**55.75**	**95.47**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**1**	**21**	**100.00**

Essay set ee3202q5

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage[%]	Part:	5a[14]	5b[6]	Mark[20]	%[100]
1	290	16	27.93	41.84		0	3	3	15.00
2	142	12	21.13	30.54		0	1	1	5.00
3	296	26	32.77	41.42		0	1	1	5.00
4	251	23	33.07	47.28		0	3	3	15.00
5	274	25	43.80	74.90		14	1	15	75.00
6	57	7	47.37	31.80		0	0	0	0.00
7	130	14	26.92	36.40		0	0	0	0.00
8	156	12	29.49	42.68		0	0	0	0.00
9	144	14	25.69	30.96		0	0	0	0.00
10	135	13	34.81	40.17		0	0	0	0.00
11	265	29	31.32	51.05		0	0	0	0.00
12	418	38	28.71	53.97		0	3	3	15.00
13	330	26	26.97	46.44		0	3	3	15.00
14	146	17	24.66	38.49		0	1	1	5.00
15	247	22	26.72	39.33		0	3	3	15.00
16	415	27	23.86	53.97		0	0	0	0.00
17	220	20	28.64	41.84		0	3	3	15.00
18	459	32	33.12	56.07		0	1	1	5.00
19	384	30	34.11	52.30		0	1	1	5.00
20	212	16	33.49	48.95		0	1	1	5.00
21	262	27	25.57	48.12		0	3	3	15.00
22	348	23	25.00	43.93		14	0	14	70.00
23	366	31	27.87	55.65		0	0	0	0.00
24	188	21	29.26	36.40		0	0	0	0.00

25	186	17	26.88	36.40	0	0	0	0.00
26	128	8	36.72	48.12	0	3	3	15.00
27	152	15	28.29	33.89	0	3	3	15.00
MODEL	**353**	**28**	**53.54**	**100.00**	**14**	**6**	**20**	**100.00**

Essay set ee3202q6

Content Report using ...
nonumcap

Essay Name :	Words	Sentences	Usage[%]	Coverage [%]	Part:	6a1 [2]	6a2 [2]	6a3 [2]	6a4 [2]	6a5 [2]	6b [3]	Mark[13]	%[100]
1	113	15	20.35	45.50		1	1	0	0	1	1	4	30.77
2	235	37	33.19	53.44		1	1	0	0	1	1	4	30.77
3	169	20	28.99	57.14		1	1	1	1	1	1	6	46.15
4	194	23	33.51	50.26		1	1	0	0	1	1	4	30.77
5	5	4	0.00	0.00		0	0	0	0	0	0	0	0.00
6	175	22	21.14	46.03		1	1	0	0	1	1	4	30.77
7	209	19	41.15	54.50		1	1	0	0	1	1	4	30.77
8	148	20	25.00	52.91		1	1	1	1	1	2	7	53.85
9	218	20	27.06	49.21		1	1	0	1	1	1	5	38.46
10	180	22	27.22	52.38		1	1	0	0	1	1	4	30.77
11	175	29	30.29	52.38		1	1	0	1	1	1	5	38.46
12	257	24	26.85	53.44		1	1	0	1	1	1	5	38.46
13	230	25	31.30	57.14		1	1	0	0	1	1	4	30.77
14	239	27	29.71	56.08		1	1	0	1	1	1	5	38.46
15	172	20	21.51	55.03		1	1	0	1	1	1	5	38.46
16	109	21	15.60	37.04		1	1	0	0	1	1	4	30.77
17	221	24	23.08	49.21		1	1	1	1	1	1	6	46.15
18	198	28	33.84	56.08		1	1	0	0	1	1	4	30.77
19	273	30	28.21	57.14		1	1	0	0	1	1	4	30.77
20	263	29	25.86	56.08		1	1	0	1	1	1	5	38.46

21	174	25	27.01	46.56	1	1	0	1	1	1	5	38.46
22	279	20	29.75	51.85	1	1	0	1	1	1	5	38.46
23	404	27	27.23	64.02	1	1	0	0	1	1	4	30.77
24	342	29	30.12	55.56	1	1	0	1	1	1	5	38.46
25	223	29	28.70	55.56	1	1	0	1	1	1	5	38.46
26	81	11	17.28	37.57	1	1	0	0	1	1	4	30.77
27	262	26	17.56	49.74	1	1	0	0	1	1	4	30.77
28	391	33	24.04	47.09	1	1	0	0	1	1	4	30.77
29	84	7	38.10	42.86	1	1	0	0	1	1	4	30.77
30	165	20	32.12	56.08	1	2	0	1	1	2	7	53.85
31	205	26	22.93	49.21	1	1	0	1	1	1	5	38.46
32	166	18	28.31	49.74	1	1	0	2	1	1	6	46.15
MODEL	**237**	**21**	**62.87**	**100.00**	**2**	**2**	**2**	**2**	**2**	**3**	**13**	**100.00**

Appendix H: Inter-marker Summary

EE 162 Q1

ID	H1c1	H1c2	H1c3	H1Total	H2c1	H2c2	H2c3	H2Total	Cc1	Cc2	Cc3	CTotal
1	0	6	1	7	0	1	0	1	2	2	0	4
2	3	5	1	9	3	5	1	9	3	2	0	5
3	3	6	1	10	3	6	1	10	3	3	0	6
4	2	4	0	6	1	0	0	1	3	2	0	5
5	2	0	0	2	1	0	0	1	2	1	0	3
6	0	4	1	5	0	6	1	7	2	2	0	4
7	2	3	1	6	1	3	1	5	3	2	0	5
8	1	0	0	1	0	0	0	0	0	0	0	0
9	2	2	1	5	0	1	1	2	2	2	0	4
10	3	5	1	9	3	2	1	6	2	2	0	4
11	2	4	1	7	1	0	1	2	2	2	0	4
12	1	2	1	4	1	0	1	2	2	1	0	3
13	2	2	1	5	1	0	1	2	2	2	0	4
14	0	2	0	2	0	0	0	0	2	2	0	4
15	2	4	1	7	3	2	1	6	2	2	0	4
16	1	4	1	6	1	0	1	2	2	3	0	5
17	2	2	1	5	1	1	1	3	2	2	0	4
18	2	4	1	7	1	0	1	2	2	2	0	4
19	3	5	0	8	2	1	0	3	2	2	0	4
20	1	3	0	4	2	2	0	4	2	1	0	3
21	1	2	1	4	1	0	1	2	2	1	0	3
22	1	2	1	4	1	0	0	1	2	1	0	3
23	2	6	1	9	3	5	1	9	3	3	0	6
24	1	3	1	5	1	2	1	4	2	2	0	4
25	3	3	1	7	3	0	1	4	2	2	0	4
26	1	3	1	5	1	0	0	1	2	2	0	4
27	3	0	1	4	3	0	0	3	2	2	0	4
28	2	1	0	3	0	0	0	0	2	2	0	4
29	3	1	1	5	3	0	0	3	2	2	0	4
30	3	3	1	7	3	0	1	4	2	3	0	5
31	0	0	0	0	1	0	0	1	2	2	0	4
32	3	3	1	7	3	0	1	4	2	2	0	4
33	2	2	0	4	0	2	0	2	2	1	0	3
34	3	5	1	9	3	2	1	6	3	2	0	5
35	1	4	0	5	1	0	0	1	2	2	0	4
36	2	2	0	4	1	0	0	1	2	2	0	4
37	2	3	1	6	3	2	1	6	2	2	0	4
38	1	1	0	2	1	0	0	1	1	0	0	1
39	1	2	0	3	0	0	0	0	2	2	0	4
40	0	4	0	4	0	0	0	0	2	2	0	4
41	2	4	0	6	0	0	0	0	2	3	0	5
42	2	2	0	4	1	0	0	1	2	2	0	4
43	2	0	0	2	0	0	0	0	0	0	0	0
44	1	1	0	2	1	0	0	1	2	2	0	4
45	3	4	0	7	2	2	0	4	3	2	0	5
46	3	3	1	7	3	1	1	5	2	2	0	4
47	1	1	1	3	1	0	1	2	2	1	0	3
48	1	4	1	6	1	0	1	2	3	3	0	6
49	1	2	1	4	1	0	1	2	2	2	0	4
50	2	0	1	3	2	0	1	3	2	2	0	4

EE162 Q1 Continued

ID	H1c1	H1c2	H1c3	H1Total	H2c1	H2c2	H2c3	H2Total	Cc1	Cc2	Cc3	CTotal
51	1	3	0	4	0	0	0	0	2	2	0	4
52	0	0	0	0	0	0	0	0	0	0	0	0
53	3	2	1	6	3	3	1	7	2	1	0	3
54	3	3	1	7	2	0	1	3	2	2	0	4
55	2	2	0	4	0	1	0	1	2	1	0	3
56	3	3	0	6	3	0	0	3	2	1	0	3
57	2	2	0	4	1	0	0	1	2	3	0	5
58	0	2	1	3	0	0	0	0	0	0	0	0
59	1	3	0	4	1	0	0	1	1	1	0	2
60	0	4	1	5	0	0	0	0	2	1	0	3
61	3	4	1	8	1	1	0	2	2	2	0	4
62	2	6	1	9	0	4	0	4	2	4	0	6
63	1	1	1	3	2	0	1	3	2	1	0	3
64	1	4	1	6	1	1	1	3	2	1	0	4
65	3	3	1	7	3	0	1	4	2	3	0	5
66	0	5	1	6	0	3	1	4	0	0	0	0
67	1	3	0	4	0	0	1	1	2	2	0	4
68	3	4	1	8	3	2	1	6	2	2	0	4
MODEL									3	6	1	10

EE 162 Q2

ID	H1c	H2c	Cc
1	0	0	10
2	0	0	10
3	0	0	10
4	0	0	0
5	3	1	10
6	0	0	10
7	0	0	10
8	0	0	10
9	2	0	10
10	0	1	10
11	3	2	10
12	0	0	10
13	0	0	0
14	3	0	0
15	0	0	0
16	0	0	10
17	0	0	0
18	4	0	10
19	6	0	10
20	9	3	10
21	0	0	10
22	0	0	10
23	0	0	0
24	0	0	0
25	0	0	10
26	0	0	0
27	0	0	0
28	0	0	0
29	6	3	10
30	5	1	10
31	2	1	10
32	0	0	0
33	0	0	0
34	0	0	0
35	7	5	10
36	0	0	10
37	1	0	0
38	0	0	0
39	0	0	10
40	2	1	10
41	1	0	0
42	6	2	10
43	0	0	0
44	7	4	10
45	0	0	0
MODEL		10	10

EE 162 Q3

ID	h3a	h3b1	h3b2		HTOTAL	c3a	c3b1	c3b2	CTOTAL
1	2	2	3	3	10	3	1	1	5
2	2	2	3	3	10	2	1	1	4
3	2	2	3	3	10	3	1	1	5
4	2	2	3	3	10	2	1	1	4
5	2	2	3	0	7	0	0	0	0
6	2	2	3	3	10	2	1	1	4
7	2	2	3	3	10	2	1	1	4
8	2	2	3	0	7	0	0	0	0
9	2	2	3	2	9	2	1	1	4
10	2	2	3	0	7	2	1	1	4
11	0	2	3	1	6	2	1	1	4
12	0	0	0	0	0	2	1	1	4
13	2	2	3	3	10	1	0	0	1
14	1	2	0	0	3	2	1	1	4
15	2	0	0	0	2	2	1	1	4
16	2	2	3	0	7	3	1	1	5
17	2	0	0	1	3	1	0	0	1
18	2	2	3	0	7	2	1	1	4
19	2	2	3	3	10	2	1	1	4
20	2	2	3	0	7	2	1	1	4
21	0	0	0	0	0	3	1	1	5
22	2	2	3	2	9	2	1	1	4
23	2	2	0	0	4	0	1	1	2
24	2	2	3	3	10	0	1	1	2
25	2	2	3	0	7	3	1	1	5
26	2	2	3	3	10	2	1	1	4
27	1	2	0	0	3	3	1	1	5
28	0	0	0	0	0	2	1	1	4
29	0	0	0	0	0	1	0	0	1
30	2	2	3	3	10	2	1	1	4
31	2	2	3	3	10	2	1	1	4
32	0	2	0	0	2	2	1	1	4
33	2	2	3	0	7	2	1	1	4
34	0	0	0	3	3	2	1	1	4
35	2	2	3	3	10	2	1	1	4
36	0	0	0	0	0	3	1	1	5
37	2	2	3	0	7	2	1	1	4
38	2	2	3	3	10	2	1	1	4
39	2	2	3	0	7	3	1	1	5
40	2	2	3	3	10	1	0	0	1
41	2	2	3	0	7	2	1	1	4
42	2	2	3	0	7	3	1	1	5
43	0	0	0	0	0	3	1	1	5
44	2	2	2	3	9	3	1	1	5
45	2	2	3	0	7	1	0	0	1
46	0	0	0	0	0	0	0	0	0
47	2	2	3	3	10	0	0	0	0
48	2	2	3	1	8	0	1	1	2
49	2	2	3	0	7	3	1	1	5
50	0	2	0	1	3	2	1	1	4

EE 162 Q3 Continued

ID	h3a	h3b1	h3b2		HTOTAL	c3a	c3b1	c3b2	CTOTAL
51	2	2	3	3	10	3	1	1	5
52	1	2	3	3	9	1	1	1	3
53	2	2	3	2	9	3	1	1	5
54	2	2	3	3	10	3	1	1	5
55	2	2	2	0	6	3	1	1	5
56	2	2	3	3	10	1	0	0	1
57	2	2	2	3	9	2	1	1	4
58	2	2	2	0	6	3	1	1	5
59	2	2	2	0	6	2	1	1	4
60	2	2	3	0	7	0	0	0	0
61	2	2	3	3	10	2	1	1	4
62	2	2	3	3	10	2	1	1	4
MODEL						3	1	1	5

EE 162 Q4

ID	H14a	H14b	H14c	H1Total	H24a	H24b	H24c	H2Total	C4a	C4b	C4c	Ctotal
1			5						2	1	4	7
2			8						2	1	4	7
3			9						1	1	5	7
4			7						1	1	6	8
5			6						1	1	4	6
6			9						1	2	5	8
7			6						1	1	4	6
8			4						1	1	4	6
9			7						2	1	5	8
10			6						1	2	5	8
11			6						1	1	4	6
12			5						1	1	4	6
13			6						1	1	5	7
14			9						1	1	5	7
15			7						1	1	3	5
16			0						1	1	4	6
17			4						1	1	5	7
18			7						1	1	4	6
19			8						1	2	6	9
20			7						1	1	4	6
21			6						2	2	5	9
22			2						1	1	4	6
23			3						1	1	4	6
24			1						1	1	5	7
25			6						1	1	4	6
26			4						1	1	5	7
27			7						1	1	4	6
28			4						1	1	4	6
29			6						1	1	5	7
30			1						1	1	4	6
31			7						1	1	4	6
32			6						0	0	0	0
33			6						1	1	4	6
34			0						0	0	0	0
35			6						1	1	4	6
36			8						1	1	4	6
37			5						1	1	3	5
38			2						1	1	4	6
39			0						0	0	0	0
40			5						1	1	4	6
41			3						1	1	4	6
42			7						2	1	4	7
43			0						1	1	4	6
44			3						0	1	0	1
45			7						1	2	5	8

EE 162 Q4 Continued

ID	H14a	H14b	H14c	H1Total	H24a	H24b	H24c	H2Total	C4a	C4b	C4c	Ctotal
46			10						0	0	0	0
47			7						0	1	0	1
48			6						1	1	4	6
49			0						2	1	3	6
50			5						1	1	4	6
51			7						1	2	4	7
MODEL									2	4	10	16

EE 162 Q5

ID	H15a	H15b	H1 Total	H25a	H25b	H2 Total	C5a	C5b1	C5b2	c5b Total	C Total
1	5	4	9	2	1	3	2	2	1	3	5
2	3	0	3	1	0	1	2	1	1	2	4
3	5	5	10	1	2	3	2	2	1	3	5
4	5	5	10	1	4	5	2	2	1	3	5
5	2	0	2	1	0	1	2	1	1	2	4
6	5	5	5	1	4	5	2	2	1	3	5
7	0	0	0	0	1	1	2	1	1	2	4
8	5	0	5	3	2	5	2	2	1	3	5
9	3	4	7	2	1	3	2	2	1	3	5
10	1	0	1	1	1	2	2	2	1	3	5
11	3	3	6	1	0	1	2	2	1	3	5
12	2	0	2	1	3	4	2	2	2	4	6
13	0	0	0	2	2	4	2	1	1	2	4
14	0	1	1	6	4	10	2	1	1	2	4
15	4	0	4	0	0	0	2	1	1	2	4
16	5	5	10	1	1	2	2	1	1	2	4
17	3	5	8	3	4	7	2	2	1	3	5
18	3	5	8	1	1	2	2	1	1	2	4
19	1	0	1	1	0	1	2	1	1	2	4
20	3	5	8	1	0	1	2	2	1	3	5
21	5	5	10	1	1	2	2	2	1	3	5
22	0	0	0	1	0	1	2	1	1	2	4
23	4	3	7	1	0	1	2	1	1	2	4
24	3	5	8	0	0	0	0	1	1	2	2
25	5	4	9	2	2	4	4	1	1	2	6
26	1	5	6	1	4	5	4	1	1	2	6
27	2	0	2	0	0	0	2	1	1	2	4
28	3	5	8	3	3	6	2	2	2	4	6
29				1	0	1	2	1	1	2	4
30	0	1	1	0	0	0	0	0	0	0	0
31	5	5	10	2	3	5	2	2	2	4	6
32	1	5	6	1	0	1	0	0	1	1	1
33	5	0	5	3	1	4	2	2	1	3	5
34	0	1	1	1	0	1	2	1	1	2	4
35	3	3	6	0	0	0	2	2	1	3	5
36	4	4	8	1	0	1	2	1	1	2	4
37	4	4	8	2	1	3	2	1	1	2	4
MODEL				6	4	10	6	2	2	4	10

EE 162 Q6

ID	H16b1	H16b2	H16c	H1Total	C6b1	C6b2	C6c	Ctotal
1			6		0	1		1
2			0		0	2		2
3			10		0	1		1
4			1		2	1		3
5			4		0	1		1
6			7		2	1		3
7			2		0	1		1
8			0		0	1		1
9			5		2	1		3
10			10		0	2		2
11			0		0	1		1
12			2		1	1		2
13			5		0	0		0
14			5		0	1		1
15			5		0	1		1
16			0		0	0		0
17			0		0	0		0
18			0		0	1		1
19			3		1	1		2
20			4		1	1		2
21			6		0	1		1
22			0		0	1		1
23			4		0	0		0
MODEL					4	2		6

OU Q7

Id	HOUq7a1	HOUQ7a2	HOUQ7b	HTOTAL	COUQ7a1	COUQ7a2	COUQ7b	CTOTAL
1	0	1	0.5	1.5	3	1	1	5
2	0.5	0	0.5	1	3	1	1	5
3	0	0	1	1	3	1	1	5
4	1.5	1	0.5	3	3	1	1	5
5	0	0.5	1	1.5	3	1	1	5
6	0.5	0.5	1	2	4	1	1	6
7	0	1	1	2	4	1	1	6
8	0.5	1	1	2.5	0	0	0	0
9	0	0	0.5	0.5	1	0	0	1
10	0.5	0.5	1	2	3	1	1	5
11	1	0	1	2	2	0	0	2
12	0.5	0.5	0.5	1.5	3	1	1	5
13	0.5	0.5	1	2	4	1	1	6
14	0.5	1	0.5	2	3	1	1	5
15	1	2	1	4	4	1	1	6
16	1	2	1	4	4	1	1	6
17	0	0	1	1	2	1	1	4
18	0	0.5	1	1.5	4	1	1	6
								0
MODEL					4	1	1	6

OU Q12

ID	h12a	h12b	h12c	h12d	hTOTAL	c12a	c12b	c12c	c12d	cTOTAL
1	6	2	3	0	11	5	1	9	2	17
2	4	0	4	1	9	4	1	10	1	16
3	5	1	5	1	12	4	2	10	2	18
4	6	0.5	5	1	12.5	5	2	9	2	18
5	8	1	6	1	16	5	1	10	2	18
6	6	0	7	1	14	4	1	10	2	17
7	8	1	7	1	17	4	1	9	2	16
8	4	0	2	0	6	5	1	10	2	18
9	7	1	6	1	15	4	1	9	2	16
10	4	1	2	1	8	4	1	10	2	17
11	6	2	7	1	16	5	1	9	2	17
12	7	1	5	2	15	5	2	10	2	19
13	6	1	4	2	13	4	1	10	2	17
14	7	1	5	2	15	5	2	10	2	19
15	7	2	4	1	14	4	1	10	2	17
16	8	2	8	2	20	5	2	10	2	19
17	6	1	2	1	10	4	2	10	2	18
18	7	2	8	2	19	5	1	9	2	17
19	6	0.5	4	1.5	12	4	1	10	2	17
20	6	1	4	0	11	5	1	10	2	18
MODEL					0	6	2	11	2	21

EE 3202 Q5

ID	H1Q5a	H1Q5b	H1Q5 Total	H2Q5a	H2Q5b	H2Q5 Total	Cq5a	Cq5b	Cq5 Total
1	2	2	4	3	4	7	0	3	3
2	5	3	8	1	3	4	0	1	1
3				7	6	13	0	1	1
4	7	4	11	12	6	18	0	3	3
5	8	2	10	14	6	20	14	1	15
6	4	0	4	4	0	4	0	0	0
7	8	0	8	7	0	7	0	0	0
8	7	1	8	7	3	10	0	0	0
9	9	0	9	6	6	12	0	0	0
10	8	3	11	6	4	10	0	0	0
11	10	3	13	14	6	20	0	0	0
12	10	4	14	14	6	20	0	3	3
13	8	3	11	10	5	15	0	3	3
14	9	3	12	2	4	6	0	1	1
15	4	3	7	8	4	12	0	3	3
16	10	0	10	14	4	18	0	0	0
17	5	2	7	10	6	16	0	3	3
18	12	4	16	14	6	20	0	1	1
19	9	5	14	14	6	20	0	1	1
20	6	3	9	12	6	18	0	1	1
21	10	4	14	14	6	20	0	3	3
22	10	4	14	14	6	20	14	0	14
23	9	4	13	14	6	20	0	0	0
24	7	3	10	10	3	13	0	0	0
25	9	2	11	12	2	14	0	0	0
26	5	2	7	10	4	14	0	3	3
27	2	2	4	4	2	6	0	3	3
MODEL				14	6	20	14	6	20

EE 3202 Q6

ID	H1a	H1b	H1Total	H2a	H2b	H2Total	Ca	Cb	CTotal
1	0	1	1	0	0	0	5	1	6
2				7	1	8	3	1	4
3	3	2	5	0	3	3	5	1	6
4	2	2	4	1	1	2	3	1	4
5	0	0	0	0	0	0	0	0	0
6	6	2	8	8	0	8	3	1	4
7	4	2	6	2	0	2	3	1	4
8	0	2	2	2	3	5	5	2	7
9	3	3	6	8	3	11	4	1	5
10	8	0	8	10	3	13	3	1	4
11	4	3	7	3	3	6	4	1	5
12	8	3	11	4	3	7	4	1	5
13	3	2	8	0	0	0	3	1	4
14	9	2	11	10	3	13	4	1	5
15	4	2	6	2	3	5	4	1	5
16	5	2	7	3	0	3	3	1	4
17	7	2	9	10	1	11	5	1	6
18	7	3	10	7	2	9	3	1	4
19	7	3	10	10	3	13	3	1	4
20	8	2	10	8	3	11	4	1	5
21	5	3	8	3	3	6	4	1	5
22	5	2	7	4	2	6	4	1	5
23	9	3	12	10	3	13	3	1	4
24	5	3	8	5	3	8	4	1	5
25	4	3	7	4	3	7	4	1	5
26	1	1	7	0	0	0	3	1	4
27	6	1	7	6	2	8	3	1	4
28	7	1	8	8	3	11	3	1	4
29	1	2	3	0	0	0	3	1	4
30	6	2	8	5	1	6	5	2	7
31	5	3	8	3	3	6	4	1	5
32	6	2	8	3	1	4	3	1	4
MODEL				10	3	13	10	3	13

Appendix I: SPSS Output

EE 162 Q1

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
H1TOTAL	68	0	10	5.15	2.234
H2TOTAL	68	0	10	2.71	2.382
CTOTAL	69	0	10	3.81	1.556
Valid N (listwise)	68				

Correlations

Correlations

		H1TOTAL	H2TOTAL	CTOTAL
H1TOTAL	Pearson Correlation	1	.704**	.594**
	Sig. (1-tailed)	.	.000	.000
	N	68	68	68
H2TOTAL	Pearson Correlation	.704**	1	.404**
	Sig. (1-tailed)	.000	.	.000
	N	68	68	68
CTOTAL	Pearson Correlation	.594**	.404**	1
	Sig. (1-tailed)	.000	.000	.
	N	68	68	69

**. Correlation is significant at the 0.01 level (1-tailed).

Nonparametric Correlations

Correlations

			H1TOTAL	H2TOTAL	CTOTAL
Spearman's rho	H1TOTAL	Correlation Coefficient	1.000	.700**	.596**
		Sig. (1-tailed)	.	.000	.000
		N	68	68	68
	H2TOTAL	Correlation Coefficient	.700**	1.000	.363**
		Sig. (1-tailed)	.000	.	.001
		N	68	68	68
	CTOTAL	Correlation Coefficient	.596**	.363**	1.000
		Sig. (1-tailed)	.000	.001	.
		N	68	68	69

**. Correlation is significant at the .01 level (1-tailed).

EE 162 Q2

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
H1TOTAL	45	0	9	1.49	2.474
H2TOTAL	46	0	10	.74	1.807
CTOTAL	46	0	10	6.09	4.934
Valid N (listwise)	45				

Correlations

Correlations

		H1TOTAL	H2TOTAL	CTOTAL
H1TOTAL	Pearson Correlation	1	.810**	.404**
	Sig. (1-tailed)	.	.000	.003
	N	45	45	45
H2TOTAL	Pearson Correlation	.810**	1	.332*
	Sig. (1-tailed)	.000	.	.012
	N	45	46	46
CTOTAL	Pearson Correlation	.404**	.332*	1
	Sig. (1-tailed)	.003	.012	.
	N	45	46	46

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

Nonparametric Correlations

Correlations

			H1TOTAL	H2TOTAL	CTOTAL
Spearman's rho	H1TOTAL	Correlation Coefficient	1.000	.740**	.376**
		Sig. (1-tailed)	.	.000	.005
		N	45	45	45
	H2TOTAL	Correlation Coefficient	.740**	1.000	.470**
		Sig. (1-tailed)	.000	.	.000
		N	45	46	46
	CTOTAL	Correlation Coefficient	.376**	.470**	1.000
		Sig. (1-tailed)	.005	.000	.
		N	45	46	46

**. Correlation is significant at the .01 level (1-tailed).

EE 162 Q3

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HTOTAL	62	0	10	6.84	3.384
CTOTAL	63	0	5	3.57	1.573
Valid N (listwise)	62				

Correlations

Correlations

		HTOTAL	CTOTAL
HTOTAL	Pearson Correlation	1	.008
	Sig. (1-tailed)	.	.477
	N	62	62
CTOTAL	Pearson Correlation	.008	1
	Sig. (1-tailed)	.477	.
	N	62	63

Nonparametric Correlations

Correlations

			HTOTAL	CTOTAL
Spearman's rho	HTOTAL	Correlation Coefficient	1.000	-.068
		Sig. (1-tailed)	.	.299
		N	62	62
	CTOTAL	Correlation Coefficient	-.068	1.000
		Sig. (1-tailed)	.299	.
		N	62	63

EE 162 Q4

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
H14C	51	0	10	5.22	2.625
CTOTAL	52	0	16	6.02	2.601
Valid N (listwise)	51				

Correlations

Correlations

		H14C	CTOTAL
H14C	Pearson Correlation	1	.238*
	Sig. (1-tailed)	.	.046
	N	51	51
CTOTAL	Pearson Correlation	.238*	1
	Sig. (1-tailed)	.046	.
	N	51	52

*. Correlation is significant at the 0.05 level (1-tailed).

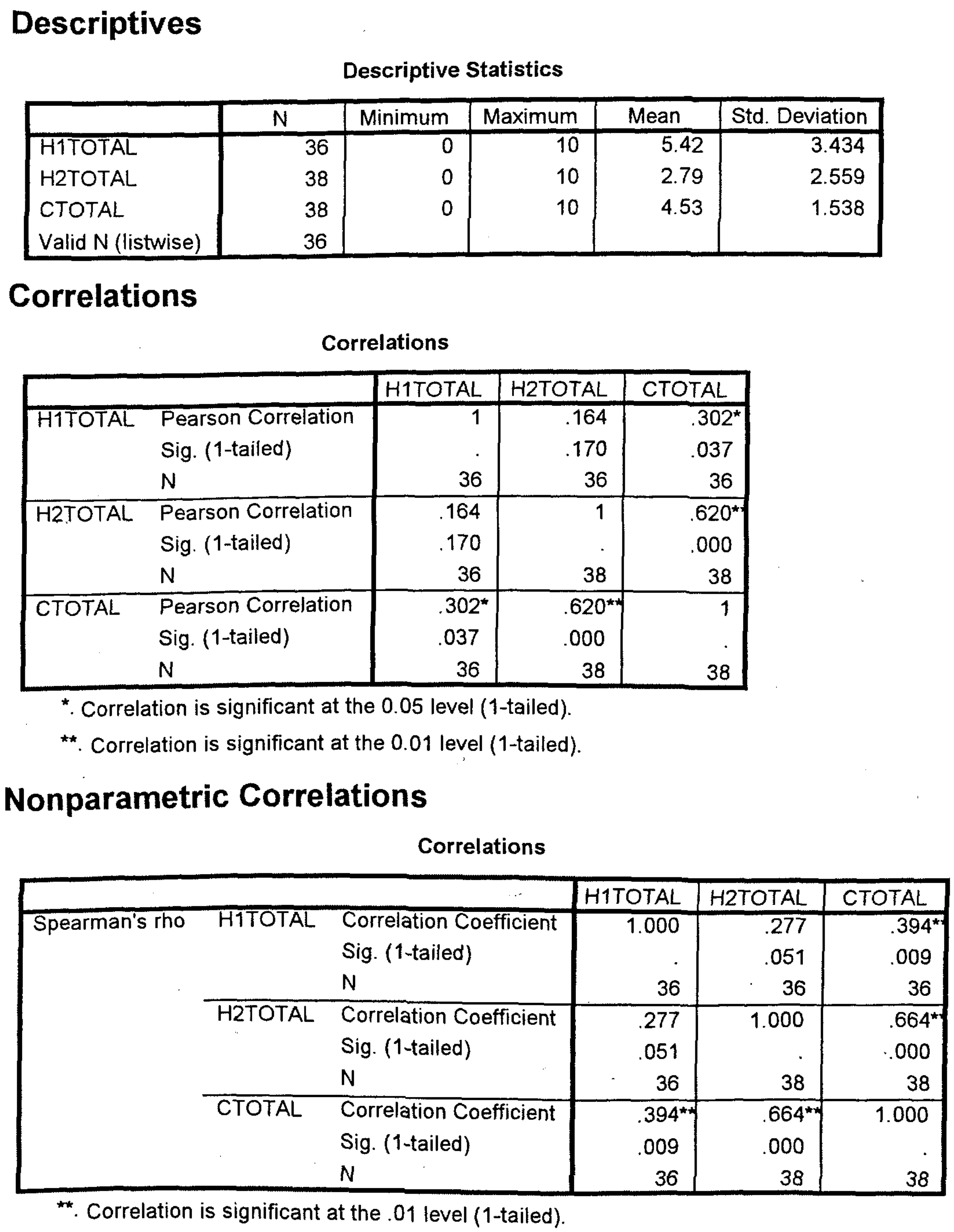
Nonparametric Correlations

Correlations

			H14C	CTOTAL
Spearman's rho	H14C	Correlation Coefficient	1.000	.336*
		Sig. (1-tailed)	.	.008
		N	51	51
	CTOTAL	Correlation Coefficient	.336**	1.000
		Sig. (1-tailed)	.008	.
		N	51	52

**. Correlation is significant at the .01 level (1-tailed).

EE 162 Q5



Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
H1TOTAL	36	0	10	5.42	3.434
H2TOTAL	38	0	10	2.79	2.559
CTOTAL	38	0	10	4.53	1.538
Valid N (listwise)	36				

Correlations

Correlations

		H1TOTAL	H2TOTAL	CTOTAL
H1TOTAL	Pearson Correlation	1	.164	.302*
	Sig. (1-tailed)	.	.170	.037
	N	36	36	36
H2TOTAL	Pearson Correlation	.164	1	.620**
	Sig. (1-tailed)	.170	.	.000
	N	36	38	38
CTOTAL	Pearson Correlation	.302*	.620**	1
	Sig. (1-tailed)	.037	.000	.
	N	36	38	38

*. Correlation is significant at the 0.05 level (1-tailed).

**. Correlation is significant at the 0.01 level (1-tailed).

Nonparametric Correlations

Correlations

			H1TOTAL	H2TOTAL	CTOTAL
Spearman's rho	H1TOTAL	Correlation Coefficient	1.000	.277	.394**
		Sig. (1-tailed)	.	.051	.009
		N	36	36	36
	H2TOTAL	Correlation Coefficient	.277	1.000	.664**
		Sig. (1-tailed)	.051	.	.000
		N	36	38	38
	CTOTAL	Correlation Coefficient	.394**	.664**	1.000
		Sig. (1-tailed)	.009	.000	.
		N	36	38	38

**. Correlation is significant at the .01 level (1-tailed).

EE 162 Q6

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
H16C	23	0	10	3.43	3.131
CTOTAL	24	0	6	1.50	1.319
Valid N (listwise)	23				

Correlations

Correlations

		H16C	CTOTAL
H16C	Pearson Correlation	1	.187
	Sig. (1-tailed)	.	.196
	N	23	23
CTOTAL	Pearson Correlation	.187	1
	Sig. (1-tailed)	.196	.
	N	23	24

Nonparametric Correlations

Correlations

			H16C	CTOTAL
Spearman's rho	H16C	Correlation Coefficient	1.000	.207
		Sig. (1-tailed)	.	.171
		N	23	23
	CTOTAL	Correlation Coefficient	.207	1.000
		Sig. (1-tailed)	.171	.
		N	23	24

OU Q7

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HTOTAL	18	1	4	1.94	.953
CTOTAL	19	0	6	4.68	1.765
Valid N (listwise)	18				

Correlations

Correlations

		HTOTAL	CTOTAL
HTOTAL	Pearson Correlation	1	.263
	Sig. (1-tailed)	.	.146
	N	18	18
CTOTAL	Pearson Correlation	.263	1
	Sig. (1-tailed)	.146	.
	N	18	19

Nonparametric Correlations

Correlations

			HTOTAL	CTOTAL
Spearman's rho	HTOTAL	Correlation Coefficient	1.000	.374
		Sig. (1-tailed)	.	.063
		N	18	18
	CTOTAL	Correlation Coefficient	.374	1.000
		Sig. (1-tailed)	.063	.
		N	18	19

OU Q12

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
HTOTAL	21	0	20	12.64	4.520
CTOTAL	21	16	21	17.62	1.203
Valid N (listwise)	21				

Correlations

Correlations

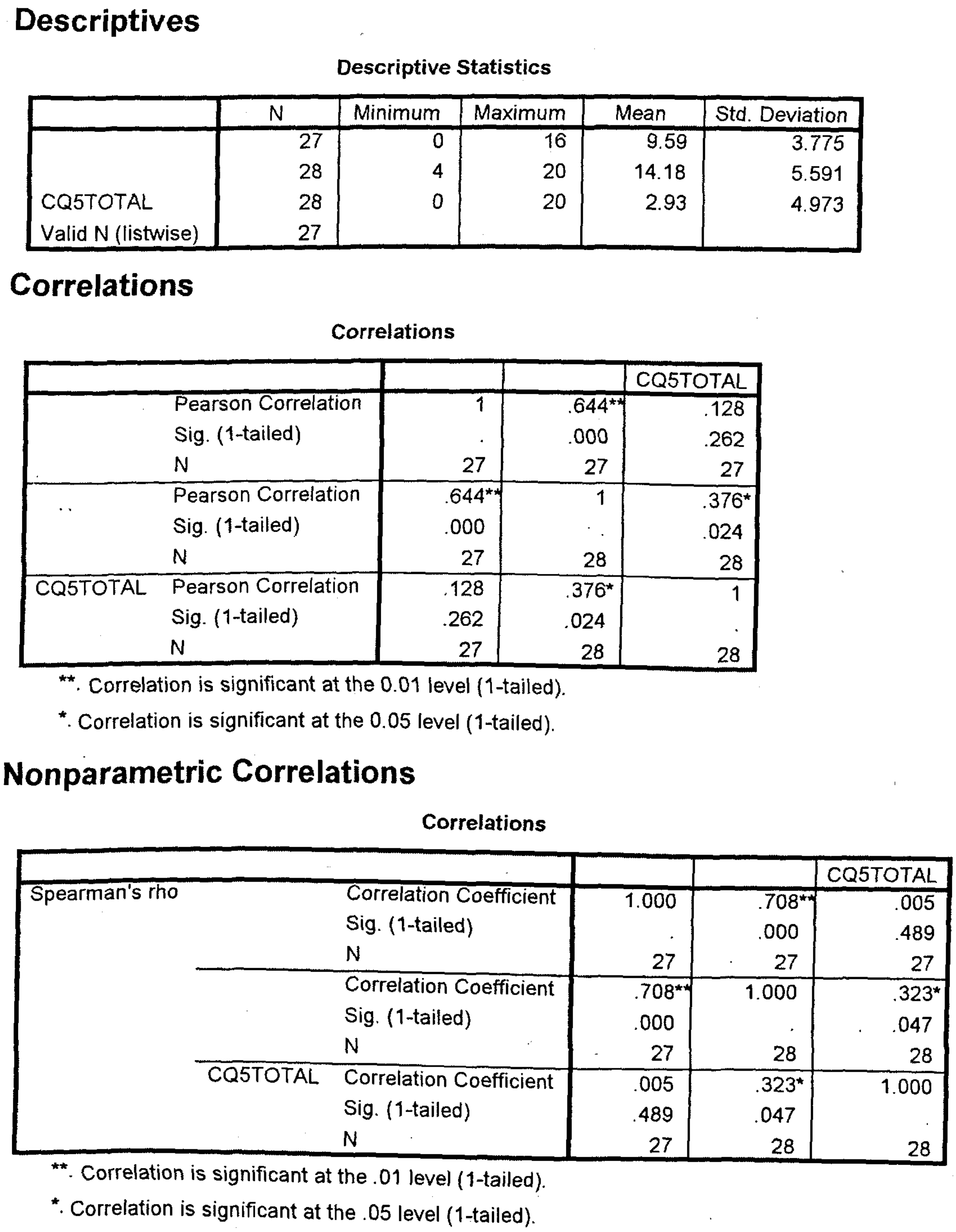
		HTOTAL	CTOTAL
HTOTAL	Pearson Correlation	1	-.348
	Sig. (1-tailed)	.	.061
	N	21	21
CTOTAL	Pearson Correlation	-.348	1
	Sig. (1-tailed)	.061	.
	N	21	21

Nonparametric Correlations

Correlations

			HTOTAL	CTOTAL
Spearman's rho	HTOTAL	Correlation Coefficient	1.000	-.097
		Sig. (1-tailed)	.	.339
		N	21	21
	CTOTAL	Correlation Coefficient	-.097	1.000
		Sig. (1-tailed)	.339	.
		N	21	21

EE 3202 Q5



Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
	27	0	16	9.59	3.775
	28	4	20	14.18	5.591
CQ5TOTAL	28	0	20	2.93	4.973
Valid N (listwise)	27				

Correlations

Correlations

				CQ5TOTAL
	Pearson Correlation	1	.644**	.128
	Sig. (1-tailed)	.	.000	.262
	N	27	27	27
	Pearson Correlation	.644**	1	.376*
	Sig. (1-tailed)	.000	.	.024
	N	27	28	28
CQ5TOTAL	Pearson Correlation	.128	.376*	1
	Sig. (1-tailed)	.262	.024	.
	N	27	28	28

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

Nonparametric Correlations

Correlations

					CQ5TOTAL
Spearman's rho		Correlation Coefficient	1.000	.708**	.005
		Sig. (1-tailed)	.	.000	.489
		N	27	27	27
		Correlation Coefficient	.708**	1.000	.323*
		Sig. (1-tailed)	.000	.	.047
		N	27	28	28
	CQ5TOTAL	Correlation Coefficient	.005	.323*	1.000
		Sig. (1-tailed)	.489	.047	.
		N	27	28	28

**. Correlation is significant at the .01 level (1-tailed).

*. Correlation is significant at the .05 level (1-tailed).

EE 3202 Q6

Descriptives

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
H1Q6Total	31	0	12	6.94	2.966
H2Q6Total	33	0	13	6.61	4.315
CQ6TOTAL	33	0	13	4.85	1.889
Valid N (listwise)	31				

Correlations

Correlations

		H1Q6Total	H2Q6Total	CQ6TOTAL
H1Q6Total	Pearson Correlation	1	.749**	.157
	Sig. (1-tailed)	.	.000	.199
	N	31	31	31
H2Q6Total	Pearson Correlation	.749**	1	.303*
	Sig. (1-tailed)	.000	.	.043
	N	31	33	33
CQ6TOTAL	Pearson Correlation	.157	.303*	1
	Sig. (1-tailed)	.199	.043	.
	N	31	33	33

**. Correlation is significant at the 0.01 level (1-tailed).

*. Correlation is significant at the 0.05 level (1-tailed).

Nonparametric Correlations

Correlations

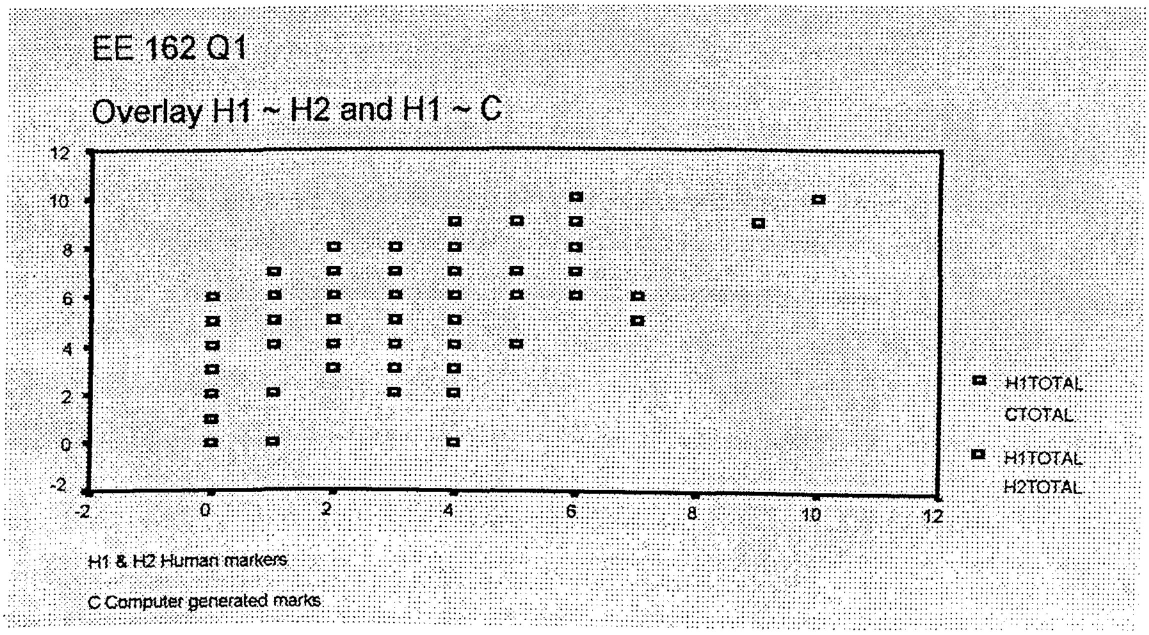
			H1Q6Total	H2Q6Total	CQ6TOTAL
Spearman's rho	H1Q6Total	Correlation Coefficient	1.000	.749**	-.004
		Sig. (1-tailed)	.	.000	.491
		N	31	31	31
	H2Q6Total	Correlation Coefficient	.749**	1.000	.134
		Sig. (1-tailed)	.000	.	.229
		N	31	33	33
	CQ6TOTAL	Correlation Coefficient	-.004	.134	1.000
		Sig. (1-tailed)	.491	.229	.
		N	31	33	33

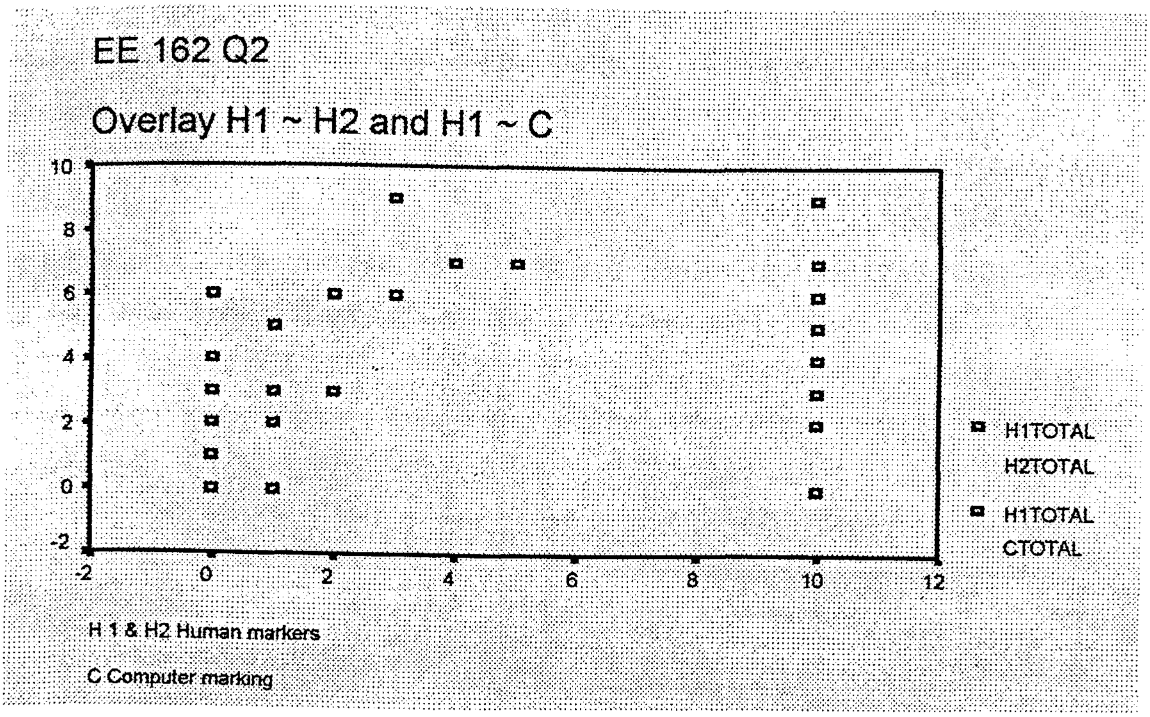
**. Correlation is significant at the .01 level (1-tailed).

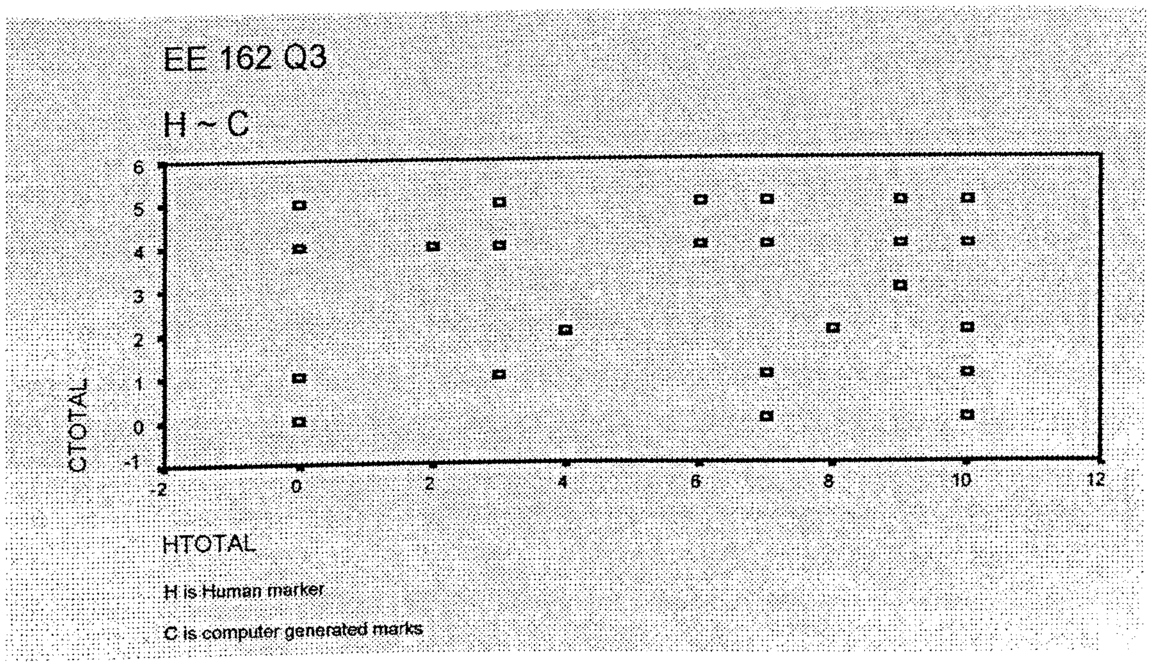
Appendix J: Scatter Plots

This Appendix is to inform the reader of the relationship between the First Marker (H1) and Computer Marking (C). Where there is Second Marker (H2) available then that has been included as well. The X-axis is always the marks as awarded by the First Marker, with the marks awarded by the Computer Marking (and Second Marker) on the Y-axis.

For a "perfect result" the Scatter Plot should appear as a straight line at 45° since each marker would be awarding the same mark as the other marker(s). Therefore the bigger the spreads from straight line then the poorer the agreement between the markers is.

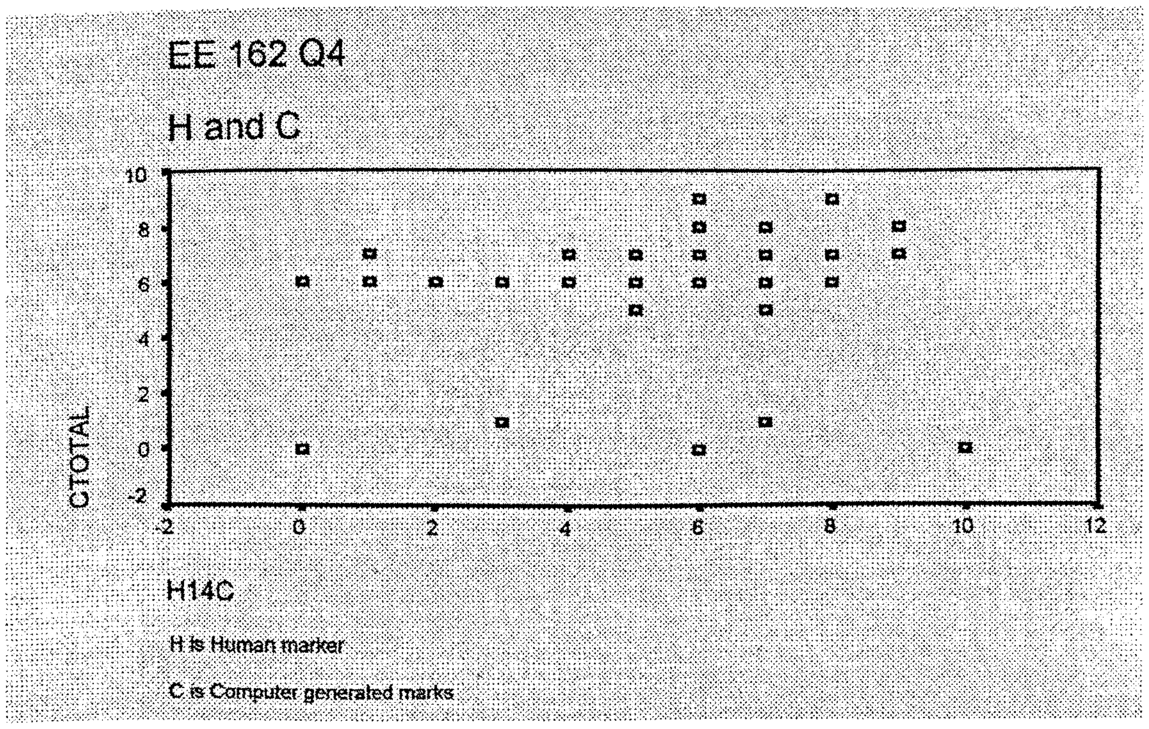






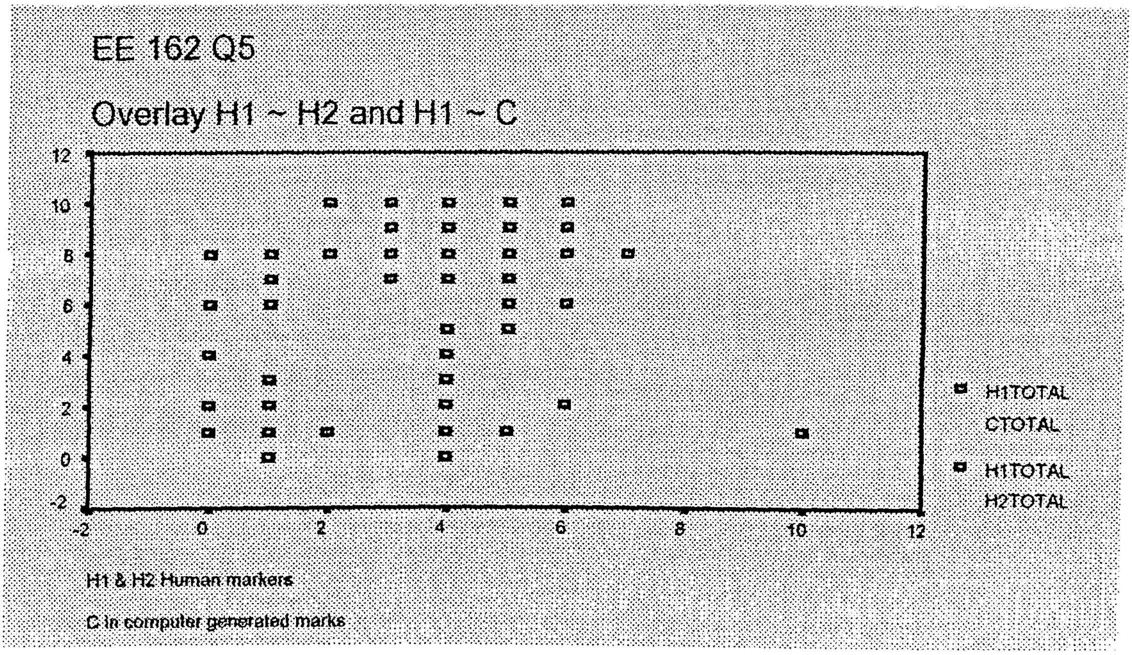
H is Human marker

C is computer generated marks



H is Human marker

C is Computer generated marks

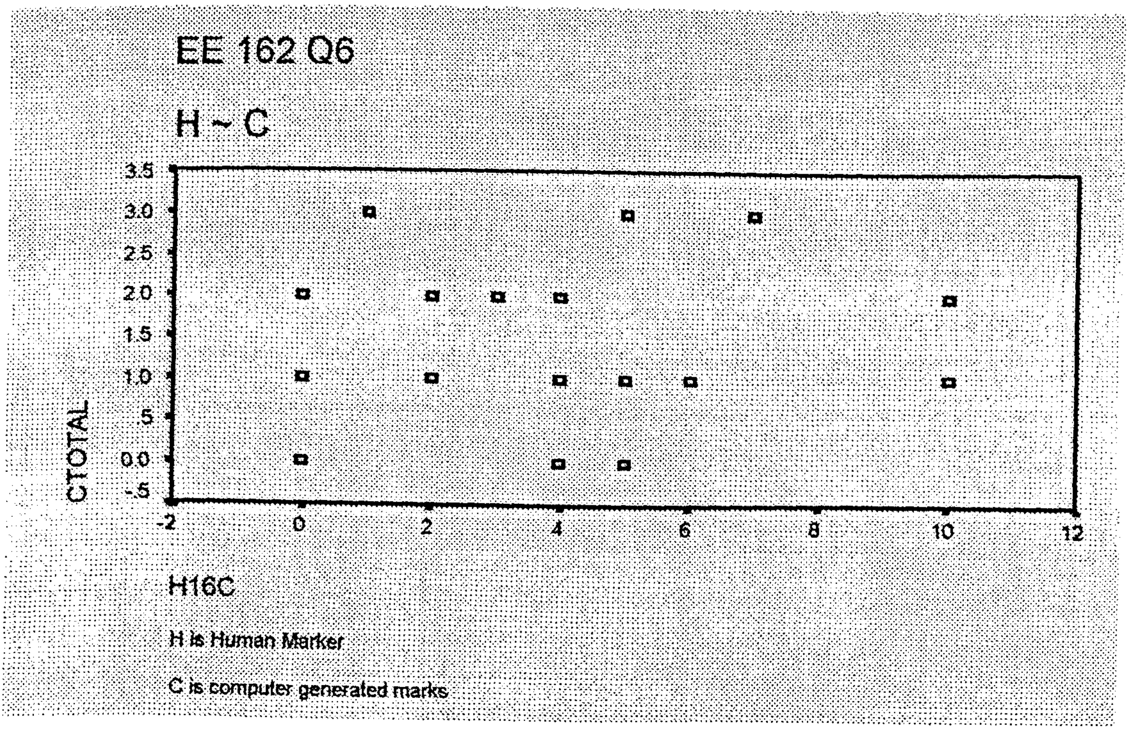


EE 162 Q5

Overlay H1 ~ H2 and H1 ~ C

H1 & H2 Human markers

C in computer generated marks

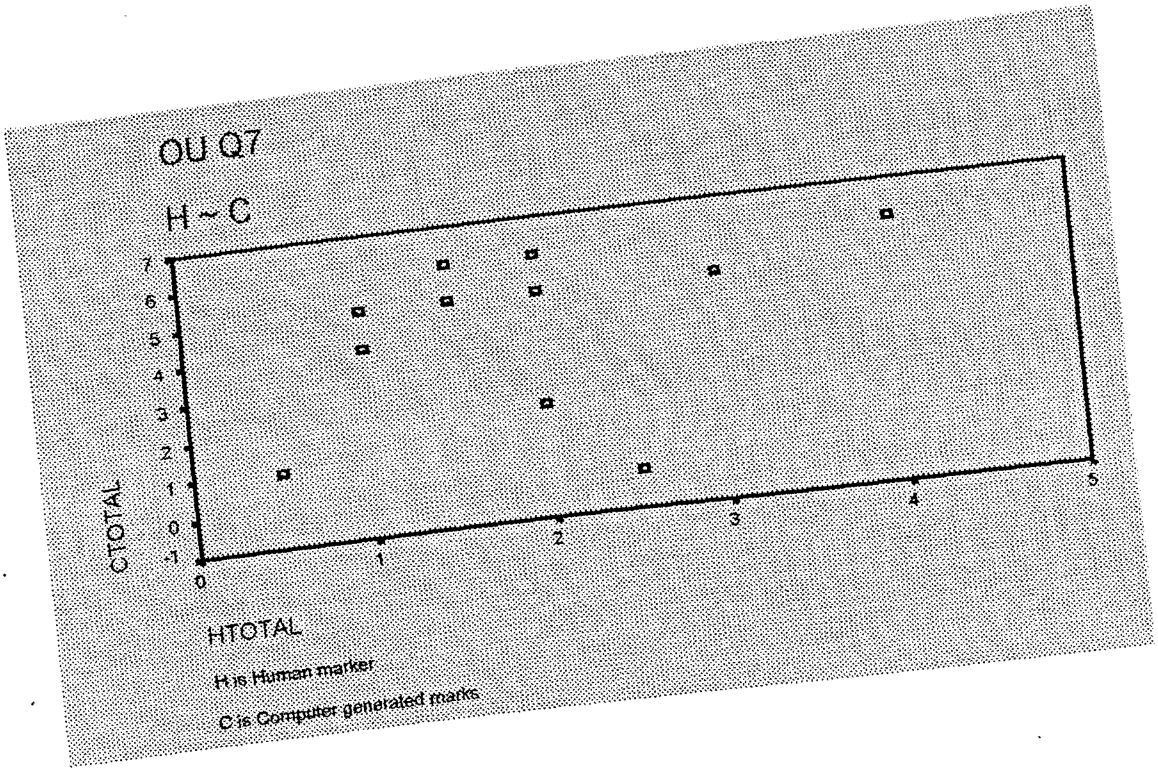


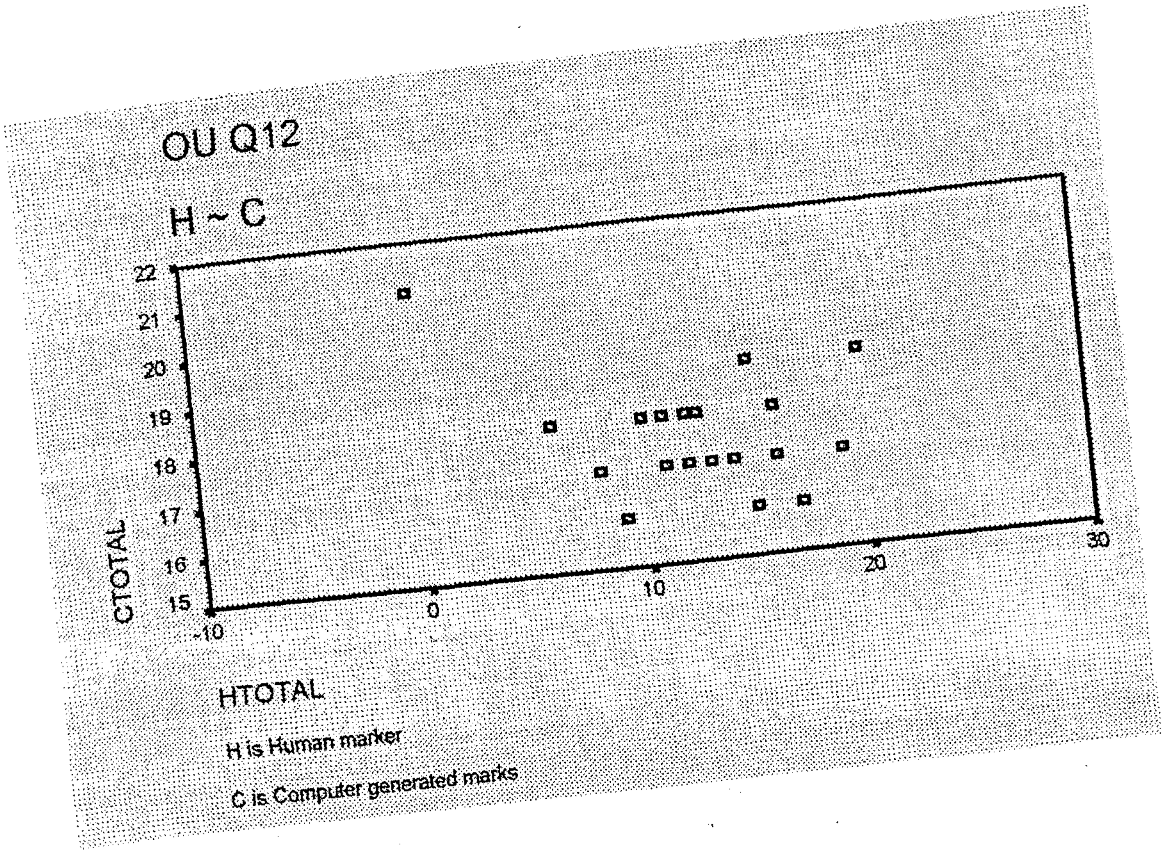
EE 162 Q6

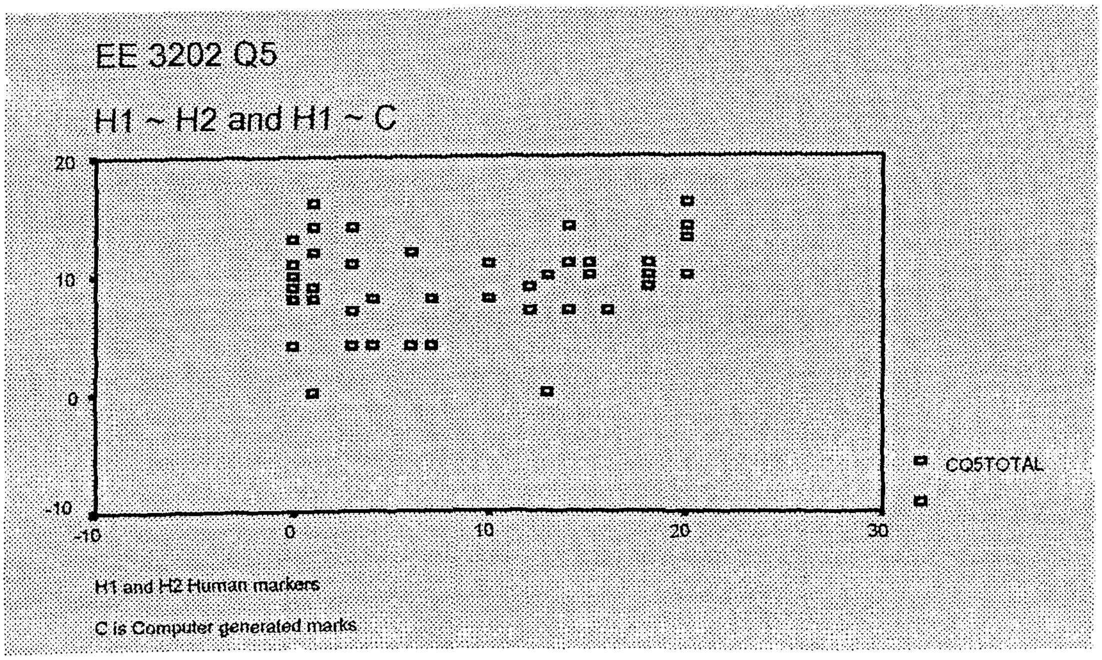
H ~ C

H is Human Marker

C is computer generated marks


OU Q7
H ~ C
7
6
5
4
3
2
1
0
-1
CTOTAL
0
1
2
3
4
5
HTOTAL
H is Human marker
C is Computer generated marks


OU Q12
H ~ C
22
21
20
19
18
17
16
15
CTOTAL
-10
0
10
20
30
HTOTAL
H is Human marker
C is Computer generated marks

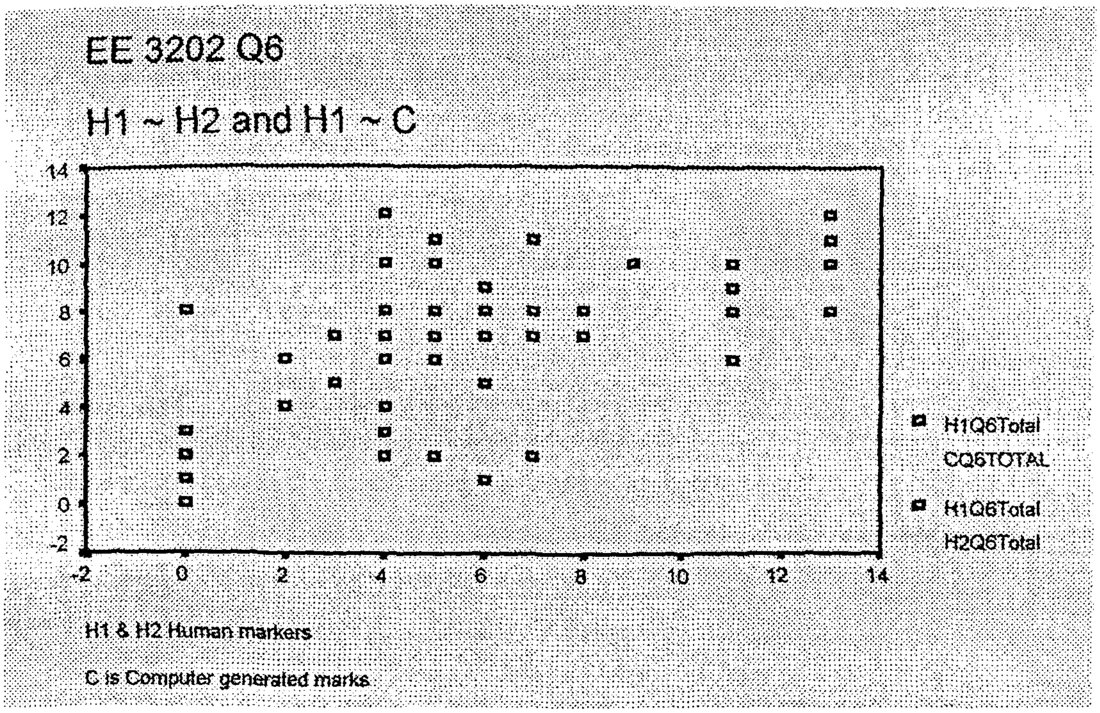


EE 3202 Q5

H1 ~ H2 and H1 ~ C

H1 and H2 Human markers

C is Computer generated marks



EE 3202 Q6

H1 ~ H2 and H1 ~ C

H1 & H2 Human markers

C is Computer generated marks

Appendix K: Description of the SEAR System & Software

Programming Language Selection

This author is familiar with a number of computer programming languages, namely ~ COBOL, Borland C/C++, Pascal / Delphi, Basic / Visual Basic. From these languages the author had to select the language that would confer the biggest range of functionality for the perceived requirements of automated assessment software. The author has kept the interface deliberately very simple, basically the interface is a console application ~ any future development of SEAR would probably have a GUI interface.

The programming language selected was Borland C/C++ as this it was felt the best language covering file handling, string [i.e. text], ASCII and non-ASCII characters, and so on that is required to develop the SEAR software. COBOL was used to sort wordlist and other similar processes when that was the most appropriate language to use.

Marking System

The system is called **SEAR**, which is an acronym for **S**chema, **E**xtract, **A**ssess and **R**eport. These four letters serve as reminders of the four main stages of assessing students in general, namely:

There must exist a **schema** to detail what is to be assessed, where the awarding of marks will be given.

On examining the student's submission the examiner must **extract** items from the submission that are worth assessing.

These extracted items are then **assess**ed against the schema for the awarding of marks.

Finally the **report**ing, or recording, of the marks awarded, at the level of the individual submission and at the cohort level, completes the marking exercise.

A fuller Assessment Life Cycle is to be found in BSI 7988 page 3.

The system has been designed to emulate this particular style of assessment, since any automated marking system has to follow a similar manual process to the examiner to be operationally acceptable.

The current software consists of a suite of programmes and other magnetic mediated files.

The software provides the storage of content marking schema, word-processed essays, extracts from these essays, assessment and reporting. Software to remark essays is also included.

The associated directory structure is:
SEAR software
Option files, one for each essay set
Wordlists
Each essay set, a folder consisting of:

Submit directory
Extract directory
Extract/Sample directory
Report directory

The **submit** directory is used as a receiving directory for word-processed essays. Essays submitted via CD ROM, or other computer files are copied into the submit directory assigned to that essay set. Essays submitted as e-mail attachments will require to be manually manipulated into their submit directory. Late essay submission(s) will be handled in the same manner as above. As essays are extracted a record will be kept.

The **extract** directory is used to store the extract taken from each of the essays in the submit directory. Late essay submissions will be extracted into the same extract directory. The same essay extracts will be used for both marking style and for marking content. Records will be of what marking has been done on these extracted files.

The **extract/sample** directory is specifically for marking style. Marking style requires a sample of essays to be manually marked as a basis for developing the weighted linear model to be used to mark the complete essay set. Keeping the sample essays physically separate from the complete essay set will ensure statistical validity.

The **report** directory is used to store the various outputs generated by the two marking processes.

System operation

Examiner(s) seeking to use SEAR will first make contact [email, person] to establish the following information ~

Their contact details,
The START date for receiving essays – assessment issue,
The FINISH date for receiving essays – the deadline for submission,
Number of essays expected,
Mode of marking – style only, content only, or both,
For content – supply both a marking scheme, and a model answer,
For style – supply a fixed size sample of human marking essays.
The essay set identification, if e-mail submission is required.

Essayist(s) who will be using e-mail submission of their essays will be given:

The dedicated email address to send their essays to
[currently set to: essays.christie@btinternet.com]
The essay set identification to be used as the subject in the email.

Essays can be alternatively submitted for SEAR by magnetic media such as various disks and tapes. CD-ROMs may be used as media for submitting essays.

Before marking [style]

The examiner must supply a marked sample of essays taken from the essay set. The sample should be chosen to reflect the range of essays present in the essay set. Picking a biased set, say, all the good essays, will lead to incorrectly processing. The size of the sample has to be at least twice the number of metrics expected used in the marking of style. This requirement is to ensure statistical validity of the final marking process. The processing of this sample is to generate a weighted linear model for the metrics.

Before marking [content]

A computer file [schema] will be generated from the examiner's marking schema to be used as the basis for content marking. To confirm that this generation has been achieved correctly then the model answer, as supplied by the examiner, is processed. The result of this processing should highlight any problems occurring in the computerised version of the original schema.

Marking process

Every essay has to undergo a two-stage process to be marked.

The first stage is to extract the essay text from the word-processed file. Certain preparation of the text occurs at this stage to facilitate the next stage. The stage is only required once per essay set regardless of which, or both, second stages are required.

The second stage is to perform the actual marking.
For style the second stage is the application of the weighted linear model generated from the manually marked essays to the whole of the essay set, essay by essay.

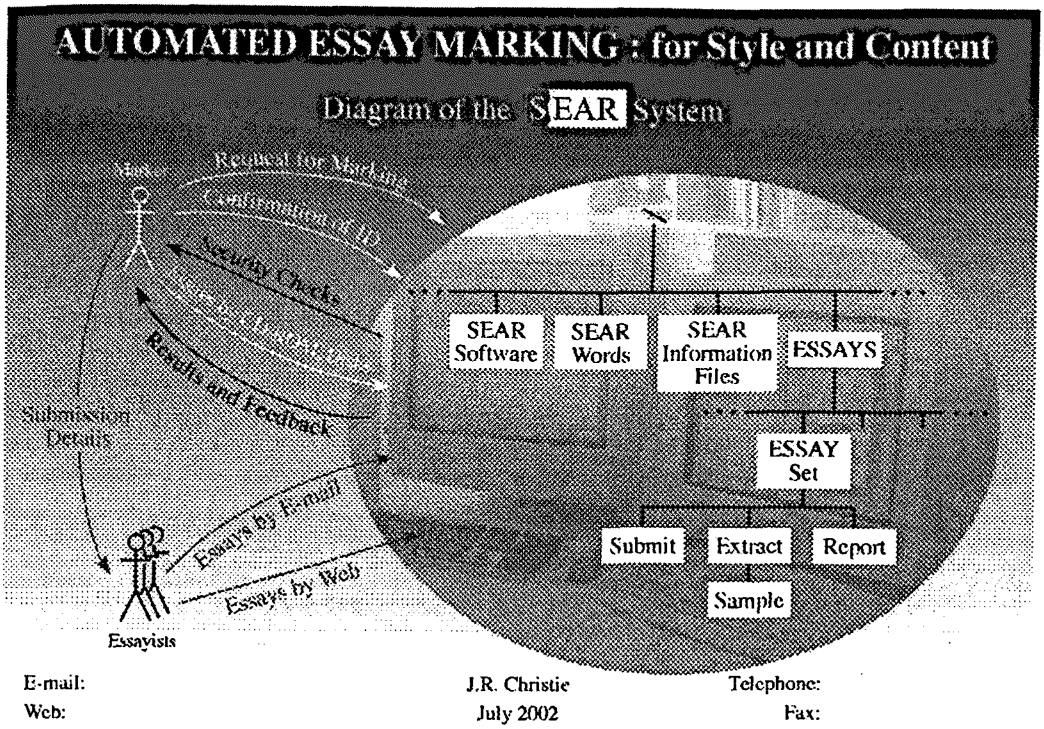
For content the second stage is the application of the confirmed computer schema against each essay in turn.

Any examiner requiring essay set(s) to be marked for both style and content will necessitate both the second stages to be performed.

Reporting Process

After marking, the results will be sent to the examiners, together with other marking details, cohort analysis, etc.

Diagram of the Current System



AUTOMATED ESSAY MARKING : for Style and Content

Diagram of the SEAR System

E-mail:
Web:

J.R. Christie
July 2002

Telephone:
Fax: