Automatic essay scoring for low level learners of English as a second language.

Mellor, Andrew

Automatic Essay Scoring for Low Level Learners of English as a Second Language

Andrew Mellor

Submitted to the University of Wales
in fulfillment of the requirements for the
Degree of Doctor of Philosophy

Swansea University

2010

Abstract

This thesis investigates the automatic assessment of essays written by Japanese low level learners of English as a second language. A number of essay features are investigated for their ability to predict human assessments of quality. These features include unique lexical signatures (Meara, Jacobs & Rodgers, 2002), distinctiveness, essay length, various measures of lexical diversity, mean sentence length and some properties of word distributions. Findings suggest that no one feature is sufficient to account for essay quality but essay length is a strong predictor for low level learners in time constrained tasks. Combinations of several features are much more powerful in predicting quality than single features. Some simple systems incorporating some of these features are also considered. One is a two-dimensional 'quantity/content' model based on essay length and lexical diversity. Various measures of lexical diversity are used for the content dimension. Another system considered is a clustering algorithm based on various lexical features. A third system is a Bayesian algorithm which classifies essays according to semantic content. Finally, an alternative process based on capture-recapture analysis is also considered for special cases of assessment. One interesting finding is that although many essay features only have moderate associations with quality, extreme values at both ends of the scale are often very reliable indicators of high quality or poor quality essays. These easily identifiable high quality or low quality essays can act as training samples for classification algorithms such as Bayesian classifiers. The clustering algorithm used in this study correlated particularly strongly with human essay ratings. This suggests that multivariate statistical methods may help realise more accurate essay prediction.

Table of contents

Chapter One: Introduction 1

1.1 Aims of the thesis 1
1.2 Background 1
1.3 Identifying features 2
1.4 Developing assessment systems 3
1.5 Influence of other fields 4
1.6 Automation 4
1.7 Outline of the thesis 5

Chapter Two: Literature Review 8

2.1 Introduction 8
2.2 Page (1994) 9
2.2.1 Summary 9
2.2.2 Comments 10
2.3 Chodorow & Burstein (2004) 12
2.3.1 Summary 12
2.3.2 Comments 15
2.4 Landauer, Laham, & Foltz (2003) 18
2.4.1 Summary 18
2.4.2 Comments 20
2.5 Larsen-Freeman & Strom (1977) 22
2.5.1 Summary 22
2.5.2 Comments 24
2.6 Arnaud (1984) 26
2.6.1 Summary 26
2.6.2 Comments 28
2.7 Engber (1995) 30
2.7.1 Summary 30
2.7.2 Comments 32
2.8 Ferris (1994) 34
2.8.1 Summary 34
2.8.2 Comments 36
2.9 Laufer & Nation (1995) 38
2.9.1 Summary 38
2.9.2 Comments 41
2.10 Meara & Bell (2001) 43
2.10.1 Summary 43
2.10.2 Comments 45
2.11 Meara, Jacobs & Rodgers (2002) 47
2.11.1 Summary 47
2.11.2 Comments 49
2.12 Discussion 50
2.12.1 Features 51
2.12.1.1 Essay length 51
2.12.1.2 Lexical diversity 52
2.12.1.3 Other considerations 56
2.12.2 Automatic scoring methods 57
2.12.3 Measures from authorship attribution 58
2.13 Conclusion 59

Chapter Three: Lexical Signatures 62

3.1 Introduction 62
3.2 Lexical signatures in low level learners 62

3.2.1 Introduction 62
3.2.1.1 Aims of the experiment 62
3.2.2 Methodology 63
3.2.2.1 Participants and task 63
3.2.2.2 Measuring lexical signatures 63
3.2.2.3 Constructing an index of uniqueness 65
3.2.2.4 Measurement reliability 65
3.2.3 Results 66
3.2.3.1 Lexical signatures 66
3.2.3.2 Index of uniqueness 66
3.2.3.3 Analysis of measurement reliability 67
3.2.4 Discussion 67
3.2.5 Conclusion 68
3.3 Investigating an index of uniqueness 68
3.3.1 Introduction 68
3.3.1.1 Aims of the experiment 69
3.3.2 Methodology 69
3.3.2.1 Participants and tasks 69
3.3.2.2 Sampling of subsets 70
3.3.2.3 Quality assessments of essays 71
3.3.2.4 Lexical signature analysis 72
3.3.2.5 Index of uniqueness 72
3.3.3 Results 72
3.3.3.1 General test characteristics 72
3.3.3.2 Measurement reliability 73
3.3.3.3 Reliability over two tasks 73
3.3.3.4 Uniqueness index scores and quality assessments 74
3.3.4 Discussion 75
3.3.4.1 Problems of reliability 75
3.3.4.2 Considerations of word selection 76
3.3.5 Conclusion 78
3.4 Conclusion 78

Chapter Four: Distinctiveness 79

4.1 Introduction 79
4.2 Distinctiveness in L2 learner texts 80
4.2.1 Introduction 80
4.2.1.1 Aims of the experiment 80
4.2.2 Methodology 80
4.2.2.1 Participants and tasks 80
4.2.2.2 Calculating distinctiveness 80
4.2.3 Results 81
4.2.3.1 Distinctiveness scores 81
4.2.3.2 Eliminating inconsistent students 81
4.2.3.3 Distinctiveness and quality assessments 82
4.2.4 Discussion 85
4.2.5 Conclusion 86
4.3 Testing reliability of a measure of distinctiveness on identical tasks 86
4.3.1 Introduction 86
4.3.1.1 Aims of the experiment 86
4.3.2 Methodology 86
4.3.2.1 Participants and tasks 86
4.3.2.2 Details of the analysis 86
4.3.3 Results 87
4.3.3.1 Distinctiveness scores 87
4.3.3.2 Consistency of learners 88
4.3.4 Discussion 88

4.3.5 Conclusion 89
4.4 Discussion 89
4.4.1 A comparison of results of the two experiments 89
4.4.2 Threats to reliability and validity 90
4.4.3 Reward structure 92
4.5 Conclusion 93

Chapter Five: Exploratory Studies 95

5.1 Introduction 95
5.2 Lexical features in L2 texts of varying proficiency 95
5.2.1 Introduction 95
5.2.1.1 Aims of the experiment 96
5.2.2 Methodology 96
5.2.2.1 Participants and tasks 96
5.2.2.2 Analysis 97
5.2.3 Results 98
5.2.3.1 Mean word length and mean sentence length 98
5.2.3.2 Entropy, Yule's K and the D estimate 98
5.2.4 Discussion 101
5.2.5 Conclusion 103
5.3 Word distribution properties of L1 and L2 texts 103
5.3.1 Introduction 103
5.3.1.1 Aims of the experiment 103
5.3.2 Methodology 104
5.3.2.1 Participants and tasks 104
5.3.2.2 Analysis 104
5.3.3 Results 105
5.3.3.1 Word frequency distribution 105
5.3.3.2 Word rank distributions 107
5.3.3.3 Word length distributions 109
5.3.4 Discussion 110
5.3.5 Conclusion 111
5.4 The type-token ratio and lexical variation 111
5.4.1 Introduction 111
5.4.1.1 Aims of the experiment 112
5.4.2 Methodology 112
5.4.2.1 Participants and tasks 112
5.4.2.2 Analysis 112
5.4.3 Results 112
5.4.3.1 Reliability over two tasks 112
5.4.3.2 Relationship to quality assessments 113
5.4.4 Discussion 114
5.4.5 Conclusion 115
5.5 LFP and P_Lex 115
5.5.1 Introduction 115
5.5.1.1 Aims of the experiment 116
5.5.2 Methodology 116
5.5.2.1 Participants and tasks 116
5.5.2.2 Analysis 116
5.5.3 Results 117
5.5.3.1 LFP 117
5.5.3.2 P_Lex 118
5.5.4 Discussion 118
5.5.5 Conclusion 119
5.6 Conclusion 120

Chapter Six: A quantity/content model 122

6.1 Introduction 122
6.1.1 Aims of the experiment 122
6.2 Methodology 122
6.2.1 Participants and tasks 122
6.2.2 Essay ratings 122
6.2.3 Analysis 123
6.2.3.1 Selection of quantity and content measures 123
6.2.3.2 Sampled features 124
6.2.3.3 Calculation of quantity and content dimensions 125
6.2.3.4 Comparison of content variables 125
6.3 Results 126
6.3.1 Basic characteristics 126
6.3.2 Quantity/content graphs 127
6.3.2.1 Sample word types TTR(100) 127
6.3.2.2 Guiraud Index 128
6.3.2.3 Yule's K 128
6.3.2.4 Hapax(100) 129
6.3.2.5 The D estimate 130
6.3.2.6 Advanced Guiraud 130
6.3.3 Correlation analysis 131
6.3.4 Partial correlation analysis 132
6.3.5 Multiple regression analysis 133
6.3.6 Results conclusion 133
6.4 Discussion 134
6.4.1 Overview 134
6.4.2 Advanced Guiraud 134
6.4.3 Classifying essays 135
6.4.4 Linearity 135
6.5 Conclusion 136

Chapter Seven: Automatic scoring algorithms 138

7.1 Introduction 138
7.2 A clustering algorithm 138
7.2.1 Introduction 138
7.2.1.1 Aims of the experiment 138
7.2.2 Methodology 138
7.2.2.1 Participants and tasks 138
7.2.2.2 Holistic assessments 138
7.2.2.3 Analysis 139
7.2.2.3.1 Selection of features 139
7.2.2.3.2 Calculation of features 141
7.2.2.3.3 Principal component analysis 142
7.2.2.3.4 Comparison of features with principal components 143
7.2.2.3.5 Clustering 144
7.2.3 Results 145
7.2.4 Discussion 146
7.2.4.1 Reliability 146
7.2.4.2 Clustering 147
7.2.5 Conclusion 148
7.3 A Bayesian algorithm 148
7.3.1 Introduction 148
7.3.1.1 Aims of the experiment 150
7.3.2 Methodology 150
7.3.2.1 Participants and tasks 150
7.3.2.2 Analysis 150
7.3.2.2.1 Selecting the training set 151

7.3.2.2.2 Bayesian classification 152
7.3.3 Results 153
7.3.3.1 Bayesian algorithm 153
7.3.3.2 Reliability of training essays 153
7.3.4 Discussion 154
7.3.4.1 Comparison of algorithm ratings with pooled hunan ratings 154
7.3.4.2 Comparison of Bayesian classifier with pooled human ratings 154
7.3.4.3 Adjustments to the algorithm 155
7.3.5 Conclusion 156
7.4 Discussion 156
7.4.1 Comparison between clustering and Bayesian algo:ithms 156
7.4.2 Inter-algorithm reliability 158
7.4.3 Essay length 158
7.4.4 Selecting an algorithm 160
7.5 Conclusion 161

Chapter Eight: Capture-recapture analysis 163

8.1 Introduction 163
8.1.1 Background to capture-recapture analysis 163
8.1.2 Application of capture-recapture analysis to L2 lexical studies 164
8.1.3 Aims of the experiment 166
8.2 Methodology 166
8.2.1 Participants and tasks 166
8.2.2 Analysis 166
8.3 Results 167
8.4 Discussion 169
8.4.1 Sampling reliability 169
8.4.2 Minimum essay length 171
8.4.3 Using capture-recapture analysis for assessment 173
8.5 Conclusion 173

Chapter Nine: Discussion 175

9.1 Introduction 175
9.2 Lexical features 175
9.2.1 Properties of lexical features for assessment purposes 175
9.2.1.1 Identifying quality differences in essays of low level learners 175
9.2.1.2 Reliability across tasks 176
9.2.1.3 Automatic calculation 177
9.2.1.4 Validity 178
9.2.2 Length of essay 180
9.3 L2 vocabulary research 182
9.4 Automating the process 185
9.5 Limitations of the present study 187
9.6 Future research 190
9.6.1 Choice of system 191
9.6.2 Selection of variables 192
9.7 Conclusion 194

Chapter Ten: Conclusion 196

10.1 Introduction 196
10.2 Uniqueness index 196
10.3 Distinctiveness 197
10.4 Other features 198
10.5 Essay length and lexical diversity 199
10.6 Clustering algorithm 199
10.7 Bayesian algorithm 200

10.8 Capture-recapture analysis 201
10.9 Conclusion 202

Bibliography **203**

Appendices **208**

1.1 A list of specially-designed computer programs 208

3.1 Cartoon Task 1 210
3.2 Cartoon Task 2 211
3.3 Sample essays for Cartoon Task 1 212
3.4 Combinations 213
3.5 Sample essays for Cartoon Task 2 214
3.6 Index of uniqueness using words occurring frequently in two different tasks 215

5.1 Sample learner essays and native text for review task 216
5.2 List of function words from the British National Corpus 218

6.1 Sample essays for TV task 219

7.1 Principal components graphed against standardized variables 220
7.2 Euclidean distance 223
7.3 Kappa statistic 224
7.4 Technical problems associated with Bayesian classifiers 225

Acknowledgements

I would like to thank my supervisor, Professor Paul Meara. Without his expert guidance, advice and encouragement, I would not have been able to complete this thesis. I would also like to thank all the other members of the Swansea PhD group, past and present, who have all helped or inspired me to some extent. I was lucky enough to start and finish this PhD program at the same time as David Coulson who kept me going on more than one occasion. Eric Bray made an invaluable contribution by his expert rating of essays and his advice as someone who had been there before. I am grateful for the support of my parents, my wife, Maki, and my children, Ibuki and Yujin.

List of Tables

Table 2.1: Expected correlation of multiple judges with six judges 10
Table 2.2: The most common features in e-rater99 13
Table 2.3: The most common features in e-rater01 14
Table 2.4: Reliability of length models, e-rater and human rater (HR) scores 14
Table 2.5: Squared partial correlations of e-rater with human raters........ 14
Table 2.6: Squared partial correlations between features and human ratings 15
Table 2.7: Objective statistics of student essays........ 23
Table 2.8: Average holistic scores by proficiency........ 34
Table 2.9: 28 quantitative, lexical and syntactic features in the analysis........ 34
Table 2.10: Means and standard deviations for 18 features by group 35
Table 2.11: Results of a stepwise regression analysis........ 35
Table 2.12: Statistics for participials in Group 1 37
Table 2.13: Mean percentages of word families at different frequency levels........ 40
Table 2.14: Correlations between the LFP and the VLT........ 40
Table 2.15: Lexical signatures for subsets of high, medium and low frequency words........ 48
Table 2.16: Distinct and unique signatures by subset size 49

Table 3.1: Words selected for analysis 63
Table 3.2: Four examples of subsets........ 64
Table 3.3: Binary representations of essays for (*came, my, sea, clothes, threw, it*)........ 64
Table 3.4: Unique signatures for subset (*came, my, sea, clothes, threw, it*)........ 65
Table 3.5: Statistics for subsets of various size........ 66
Table 3.6: Comparison of six-word subset to the original study 66
Table 3.7: Summary of uniqueness index scores for 64 essays 67
Table 3.8: Average, highest and lowest reliability coefficients for ten iterations........ 67
Table 3.9: Possible subset combinations by sample and subset size........ 70
Table 3.10: Sampling rates by number of subsets for a six-word subset........ 70
Table 3.11: Number of subsets to achieve sampling rate of 1.68% 71
Table 3.12: Candidate grades for Tasks 1 & 2 71
Table 3.13: Thirty selected words for Tasks 1 & 2........ 72
Table 3.14: Task characteristics 72
Table 3.15: Reliability of basic features over the two tasks 73
Table 3.16: Unique lexical signatures and uniqueness index scores for both tasks 73
Table 3.17: Number of thirty words per essay........ 77
Table 3.18: Thirty words that appeared frequently in both tasks........ 77

Table 4.1: Distinctiveness scores and lexical features for two tasks........ 81
Table 4.2: Distinctiveness and other measures for revised analysis 82
Table 4.3: Holistic ratings of written tasks........ 82
Table 4.4: Distinctiveness scores with/without unique words........ 87
Table 4.5: Total words, word types and type-token ratio........ 88
Table 4.6: Reliability for different tasks and same tasks 89

Table 5.1: Mean scores for MWL and MSL for each group 98
Table 5.2: Mean scores for entropy, Yule's K and the D estimate for each group 98
Table 5.3: Top five Zipf ranked words for each text........ 108
Table 5.4: Mean word length for native and learner texts........ 110
Table 5.5: Mean, maximum and minimum values for TTR and LV........ 113
Table 5.6: Ratings of essays ordered by lexical score........ 114
Table 5.7: *Not in lists* words for Tasks 1 & 2 116
Table 5.8: LFPs for 14 learners on two tasks........ 117
Table 5.9: Reliability statistics for LFP over Task 1 & 2........ 118
Table 5.10: P_Lex, and difficult/easy words for Tasks 1 & 2 (T1 & T2)........ 118

Table 6.1: The number of essays for each rating category........ 123
Table 6.2: Essay characteristics of 34 learner essays........ 126

Table 6.3: Correlation of essay ratings, essay length and lexical diversity 131
Table 6.4: Partial correlations with essay ratings controlled for essay length 132
Table 6.5: R and r^2 values for two variable model with essay length 133

Table 7.1: Percentage of variance accounted for by the PCs 143
Table 7.2: Features and weightings used in the clustering 145
Table 7.3: Decision agreement (DA) and Kappa for clustering algorithm 146
Table 7.4: Automatic treatment of essays agreed on by raters 147
Table 7.5: Essays identified by the training set conditions 152
Table 7.6: Decision agreement and Kappa reliability for Bayesian algorithm 153
Table 7.7: Training set essays rated high by human raters 153
Table 7.8: Automatic treatment of essays agreed on by raters 154
Table 7.9: Bayesian classifier treatment of essays agreed on by raters 155
Table 7.10: Decision agreement between raters and automatic methods 156
Table 7.11: Classification of essays agreed on by both raters 157
Table 7:12: Agreed essays by raters and algorithms 157
Table 7.13: Agreement of two automatic algorithms 158
Table 7.14: Agreement of essay length with raters and algorithms 159

Table 8.1: Results of full capture-recapture analysis 167
Table 8.2: Capture-recapture analysis using words only occurring once in a task 167
Table 8.3: Values of Ñ for two simulations 170

Table 9.1: Variables from Ferris (1994) in terms of their ease of calculation 178

Table A1: Thirty words appearing frequently in both tasks 215
Table A2: Unique lexical signatures and uniqueness index scores for both tasks 215
Table A3: Two rater representation 224
Table A4: Example two rater representation 224

List of Figures

Figure 2.1: A graph of a 75%-10%-10%-5% LFP 42

Figure 3.1: Uniqueness index scores for selected tasks 74

Figure 4.1: Distinctiveness for selected learners on both tasks 83
Figure 4.2: Essay length for selected learners on both tasks 84
Figure 4.3: Number of word types for selected learners on both tasks 84
Figure 4.4: The relationship between distinctiveness and quality 91
Figure 4.5: Distinctiveness weighting using reciprocal 93
Figure 4.6: Distinctiveness weighting using reciprocal square root 93

Figure 5.1: Entropy for increasing essay length for two essays 99
Figure 5.2: Yule's K for increasing essay length for two essays 100
Figure 5.3: The D estimate for increasing essay length for two essays 101
Figure 5.4: Word frequency distribution for learner essays 105
Figure 5.5: Word frequency distribution for native samples 105
Figure 5.6: Hapax legomena for native and learner essays 106
Figure 5.7: Zipf ranked words 1 to 15 107
Figure 5.8: Deviations from typical occurrence of *a* and *the* 108
Figure 5.9: Distribution of word lengths for native and learner essays 109

Figure 6.1: Four quadrants in a standardized variable graph 126
Figure 6.2: Quantity v content using TTR(100) as content 127
Figure 6.3: Quantity v content using Guiraud Index as content 128
Figure 6.4: Quantity v content using Yule's K as content 129
Figure 6.5: Quantity v content using Hapax(100) as content 129
Figure 6.6: Quantity v content using the D estimate as content 130
Figure 6.7: Quantity v content using an estimate of AG as content 131

Figure 7.1: Scree chart of PCs 143

Figure 8.1: Estimates for each essay on both analyses 168
Figure 8.2: Range of N estimates for various sample sizes for population size 110 171
Figure 8.3: Number of word types and essay length 171
Figure 8.4: Number of types occurring only once and essay length 172

Figure A1: Uniqueness index scores using words from both tasks 215
Figure A2: PC1 v essay length 220
Figure A3: PC2 v TTR(100) 220
Figure A4: PC3 v mean sentence length 221
Figure A5: PC4 v error estimate 221
Figure A6: PC5 v mean word length 222
Figure A7: PC6 v Hapax(100) 222

Chapter One: Introduction

1.1 Aims of the thesis

There are two main aims of this study. The first is to identify quantitative lexical features that can contribute to assessments of quality of second language (L2) learner essays within an automatic scoring system. The second aim is to evaluate some systems for assessing L2 essays incorporating some of these features. An additional, more general, aim is to contribute to L2 vocabulary research. This contribution may involve the relationship between features and quality in L2 productions, the complementary properties of various features of L2 writing and the use of statistical techniques to aid research in L2 vocabulary.

1.2 Background

There are several reasons why this study is important. Some relate to practicalities of assessment and some to more general research. Firstly, there is a need for tools to evaluate learner essays that can be used in a variety of classroom and testing situations. Automatic assessment may play a number of roles. It can be used instead of human rating possibly for reasons of cost, time or resources. Human raters may be expensive to recruit and train. Human rating takes time. Since a lot of rating involves teachers, this often diverts time away from other activities. Also, sometimes it may be difficult to recruit human raters. Trained raters are a scarce resource. Automatic assessment may also be used to complement human raters. In some testing situations, multiple raters are deployed to ensure reliability. However, in a lot of classroom situations, a second rater is not an option. An automatic rating system could be used as a second rater to ensure reliability. It can also complement human rating when objectivity becomes an issue. Teachers sometimes find it difficult to objectively assess their own students. In these situations, an automatic system can provide a check against bias as well as boost reliability.

A second reason why this study is important is that automatic assessment enables teachers to promote written activities and assessments. Many teachers would like to include more writing in classroom and assessment situations but may be put off by the time and effort involved with assessment. Automatic assessment is also a useful technology for learners to manage their own learning. Learners could use automatic assessment applications to assess their own writing and to receive some simple feedback.

This study also makes contributions to research in a number of ways. For example, research is necessary to help understand more clearly questions relating to the assessment of essay quality and language ability through written tasks. However, human rating of essays can be a problematic area. It is notoriously unreliable and often it is not clear what aspects of an essay influence rating. Even experienced raters are sometimes unable to articulate clearly the reasons for their decisions. An automatic essay scoring system will not operate the same way as a human rater. Some features that appeal to human raters may be difficult for a computer to measure. It may be necessary to find other features of essays accessible to computers approximating features that appeal to human raters. By examining some of the essay features that vary according to quality, we may be able to shed light on some of the distinctions between good essays and poor essays. Statistical patterns can help us uncover hidden truths about language use. This is demonstrated by the success of corpus linguistics to identify patterns of language use. Sinclair (1991) notes that analyses made possible by large corpora have shown that people's intuitions about language do not always correspond with actual language use. Human intuitions are not necessarily false but they may represent only part of the picture. Statistical analysis may help complete this picture by identifying patterns not immediately obvious through human analysis. It is quite possible that statistical analysis can similarly highlight patterns of language use that impact on essay quality ratings.

However, although this study may shed light on some aspects of language use in written productions and on which essay features influence raters, it is important to stress this is not the primary aim of the study. The aim is to explore the features that might help *predict* human ratings of essays rather than attempt to *explain* why essays are rated in a particular way.

A final reason why this study is important is to make a contribution to research into lexical features more generally. A lot of research has been done into various features but not so much into how they work together. Nation (2007) calls for a greater use of multi measures. One aspect of this is recognizing the complementary nature of various features. This involves understanding more about properties of features and their strengths and weaknesses as well as how they operate with other features. This research can help in that direction.

The experimentation in this thesis is split into two parts. The first part involves identifying features associated with essay quality. The second part involves evaluating systems that could help assess essays.

1.3 Identifying features

First of all, some basic features of essays and their relation to essay quality are considered. In this study, an essay feature has a broad meaning including any property of an essay such as length or a count of particular words or classes of words. It also encompasses more complex measures such as lexical diversity. There has been much research into various features and their relationship to essay quality. This can help guide the experimentation. Some very simple features have been found to have a strong relationship with essay quality. These include the total number of words in an essay (tokens) or the number of different words (types). Similarly, many studies have considered various measures of lexical diversity such as the type token ratio (TTR). Recently, more complex measures of lexical diversity such as the D estimate (Malvern & Richards, 1997) and Lexical Frequency Profiles (Laufer & Nation, 1995) have been developed.

Features may be evaluated in their own right as indicators of essay quality or as one of a battery of measures. The latter case involves concentrating on the complementary aspects of features. This also underlines the importance of considering a variety of measures. Any assessment system is unlikely to be fixed. Each situation may vary according to the task, the students, the purpose of the assessment, assessment conditions etc. Each possible scenario may benefit from a slightly different type of assessment system including different features.

1.4 Developing assessment systems

The second part of the experimentation considers some basic systems for assessing essays. Recently, automated essay scoring has become a field of its own and a number of automated systems have been developed to grade essays of both first language (L1) students and L2 learners. Although these systems have shown some success, they tend to be geared towards large-scale testing situations and are often unsuitable for smaller scale L2 situations. There are two notable characteristics of these systems. The first is that they require training samples to develop a model to grade the essays. These training samples may consist of graded essays, model essays or expert material. The second characteristic is the large number of essays required for these systems to operate effectively. This is partly a result of the statistical procedures involved in these assessment systems and partly because of the requirement for training data. This number of essays is typically in the many hundreds or even thousands. The result is these systems are often unwieldy and inflexible. They have been criticized for their inflexibility and their unforeseen effects on educational environments (Herrington & Moran, 2001). For example, the requirement of training samples to build a model means that, in practice, essay topics are set by the testing organization. Therefore, teachers lose control over the writing process. This requirement of training samples makes these systems difficult to use where only a small number of essays exist and entails time and planning for training samples to be produced.

This often precludes these systems from being used in real time. Nevertheless, there are many things that can be learned from an examination of these existing automatic systems. However, there is a need for a more flexible form of automatic assessment that can be used for a variety of topics, with smaller numbers of essays and which can be deployed in a much quicker timeframe.

1.5 Influence of other fields

This study is based in applied linguistics. It aims to deploy a variety of linguistic features to help assess L2 written tasks. In particular, it draws heavily on research and techniques used in L2 vocabulary. Lexical features are useful because they are wide ranging, easy to calculate and there is a large body of quantitative research dedicated to them. At the same time, this study also draws on other fields. Statistical techniques are a core ingredient of most assessment systems. Some of these statistical techniques have also been developed in state-of-the-art fields such as information retrieval and pattern recognition.

The field of authorship attribution also provides inspiration. This field involves the use of statistical analysis of linguistic features to identify authorship of disputed texts. Authorship attribution has a long history and many statistical measures and procedures have been developed in this field. The aims of the field may be different but some of the features and statistical tools may be applicable to assessment. The methodology in this field may also be instructive. Typically, absolute proof of authorship is difficult to achieve but the aim is to produce a body of overwhelming evidence from various sources so that one can make decisions with a high degree of probability of being correct. A similar approach to assessment may be necessary. Evidence collected from a variety of features and techniques is likely to give a more reliable assessment estimate.

1.6 Automation

In line with the aims of this study, the analyses conducted in all the experiments are automated as far as possible using specially designed computer applications. Applications were designed using Delphi 6 software (Borland, 1983). Where other means are used, they are acknowledged. To use a computer program to analyze data offers many advantages not least speed and ease of calculation. However, computers also place some limitations on the analysis. There are certain things that are easy for a human rater to do but that are difficult for a computer to do and so alternative strategies may be required.

1.7 Outline of the thesis

This thesis can be split into three parts as follows:
Literature review (Chapter Two)
Experimental work (Chapters Three to Eight)
Discussion and conclusion (Chapters Nine and Ten)

In the first part, Chapter Two, a literature review lays out the background to this thesis. Some experiments that are considered important in presenting the case for automatic assessment are summarized and critiqued.

The first three papers in Chapter Two consider some of the essay scoring applications currently in use. The paper by Page (1994) reports on Project Essay Grade, an automatic system used for rating native speaker essays. This report makes a strong case for the efficiency of automatic assessment. Chodorow and Burstein (2004) analyze some of the features of another automatic system, e-rater, and its use on L2 writing in the TOEFL written test of English. They also address the charge that e-rater is overly influenced by essay length. Landauer & Dumais (1997) describe The Intelligent Essay Assessor (IEA) and the dynamics of latent semantic analysis which offers an alternative approach to automatic assessment based on semantic relations.

The next group of papers relate to L2 learners and the use of specific objective measures and their relationship to quality assessments. The paper by Larsen-Freeman & Strom (1977) highlights the usefulness of essay length and error free measures of grammatical complexity. Arnaud (1984) introduces some measures of lexical diversity and considers their relationship with essay quality. The cases for lexical diversity and error measurement are developed by Engber (1995) who argues that lexical choice together with error exert a greater influence on holistic impression than either feature individually. Ferris (1994) looks at a wider range of features of L2 writing and the link between these features and quality judgments. This paper also introduces some useful statistical techniques that can be used to analyse multi-feature data.

The remaining three papers look at some methods that have been developed to measure L2 vocabulary knowledge but which can also inform about essay quality. The first two papers involve two extrinsic measures of lexical diversity, Lexical Frequency Profiles (Laufer & Nation, 1995) and P_Lex (Meara & Bell, 2001). These measure lexical diversity in relation to the frequency of words in general English. They may inform about essay quality under an assumption that essays including more infrequent vocabulary are better essays. Finally, Meara, Jacobs & Rodgers (2002) introduce a new type of lexical metric that measures the unique lexical choices of learners in essays. They argue that essays which include more unique lexical items are likely to indicate learners with a larger vocabulary. Chapter Two is then concluded with a discussion of how this empirical work shapes the experimentation reported on in the following chapters.

The second part of the thesis describes the experimental work. This experimental work itself can be split into two parts. Chapters Three to Chapter Five concentrate on lexical features that could be incorporated into an automatic scoring system. Chapters Six to Chapter Eight concentrate on some simple ideas for automatic systems.

Chapter Three includes two experiments involving the lexical signature approach of Meara, Jacobs & Rodgers (2002). Using this approach, an index of uniqueness is developed and its suitability for assessment purposes investigated. Chapter Four is an extension of this research and develops a similar concept of distinctiveness as an alternative to uniqueness. Chapter Five contains four experiments investigating various lexical features often used in authorship attribution or elsewhere in applied linguistics. These features include essay length, lexical diversity, and features of word distributions.

Chapters Six to Eight look at some simple systems that could be used for assessment and which may include some of the features identified in the earlier chapters. Chapter Six proposes a simple two-dimensional 'quantity v content' model to assess essays. The quantity dimension is measured by essay length and the content dimension by lexical diversity. An experiment is conducted to gauge the efficiency of this model and to examine various measures of lexical diversity as the content dimension. Chapter Seven includes two experiments involving more complex systems of assessment. Two different assessment algorithms are used on the same set of essays and the results compared. One algorithm utilizes cluster analysis and the other a Bayesian classifier. In the first algorithm, cluster analysis is used alongside principal component analysis to try to identify groups of essays according to spatial representation in a number of dimensions. In the second algorithm, a Bayesian classifier is used to categorize the essays as good essays or poor essays. The latter algorithm gets around the need for an external training set of essays or expert materials by using a training set identified from within the set of essays. The two algorithms using cluster analysis and Bayesian analysis are tested on the same set of learner essays to get an insight into inter-algorithm reliability. Chapter Eight investigates a method taken from biology called capture-recapture. It is proposed as a method of assessing essay quality by making predictions of the vocabulary size of the writer of the essay. It could be used in assessment situations where more than one piece of writing from a learner is available.

The final part of the thesis includes two chapters. Chapter Nine is a discussion of some of the issues raised by the experimental chapters. It focuses on reliability and validity concerns and identifies some limitations of this study. It also evaluates the contribution of this study to research in L2 vocabulary acquisition and looks at how this research could be continued in the future. Chapter Ten is a brief conclusion which summarizes the main findings of this thesis.

Chapter Two: Literature Review

2.1 Introduction

In this literature review, some papers describing existing automatic scoring systems and potentially useful quantitative features are reviewed. The suitability of these systems and features for a more flexible role in L2 assessment is considered.

The first three papers look at three systems of automatic essay scoring. Page (1994) makes an excellent case for automatic essay scoring and describes the development of Project Essay Grade, a system for rating native speaker essays. Chodorow & Burstein (2004) describe e-rater, a system developed by Educational Testing Service that has been used to automatically rate the TOEFL Test of Written English. They also investigate the charge that the system is overly dependent on essay length. Landauer, Laham & Foltz (2003) report on experiments involving The Intelligent Essay Assessor which employs latent semantic analysis to assess essays according to their semantic content.

The remaining papers involve various features and methodologies that may be useful in automatic assessment. Some relate features directly to essay quality while others focus on general proficiency. In the first paper, Larsen – Freeman & Strom (1977) assess various features as potential indices of L2 development. The next two papers by Arnaud (1984) and Engber (1995) focus on measures of lexical diversity. Arnaud explores the relationships of both lexical diversity and essay length with essay quality. Engber also investigates lexical diversity but focuses on how, alongside lexical error, it can account for essay quality. Ferris (1994) investigates a wider array of features using a discriminant analysis and stepwise multiple regression to identify essay features affecting quality. A large number of lexical and syntactic features are incorporated into the analysis to discriminate between essays of two proficiency groups.

The papers by Laufer & Nation (1995) and Meara & Bell (2001) argue that learners tend to acquire higher frequency vocabulary before lower frequency vocabulary. They have developed techniques to measure proficiency by analyzing the frequency level of vocabulary used in essays. The paper by Meara, Jacobs & Rodgers (2002) proposes an innovative approach to essay evaluation using lexical signatures. This lexical signature approach is based on a concept of uniqueness and the authors argue that more proficient learners are likely to produce more unique patterns of lexis. Measuring the degree of uniqueness of lexis in an essay may give insights about the proficiency of the writer and enable us to make predictions about the quality of the essay.

2.2 Page (1994)

2.2.1 Summary

This is a progress report on Project Essay Grade (PEG), an ongoing project using computers to evaluate L1 writing of American high school students.

Page argues that, ideally, essay quality should be judged by humans but there are some problems with this in practice. One problem is that teachers are busy and often do not have enough time to rate essays. This may mean that many essays either go unmarked or are not assigned in the first place. Another problem may be that essay ratings suffer from low inter-rater reliability. This may in part reflect inaccuracy and bias. Increasing the number of judges improves reliability by reducing this judge inaccuracy and bias. However, as the number of judges increases, so do costs often rendering multiple judges infeasible.

One source of inter-rater variation arises because human judges do not always use the same criteria to judge essays. Criteria are also a major problem when devising a computer analysis. Many criteria that appeal to human judges may be infeasible for computers to assess. Page refers to intrinsic qualities of an essay that might appeal to readers and judges as *trins*. In contrast, Page refers to features that can be measured and may be markers of these intrinsic qualities as *proxes*. *Trins* may include criteria such as fluency, diction, grammar, punctuation. For the *trins* of diction, a *prox* of word length might be suitable since longer words are generally less common. For the *trins* of complexity of sentence structure, *proxes* such as the number of prepositions or the number of relative clauses can be considered.

To test the effectiveness of some of these *proxes*, an experiment was conducted using 599 essays from 12th grade students collected as part of the National Assessment of Educational Progress (NAEP) in 1990. The task involved responding to a question about whether the city government should spend money fixing abandoned railway tracks or converting old warehouses. The essays had two ratings supplied by NAEP, one for task completion and the other for overall quality. Page utilized the overall quality rating and recruited more judges to rate on a six-point holistic scale. The pooled ratings of six judges were taken as the holistic scores for the essays. The ratings of any two judges correlated with each other at an average of $r = 0.564$. Page went on to investigate reliability ratings of groups of judges. As the number of judges rises so should reliability. Given that the reliability of one judge was r = 0.564, the expected reliability of six pooled judges can be predicted using the Spearman-Brown formula as r = 0.886.

An automatic prediction of essay quality using a number of variables was carried out using a multiple regression analysis. Unfortunately, the variables included in the analysis were not specified in detail in this paper. However, Page reports that, in previous studies, essay length was found to be the most important variable. In this regression model, the fourth root of essay length was used. The rationale was that judges expect and credit a certain amount of writing but credit for writing over and above that amount tails off very quickly. This dynamic was best modeled by using the fourth root of essay length. The correlation between the computer prediction generated by the multiple regression analysis and the holistic score of the pooled judges was r = 0.877. This fared very well in comparison with the expected inter-rater reliability of the six judges (r = 0.886).

Table 2.1: Expected correlation of multiple judges with six judges

Number of judges	Expected correlation with six judges
1	.706
2	.799
3	.839

Page then investigated how many judges would be required to achieve the same reliability as the computer prediction with the results in Table 2.1. The computer method seems to be more accurate than a panel of two or three judges rating each essay. In fact this computer method was consistently more reliable than *any* group of two or three raters from the group of six raters.

To check the validity of regressions, the set of essays was randomly split into two groups of size 1/3 and 2/3. Two-thirds were used to establish weights for the predictor variables and the remaining one-third was used for testing these predictor variables against the ratings of the judges. This procedure was repeated three times and correlations of r = 0.864, 0.863 and 0.849 were returned. Finally, validity of results was tested across years. A sample of 495 essays on the same topic from 1988 was used to establish weightings which were then tested against the judge ratings of 1990 essays. The correlation from 1988 predictors and 1990 judge ratings was r = 0.828.

2.2.2 Comments

This study sets out a very strong case for the use of computer methods to assess written tasks in an L1 context. On the whole, many of these same arguments are relevant in L2 contexts. Human rater reliability is a major concern. In addition, although L2 rating criteria may mean different predictor variables are used, essay length is still likely to be important in L2 essays. However, whether the statistical methods used in this experiment are appropriate for smaller scale testing situations is in some doubt.

The performance of computer analysis to predict holistic scores in this analysis is quite remarkable. The computer method performs consistently better than groups of two or three independent judges. This is largely because of weakness in ratings of individual judges. The sources of this weakness include inaccuracy and bias in individual judges as well as different rating criteria applied by judges. This weakness leads to low inter-rater reliabilities between judges. In order to limit this low reliability, multiple judges can be used. But the three or more judges needed to raise the inter-rater reliability to a level achieved by automatic means are, in practice, economically not viable. Continued research can only be expected to further improve the performance of automatic methods.

Whether this method that works so well in an L1 context is applicable in L2 contexts is an important question. The basic concerns seem applicable to L2. Firstly, as in L1, in L2, many teachers find grading essays a time consuming task. As in L1, there are few opportunities to have another teacher rate or help rate essays and costs also impinge on the use of judges. Although rating criteria vary between L1 and L2, raters are subject to the same kinds of error, bias and inconsistency in the rating process which adversely affects inter-rater reliability. An effect not addressed directly in this study is that of rater independence in relation to students. When teachers act as raters, it is sometimes difficult for them to judge their students' output objectively. Knowledge of students may introduce bias into the grading and rating of essays.

The actual variables used in this experiment were not specified. The *trins*, intrinsic qualities that identify a good essay, are likely to be different in L1 and L2. In L2, emphasis will be more on language ability rather than content and ideas. Therefore, different *proxes*, indicators of these *trins*, will probably be used as independent variables. However, essay length, which Page found to be influential, is also likely to be important in L2 essays. Page found that essay length was most important in short essays but less important in longer essays. L2 texts have a tendency to be short and many studies have suggested an association between essay length and holistic rating, for example, Larsen-Freeman & Strom (1977).

The argument for automatic assessment delivered by this experiment is strong. However, there is one weakness. The statistical methods involved favour testing situations with a large number of essays. Multiple regression is a statistical analysis which demands a high number of datasets. In this case, the number of datasets corresponds to the number of essays. In fact, Hatch & Lazaraton (1991) suggest thirty datasets for each independent variable. So, for example, if five variables were used, at least 150 datasets would be necessary. Although Page does not specify how many independent variables were used in this experiment, the high number of essays available, 600, was more than sufficient. However, it is not uncommon for teachers to have far fewer essays that require rating. In these situations, multiple regression may not be the most appropriate form of analysis.

This study shows that unless there are upwards of three teachers to rate each essay, automatic rating may be technically superior to human rating. Use of human judges on such a scale is prohibitively expense in most cases. This suggests that automatic assessment may be useful for more than speedy online applications where human rating is impossible. It may also be superior to using humans in more general testing situations. However, the challenge is to make automatic assessment a more versatile alternative for use in a wider variety of test formats.

2.3 Chodorow & Burstein (2004)

2.3.1 Summary

Chodorow & Burstein describe e-rater, an automated scoring system developed by Educational Testing Service (ETS). E-rater has been used in large-scale L1 testing environments and also on the TOEFL Test of Written English. In this experiment, Chodorow & Burstein address the effects of essay length on TOEFL written test scores awarded by human raters and by e-rater.

This study was concerned with two points related to e-rater and its role in rating the written portion of the TOEFL test. Firstly, Sheehan (2001) argues that an early version of e-rater, e-rater99, depends on features influenced by essay length rather than writing quality. Secondly, there was evidence of variation in TOEFL written scores according to native language. Accordingly, this experiment investigates the following:

1) performance of e-rater against human raters when essay length is controlled.
2) features of essays that are effective at scoring essays when length is controlled.
3) whether differences in TOEFL written scores according to native language are due to essay length.

4) whether e-rater scores show the same difference by native language as human raters.

TOEFL written answers to seven prompts were used including the following example:
Do you agree or disagree with the following statement? Playing a game is fun only when you win.

Students taking the test had thirty minutes to write an essay on one of the given prompts. Essays were scored by two human raters on a scale of one to six. The final score awarded was the mean of the two rater scores unless they varied by more than one point. In that case, a third rater graded the essay and the score was the mean of the third rater and the closest score from the first two raters. ETS provided a scoring rubric which contained no reference to essay length.

E-rater measures over fifty features from four types: syntactic, discourse, topical and lexical. The lexical features were added for the e-rater01 version. However, there was no direct measure of essay length included. In this experiment, two versions of e-rater, e-rater99 and e-rater01, were trained on a set of 265 essays for each prompt. This training set was graded by human raters and used by e-rater to create a stepwise linear regression model. This model found eight to ten features to best score each prompt giving a unique model for each prompt. The score produced was rounded to the nearest whole number for the predicted score. The training essays included essays with all possible scores from human raters. A set of 500 mixed essays with proportions of each score matching the general scoring profile of TOEFL test-takers was also rated for each prompt using the model. Also, large sets of essays for each prompt from three language groups of Spanish, Japanese and Arabic speaking test-takers were rated.

The most common features in e-rater99 and e-rater01 appearing in four or more of seven models are shown in Table 2.2 and Table 2.3.

Table 2.2: The most common features in e-rater99

Feature	No. of models (out of seven)	Feature Type
number of auxiliary verbs divided by number of words in essay	7	syntactic
number of auxiliary verbs	7	syntactic
score from topical analysis by argument	7	topical
number of verb+ing words	6	discourse
score from topical analysis by essay	5	topical
number of argument development contrast phrases	4	discourse

Table 2.3: The most common features in e-rater01

Feature	No. of models (out of seven)	Feature Type
number of words of various lengths 5,6,7,8 characters	7	lexical
number of different word types	7	lexical
number of auxiliary verbs in essay	5	syntactic
score from topical analysis by argument	6	topical
number of argument development words in first paragraph	5	discourse

To investigate the relationship of essay length to human rater and e-rater scores, a length based model using polynomial regression was fitted to the same training essays and mixed essays. This length based model incorporates a measure of the number of words and the number of words squared. The proportion of exact agreements and adjacent agreements with the human raters was calculated. Performance reliability was also measured as a Kappa statistic (K), the proportion of actual agreement of model scores and human rater scores after chance agreement has been removed. Similarly, the agreements of the two versions of e-rater, e-rater99 and e-rater01, and the correlations of the two human raters were calculated. This information for all the prompts combined is shown in Table 2.4.

Table 2.4: Reliability of length models, e-rater and human rater (HR) scores

	Combined r^2	Exact agreement (K)	Adjacent agreement (K)
length model	.53	.49 (.31)	.95 (.84)
e-rater99	.50	.46 (.27)	.92 (.76)
e-rater01	.60	.53 (.37)	.96 (.88)
HR1 v HR2	.59	.56 (.42)	.96 (.86)

The length model outperforms e-rater99 by correlating better with human raters and having a higher proportion of exact agreement and adjacent agreement with human raters. However, e-rater01 performs better than both the older version of e-rater and the length based model and compares with the agreement between the two raters.

To investigate the essay length effects further, squared partial correlations between e-rater and human raters with essay length removed were calculated. The results are shown in Table 2.5 for the combined prompts.

Table 2.5: Squared partial correlations of e-rater with human raters

	Human raters	Single rater
e-rater99	.06	.04
e-rater01	.15	.11

partial variance between two raters = .26

The low partial correlations between e-rater and human raters support the idea that e-rater is based largely on length of essay. E-rater99 shows little variance outside that provided by essay length but e-rater01 is slightly better. In order to find out how sensitive individual features were to essay length, the squared partial correlations between each feature and the human rater score were calculated and are shown in Table 2.6.

Table 2.6: Squared partial correlations between features and human ratings

E-rater model	Feature	Sq. Partial correlation
99	number of auxiliary verbs	<.001
99	number of auxiliary verbs divided by number of words in essay	<.001
99	number of verb+ing words	.011
99 / 01	score from topical analysis by argument	.076
99	score from topical analysis by essay	.075
01	number of different word types	.099
01	number of words of various lengths, 5,6,7,8 characters	.134

All the features have low correlations but the new lexical features in e-rater01 perform relatively well.

The study also investigated the relationship between essay length and native language of test-takers. Spanish speaking test takers tended to score higher than Japanese and Arabic test-takers but also tended to write longer essays. Results of an ANOVA suggested that the length of essay was not entirely the reason for the higher scores by Spanish students and that a native language effect did exist.

2.3.2 Comments

This study gives a rare insight into how one of the main automatic assessment packages actually works and raises several points of interest. Firstly, although the effect of essay length on automatic rating can be seen as troublesome from a validation angle, there could actually be ways to positively incorporate it into the assessment model. In addition, there are issues relating to inter-rater reliability in a categorical rating system. Finally, although this type of large-scale testing situation is quite atypical, there are aspects that can be adapted to smaller scale applications such as type of feature selected.

This study highlights the strong correlation between length of essay and ratings by humans and automatic rating systems. However, it is not clear to what degree human raters are actually influenced by the length of an essay. There is some evidence that raters take essay length into account when rating essays. Cumming, Kantor & Powers (2002) found that "assessing quantity of production" was a behaviour often exhibited by raters assessing TOEFL essays in a "think aloud" study. This study seems to show that raters actively assess essay length but it does not really tell us how much this influences their rating. Perhaps longer essays simply give the rater more chance to notice positive aspects of the essay. Or perhaps raters are not influenced by essay length at all and the correlation is due to the fact that longer essays just tend to be better essays.

The case with automatic ratings is different. Here the influence of essay length may be direct. Automatic systems are based on predicting the scores of human raters, and because essay length is often the best single predictor of human scores, automatic systems will identify it as the most efficient feature to predict human ratings. In this experiment, features selected at the training stage may be effective only because they are indirect measures of length. Six of the seven most common features of e-rater99 indirectly reward length because they are simple totals that increase along with essay length. In fact, two of the three features included in all models are the same except that one is controlled for length of essay (number of auxiliary verbs divided by number of words in the essay) while the other is not (number of auxiliary verbs). In this case, the latter measure can only be a proxy for essay length. Even the features in the better performing e-rater01 seem to be mostly totals that will increase with essay length.

The assessment community has an uncomfortable relationship with essay length. It is recognized that human ratings are highly correlated with essay length, but many assessment organizations are at pains to emphasize that rating rewards quality rather than quantity. Accordingly, length is not included in the scoring guidelines for the TOEFL written test. While human rating mechanisms are not clear, the features employed in automatic assessment are open to scrutiny. Both the reliability and validity of automated systems such as e-rater depend on the scores correlating highly with those of human raters. To achieve high levels of reliability and validity, the trade off designers must accept is the ceding of some control of feature selection, in this case, to the stepwise regression analysis which selects the best combination of features on the basis of statistical efficiency. In this case, although designers selected the initial fifty features as input to the regression analysis, the combination of best features is left to the statistical analysis. This combination may undermine validity of the method when examined with hindsight.

Most academic papers written by the main producers of assessment software tend to read like commercial brochures. They are full of claims about their product and its reliability compared with human raters but short on details of how their systems actually work and how they came about these claims of reliability. This paper clearly shows the features used and how their reliability statistics are calculated. The results may also give us a clue as to why, in general, these companies feel such commercial secrecy is necessary. It may not be so much stopping the opposition knowing their model but rather not letting the consumer know what they are doing because it might put them off.

Nevertheless, there may be an argument that essay length has some merit when evaluating second language learners particularly in a timed essay format. Many researchers have proposed a model of lexical competence that includes three dimensions of vocabulary size, richness and fluency, for example, Daller, Milton & Treffers-Daller (2007). Since second language students who can access and retrieve language forms more readily are likely to write relatively more in a timed format, then essay length could act, in Page's words, as a *prox* for the *trins* of lexical fluency.

This study also highlights the problems of inter-rater reliability for categorical data when there are few categories. In this study, reliability is based on the proportion of exact and adjacent matches with human raters. Using adjacent matches, the Kappa reliability adjusted for chance agreement is .88. This compares to a Kappa of only .37 for exact matches. Within a narrow scoring framework as in TOEFL with a scoring system from one to six, and particularly when a high proportion of the scores fall into bands 3, 4 and 5, then reliability based on being within one score of human raters seems to be less than ideal. Proportions of the total test-taking community getting TOEFL scores 3, 4 and 5 are 20%, 37% and 26% respectively. These bands cover 83% of TOEFL test takers. According to the adjacent match reliability statistic, a rating of 4 in this study means we can be only be 88% sure that this candidate is a score 3, 4 or 5. Other testing situations may require fewer categories. In those situations it may be necessary to find ways to boost the exact agreement ratios because adjacent agreement will not carry much meaning. Interestingly, the e-rater reliability is better than human rater reliability under adjacent matches, .88 to .86 but inferior under exact matches, .37 to .42.

The large number of essays available (nearly 11500 for seven prompts) for this study is impressive. It allows for the building of models using many training and cross validation essays. Outside of a few big institutionalized examinations, assessment more often involves a much smaller number of essays. Those situations may not afford the luxury of a training set of pre-rated essays. Perhaps this type of large scale study can help identify some core features that may be useful in other testing situations. For example, in this study, e-rater99 produced four features which were incorporated in all models while e-rater01 found two features that were in all seven models. Of course these features may only effectively work for these particular essays but it may be worth investigating how they work more generally.

2.4 Landauer, Laham, & Foltz (2003)

2.4.1 Summary

The Intelligent Essay Assessor (IEA) is based on the process of latent semantic analysis (LSA) and has been used in the assessment of L1 essays. It aims to measure conceptual content of essays rather than grammar, style or mechanics.

LSA is the mathematical representation of co-occurrence relations among words and passages using statistical computations involving singular value decomposition (SVD). In LSA, each essay is transformed into a vector in semantic hyperspace and compared to other essay vectors. Two measurements are derived to help rate essays. The cosine of each essay vector to other vectors is used to determine the quality of that essay. This is complemented by another measurement, vector length, which measures domain relevance in relation to the semantic space created. In these experiments, three methods of determining the range of the semantic space and judging quality are described. The main method is to use a set of essays pre-rated by humans. However, two other methods are also explored. In one, expert model essays or outside sources are used instead of pre-rated essays and in the other, the internal composition of the set of essays determines the parameters without any other outside input.

This study reports on three main experiments:

1) Heart studies – involving 94 undergraduates writing essays both before and after reading for a psychology course. The tests were scored by two human raters.
2) Standardized tests – 787 essays on two topics for the GMAT and 900 essays on a single topic for grade school children.
3) Classroom studies – 845 undergraduate essays on six different topics from two institutions.

In the experiments involving standardized tests and classroom tests, the semantic space was constructed with essays rated by humans. In these experiments, the IEA score agreed with raters as well as single raters agreed with each other. Inter-rater reliability was checked between the two human raters, IEA and single raters, and IEA and pooled raters. For standardized essays, the values were 0.86, 0.85 and 0.88 respectively. For classroom essays, the values were 0.75, 0.73 and 0.78 respectively. The effect of the number of pre-scored essays on reliability was also investigated. Reliability values varied from 0.53 with six pre-scored essays to 0.69 with 25 essays to 0.75 with 400 essays.

For the heart studies experiment, the semantic space was constructed using paragraphs about the heart from an encyclopedia and high school biology textbooks. Each essay was scored by finding the ten most similar essay vectors and taking an average cosine weight. The scores correlated at 0.65 with pooled scores from two raters. In an alternative method using these same essays, the semantic space was constructed using the student essays themselves and scores were calculated through comparison with other similar essays. The scores using this method correlated at 0.62 with pooled scores from two raters.

The correlation of scores with length of essay was also checked. Standardized essays tended to have a higher correlation than the classroom essays.

Although the correlation between LSA scores and individual rater scores was not as high as that between LSA scores and pooled rater scores, it was still relatively high. The authors suggest using LSA with one rater when a second rater is not an option in order to get a more reliable estimate of the score or to spot inconsistencies in human grading.

The following features of LSA are also of interest:

1) Confidence measures for score quality can be estimated by comparing the cosine between the vector of an essay and a group of the most similar essays. If the nearest neighbour grades are too variable or the cosine value is too low, it may mean that the score for that essay may not be reliable.
2) Plagiarism can be detected by noting essay vectors that are unusually similar. Suspicious essays can then be picked out and examined by hand.
3) Coherence in an essay can be evaluated by comparing similarity of vectors for each successive sentence in the essay.

2.4.2 Comments

LSA is a complex process. To understand it more clearly, it is useful to differentiate between the statistical analysis and the linguistic model. These two aspects both potentially enhance the validity of this automatic assessment technique for L1 learners. However, the suitability of this technique for L2 learners is less clear but there may be useful aspects that could be adapted to an L2 context.

As a statistical analysis, LSA seems to be a form of principal component analysis (PCA). PCA is basically a technique to streamline sets of multivariate data. PCA helps reduce the dimensionality of such data by finding inter-correlations between variables. The result of a PCA is a reduced set of new variables which account for most of the variation in the original data. This smaller number of variables can then be used in other techniques such as clustering. In an analysis for rating essays, the original set of variables could be all the words in the whole set of essays and the vector for each essay would show the occurrence or non-occurrence of each word in that essay. After SVD, a smaller number of new variables in order of variability between essays can be produced. These new variables, although perhaps lacking any obvious intrinsic meaning, could form the dimensions of a semantic space in which the essays could be mapped and evaluated according to their proximity to 'expert' essays. The new set of variables in LSA is typically in the range 100-400. It is worth noting that, in LSA, the exact number of new variables selected to form the semantic space can have a marked impact on the effectiveness of the results. Both too many and too few variables may lead to less effective results (Landauer & Dumais 1997).

LSA is also presented as a linguistic model of lexical acquisition and semantic representation. Landauer & Dumais (1997) argue that this model may help explain the "poverty of the stimulus" problem whereby children appear to acquire far more language than they have direct exposure to through learning. If learning involved processes similar to the mathematical processes in LSA, then children would be able to learn a lot of words through indirect learning. This latent semantic learning is achieved through identifying contexts where words do or do not co-occur with other words. Landauer & Dumais (1997) have found through modeling lexical acquisition with LSA that three quarters of learning could be indirect.

LSA research has mostly focused on native language testing situations. It is supposed to tell us about the writer's conceptual knowledge of a particular topic by the degree of similarity to expert writers, subject material or other essay writers. It is likely that the words that are important in such an analysis may be those with particular relevance to that topic or field. For second language learners, essay grading is not so much about the conceptual content of essays but about language ability. Question topics are likely to be more general and the vocabulary less topic specific. Whether LSA has a role to play in L2 testing situations is still unclear as there seems to have been little research in that area yet. However, LSA has been used to track lexical development in L2 learners (Crossley et al., 2008). In a longitudinal study of six L2 learners of English, the LSA scores of spoken utterances increased over the one year span of the experiment in line with other measures of lexical development.

If we compare IEA to other automatic methods of rating essays such as PEG and e-rater, IEA has a definite advantage in terms of validity if one buys into the linguistic model of LSA. Moreover, the form of the analysis itself seems to provide more validity than these other models. The methodology used by Page or ETS depends mostly on various types of word counts. In LSA, the semantic relationships of words themselves are analyzed. So, although the word order information is still lost in LSA, word meaning information is preserved. This seems to be a step up from largely count-based analyses. However, one drawback to LSA is that despite a more complex model, which may incorporate more information, it does not appear to perform any better than the other models. It would be interesting to see if procedures such as LSA could be utilized alongside other measures to give an enhanced methodology. In this respect, the other methods seem to have an edge in flexibility. Because they are not constrained by any theory, they are free to incorporate anything that works into their models. In fact, the latest version of e-rater seems to have incorporated an LSA-like feature (word-vector score and word-vector cosine correlation) as one part of its analysis (Lee, Gentile & Kantor, 2008). Another disadvantage that LSA shares with other patented commercial applications is that it is difficult to find out exactly how it works. It is marketed for large-scale testing situations and opportunities to experiment with it are limited to online sites such as University of Memphis Coh-Metrix website (McNamara et al., 2005) which provide some simple LSA-like output.

These drawbacks not withstanding, LSA type procedures offer some hints for the creation of a more user-friendly versatile testing instrument. One thing holding back automatic assessment being used in a greater variety of testing situations is the necessity of having pre-rated essays or model essays. Therefore, evaluating essays solely by proximity to other test essays holds promise. One such method employed in this experiment reported high levels of reliability, albeit a little lower than when compared with ‘expert’ papers. This is an approach that should be explored in more detail. Procedures such as PCA might also be helpful more generally by helping select words that offer maximum information for other analyses such as cluster analysis.

2.5 Larsen-Freeman & Strom (1977)

2.5.1 Summary

This study aims to construct an index of development that can be used to identify the proficiency of second language learners. This index is targeted primarily at researchers so that results from different studies can be compared more meaningfully. It would be applicable to different ages and to different native language groups. It is also hoped that it will not be restricted to English as an L2. In this experiment, written essays were considered because texts are easy to work with but it is hoped the index can also be used with spoken English.

The experiment involved 48 undergraduate learners of English as an L2 with no more than two of these learners sharing the same native language. Twenty eight of the subjects were male and twenty were female. Essays were written for the UCLA ESL Placement Examination. The test was conducted under examination conditions but with no time limit. However, a minimum essay length was set (but not directly specified in the paper). The subject of the compositions was not specified. The essays were assessed by two researchers and assigned to five quality levels: poor, fair, average, good, and excellent. There was a high level of agreement in these ratings with an inter-rater reliability of 0.9713 which was statistically significant.

Essays were examined for a variety of features. Examples of errors and native-like usage were noted. Writing mechanics were assessed in terms of punctuation, capitalization and spelling. Organization, clarity, syntactic sophistication and lexical choice were also assessed. The ability to write grammatically was examined by looking at morphology, syntax, prepositions, tenses, aspect articles, subject verb agreements, case and negation. Sentence construction problems such as sentence fragments and run-on sentences were also identified.

Forming an index from all these features proved problematic for various reasons. Firstly, it is difficult to weight the relative importance of errors. For example, it is not easy to say if misuse of an article is a more serious error than inappropriate use of a lexical item. Secondly, an index based on errors may be problematic. It might not be safe to assume a linear relationship between error frequency and proficiency. A decreasing number of errors might be expected as proficiency increases. However, error is not always linear in its relationship to quality. In this study, essays rated poor contained fewer article errors than essays rated fair. Only essays rated good or excellent contained fewer errors than essays rated poor. Thirdly, if an index is based on particular features, those particular features may be biased against certain native language groups which experience interference with that feature.

After considering these problems, it was decided to examine some objective features. The following features were investigated: essay length, the average number of words per T-unit and the number of error free T-units. A T-unit is defined by Hunt (1965) as a "minimal terminable unit,..minimal as to length, and each...grammatically capable of being terminated with a capital letter and a period." For the number of error free T-units, the count was controlled for length of essay and so only 37 of the 48 essays written to a length of 200 words were examined for this feature.

The results of the examination of features found the following trends.

1) There was a decreasing number of spelling errors with proficiency despite the use of more sophisticated vocabulary.
2) Syntactic sophistication improved with proficiency from simple sentences in essays rated poor to complex sentences with relative clauses and multiple embeddings in essays rated excellent.
3) Tense usage improved with proficiency.
4) Errors in morphology decreased slightly from essays rated fair to essays rated excellent.
5) There was no obvious trend in errors related to prepositions.
6) Essays rated fair had more errors than essays rated poor but then the number of errors decreased with increasing proficiency.

The objective analyses produced the results summarized in Table 2.7.

Table 2.7: Objective statistics of student essays

	Poor	Fair	Average	Good	Excellent
Average number of words per essay	132.53	150.55	177.33	218.91	228.00
Number of essays (n=48)	(11)	(12)	(6)	(14)	(5)
Average number of words per T-unit	11.58	12.50	12.92	14.28	14.46
Standard deviation	4.90	2.82	4.22	2.15	3.37
Number of essays (n=48)	(11)	(12)	(6)	(14)	(5)
Average number of error free T-units	3.5	11.7	17.2	25.5	28.5
Standard deviation	3.1	5.1	9.7	7.5	4.9
Number of essays (n=37)	(9)	(9)	(3)	(12)	(4)

Even though there was no time constraint on the essay, there was a steady increase in average essay length with increasing proficiency. Although there was a consistent increase with proficiency in the average number of words per T-unit, the result was not statistically significant. The range of values in each group was relatively high as shown in the standard deviations. Some extreme values may have undermined the trend. For example, the longest average number of words per T-unit was 23 which occurred in a poor essay and some excellent essays had very short T-units. Among the 37 essays examined for error free T-units, there was a highly significant relationship between average number of error-free units and proficiency.

The authors conclude that two features, essay length and the number of error free T-units, show great promise for an index of development. They hope to continue their research by using a similar approach to analyze transcripts of spoken interviews.

2.5.2 Comments

This paper raises many points of interest about the construction of an index of development. In particular, it highlights some of the main differences between the way humans rate essays and how essays might be rated automatically by computer. One aspect of this is the type of feature that can be incorporated into an automated computer application. Another aspect is how large amounts of information can be processed.

This experiment was conducted very thoroughly and a lot of information produced. This information was split into two types. Firstly, a subjective analysis reported on various trends over the different proficiency levels in terms of writing mechanics, lexical choice, and grammar. A second more objective analysis targeted three counts: essay length, the average number of T-units and the number of error free T-units. However, the more we look at these two analyses, the less obvious the differences become. The difference seems to be more in the way the data is reported than the nature of the analyses themselves. In the former analysis, there are subjective elements, as in the decisions about errors but there is also an objective element in that these features are counted. Similarly, the second analysis involves one count that is very objective, essay length, and two others that are more subjective. Both splitting an essay into T-units and deciding whether these T-units are error free or not are subjective decisions. In addition, Nihalani (1981) notes that identifying T-units in low level essays is particularly difficult. This is echoed by various other researchers such as Gaies (1980) who argue that T-units are not suitable for analysis of output of low level learners.

Another way of defining objective and subjective features may be to divide them into features that can be analyzed automatically by a computer and those that require human judgment. Using this distinction, essay length is definitely objective but the other features may be subjective to a certain extent. However, we can probably split the subjective features up further into features that can reasonably be estimated automatically and ones that are more problematic. Clarity would probably belong in the latter category. It is likely, though, that most of the others could be candidates for estimation in some way. For example, perhaps syntactic sophistication could be predicted by the presence of words that typically occur in constructions of a certain sophistication. For example, mastery of relative clauses might be flagged by the occurrence of relative pronouns. Perhaps run-on sentences could be identified by a count of conjunctions between successive full stops. Perhaps lexical choice could be estimated by an analysis of words by their frequency in general English. Less frequent word usage may be a predictor of superior lexical choice. Unfortunately, we would probably lose the very robust T-unit to the category of features difficult to estimate. It is not easy to see how a T-unit could be identified automatically. A measure of average sentence length can be easily estimated by means of a word count and a punctuation count. Although a T-unit is related to a sentence, its difference is likely to be crucial to its efficiency as an indicator.

Although many of these features may be estimated automatically, it is clear that human analysis is able to consider a lot more aspects of an essay than an automatic analysis. However, inter-rater reliability is often a problem with human rating of essays. Some of this inter-rater variability may come from the different ways raters prioritize and process all this information. In this study, the authors were faced with making arbitrary decisions about importance of errors that may lead to bias against some L2 learners. When trying to decide which features were the best to incorporate into an analysis, the authors effectively reached the conclusion that all information might be important. It seems the problem with human rating may not be in identifying important aspects of an essay, but in processing all this information to make decisions consistently and reliably.

Processing information consistently and reliably is something that computers ought to be able to do well. Occurrence of particular features over different essays can be more reliably compared automatically than by humans. It may also be possible for computer analysis to find patterns in data that might not easily be noticed by humans. This may help identify key features that discriminate well between essays of different quality. Computers can also probably make better judgments than humans against outside criteria. For example, it is easy for a computer to compare vocabulary used in an essay against a list of the most frequent words in English.

This study recognizes two features that may be useful, essay length and the number of error free T-units. Even though the latter measure is difficult to use in an automatic program, there are plenty of other features that can be employed. The key to effective automatic assessment could well be not in the individual features but in identifying a number of complementary features that can be automated and then incorporating these features into one analysis.

2.6 Arnaud (1984)

2.6.1 Summary

This paper considers the validity of discrete item vocabulary tests for second language learners by investigating how they reflect performance in real-life writing situations. Vocabulary test scores are compared with some measures of lexical diversity elicited from a written task.

For this experiment, Arnaud had one hundred French students of English complete a 25-item vocabulary test which involved providing English translations for native language words. The students also completed a written task based on the following question:

What do you suggest to improve secondary education in France?

The vocabulary test and the written task were completed with an interval of one week. In addition, for comparative purposes, four American native speakers wrote an essay based on the following question:

What do you suggest to improve high school education in the U.S.A?

Before the main analysis, the essays were reviewed subjectively. Arnaud selected the exact form of the lexical diversity statistics after making several observations about various characteristics of the essays. Lexical diversity in its basic form is the number of different types as a proportion of the number of tokens as follows:

Lexical diversity = types/tokens

Firstly, it was noted that lexical diversity can be affected by non-lexical aspects of language such as cohesion of the text. An essay including repetitions caused by a lack of cohesion may get lower lexical diversity scores. It was also found that lexical items seemed to create the most effect in the essays and some essays included only a very limited range of lexical items. Therefore, lexical variation was deemed as an appropriate measure using lexical types in relation to total tokens as follows:

Lexical variation = lexical types/tokens

Arnaud also noticed that some students had a vocabulary that was insufficient to express value judgments beyond the use of general words like *good* and *bad*. This, Arnaud argues, is related to rareness of vocabulary and it was decided to include a measure of lexical rareness. Rareness was appraised via a French Ministry of Education list of 1522 words which should be known by French students. Any word not included in this list was judged to be rare. Vocabulary rareness was calculated as rare types as a proportion of total lexical types as follows:

Vocabulary rareness = rare types/lexical types

It was also noted that the communicative effect of some essays was hindered by a large number of errors. Therefore a count of lexical errors was included. Non-lexical errors were not included in this count.

Results of the experiment were as follows. Essays varied in length from 180 words to nearly 600 words with a mean length of 313 words. The four native speaker essays varied from just under 300 words to just over 550 words.

Lexical diversity measures are known to be affected by length of the essay. Arnaud demonstrated that this effect occurs in this set of essays. Accordingly, lexical measures were calculated on samples of equal length from each essay. The sample size was equal to the length of the shortest essay, i.e. 180 words. Rather than taking the first 180 words in each essay, 180 words were randomly sampled from each essay. The sampling of essays itself brings statistical error into the process. In order to gauge this, the reliabilities of each measure were calculated by correlations of repeated samplings over twenty essays. Correlations varied from r = .73 to .97. Because essay length was standardized for analysis, the measure of lexical variation simply became the number of lexical items in the 180 word essay.

The distributions of the three lexical measures, lexical variation, lexical rareness and lexical error were investigated. Lexical variation scores ranged from less than 40 to over 70 with a mean value of 55. Some of the French students scored as well as the native speakers in terms of lexical variation. The distribution of lexical rareness ranged from 0.1 to 0.6 with a mean of 0.28. The distribution of lexical error ranged from slightly over 0 to about 18 with a mean of 6.57.

There were no significant correlations between essay length and any of the three lexical measures. Comparisons with vocabulary test scores showed significant correlations between test scores and lexical variation (r = .36) and test scores and lexical error (r = -.21). There were also significant correlations between lexical variation and lexical rareness (r = .27) and also lexical variation and lexical error (r = -.24). Because of the correlations between test scores and lexical variation, and test scores and lexical error, Arnaud concludes that discrete item vocabulary tests are valid indicators of real life language performance. The author suggests checking these lexical diversity measures over two essays written by the same students. He also suggests using the following formula for lexical diversity which he terms vocabulary richness (VR):

$$VR = LV + LR - 2 \times LE$$

where LV is lexical variation, LR is lexical rareness and LE is lexical error.

2.6.2 Comments

Although there may be doubts as to whether this experiment fulfills the stated aim of validating discrete-item vocabulary tests by a comparison with measures of lexical diversity, we nevertheless gain a lot of valuable insights about measures of lexical diversity and their relationship with essay length and evaluations of essay quality. In particular, it is demonstrated that simple measures such as lexical diversity, length of essay, and lexical error are individually inadequate either to predict quality in practice or to provide a theoretical model for a valid assessment system. However, these simple measures together may offer a simple yet powerful model for evaluating overall essay quality.

Ultimately, the experiment fails in achieving the stated aim. Whether vocabulary tests can be sufficiently validated by measures of lexical diversity seems to be debatable. The rationale was to validate these tests by providing comparison with actual productive language use but lexical diversity may be too vague a concept for measuring productive vocabulary knowledge. This approach to validity involves arguing that a feature/test measures a trait by showing that it correlates with an already accepted method of measuring that trait. The problem in this case is that it is not clear whether any measure of lexical diversity is generally considered a measure of vocabulary knowledge. This is epitomized by the fact that Arnaud has not exactly decided how to measure lexical diversity at the beginning of the experiment. As the experiment proceeds, so the concept of lexical diversity evolves from a simple type-token ratio to a multidimensional concept including lexical variation, rareness and error. This study, in fact, highlights all too clearly many reasons why lexical diversity may not be a suitable candidate for validating other measures of vocabulary. Firstly, there is confusion in the literature about how measures of lexical diversity should be calculated. Different authors use different methods of calculations. In addition, lexical diversity is clearly affected by the length of the essay. Once one gets past these problems of how to calculate lexical diversity, one runs into the thorny issue of deciding what it actually measures.

There is one thing that might have been done differently in such an experiment. From an experimental design point of view, it might have been better to decide the form of the lexical features to use before subjectively analyzing the essays. This pre-analysis could have been done using essays from outside the study. By selecting features according to characteristics of the essays under analysis, the generalizability of the results to other sets of essays is compromised.

Although the study may not have achieved its initial aims and the methods employed may be less than ideal, it does, however, address some very important and interesting issues. One is the role of essay length and simple measures of lexical diversity in essay assessment. Another is the fact that simple multidimensional approaches to measuring vocabulary knowledge are far superior to single measure approaches both in practical terms and in terms of validity.

This study clearly shows the merits and limitations of using essay length or a simple measure of lexical diversity to assess essays. Both these measures are useful to a certain extent. For example, their simplicity lends them to easy interpretation. Lexical diversity is versatile in that it can be changed in tune with a particular context as Arnaud's selection of measures shows. The disadvantages of lexical diversity measures are well documented. They can be affected by essay length and sometimes comparisons are difficult across studies because different calculations have been used by different researchers.

Arnaud also supplies more evidence to appraise the value of essay length as a predictor of essay quality. Arnaud reports that short essays are good indicators of poor quality essays while average to long essays do not seem to fit a length-reflects-quality argument. Long essays were often assessed as average while essays assessed as high were often not as long. Contradictory findings relating to essay length and quality assessments over various studies simply reflect the commonsense reality that essay length is related to quality to some extent. This extent may vary from case to case. For example, there may be a limit to the amount of text a very low-level learner can produce in a timed format. Similarly, common sense should suggest that in a timed format, a longer essay is more likely to be written by a learner of high proficiency than a learner of low proficiency. But it is just as easy to see that essay length cannot be equated to quality. For example, one might imagine two learners with different writing styles where one writes freely as much as possible without editing while the other writes and leaves time for reviewing and editing. If they are of a similar proficiency, the first may produce more text but with more errors and more internal problems that detract from the quality whereas the latter may have less text of a higher quality. So, essay length may be an indicator of quality to varying degrees depending on the conditions of the test and the nature of the learners. It is clear, however, that both essay length and simple measures of lexical diversity are inadequate by themselves to assess a complex trait like vocabulary knowledge or language ability.

Although features such as essay length or lexical diversity cannot predict total essay quality by themselves, a simple combination of such features may be able to account for assessments in a far better way practically. For example, even a simple "quantity" and "content" two-dimensional model including essay length and lexical diversity seems powerful and in tune with assessment realities. From a rater's viewpoint, a long essay which has good vocabulary usage, an average length essay with average vocabulary, or a short essay with a poor use of vocabulary should all be easy to assess. A short essay exhibiting good vocabulary or a long essay exhibiting poor vocabulary may be more problematic to assess. This kind of "quantity" versus "content" two-dimensional model would probably perform quite well for assessing low level learners in a timed essay format. Similarly, the more complex concept of vocabulary richness proposed by Arnaud at the end of the study is a more complete model of vocabulary knowledge than a unitary measure of lexical diversity. It includes measures of lexical variation, rareness and error. It seems to correspond more with current studies that emphasize the multidimensional facets of vocabulary knowledge such as vocabulary size, depth of knowledge and lexical fluency. A challenge is to relate these theoretical dimensions of vocabulary knowledge to easily measurable traits of essays and likewise to assessment realities. The theoretical dimensions can help guide the selection of variables. Conversely, statistical measures found to discriminate between essays may also help inform theory.

2.7 Engber (1995)

2.7.1 Summary

This paper investigates the relationship between lexical proficiency and reader quality ratings of timed essays. In particular, it highlights the role of lexical diversity and lexical errors on quality judgments.

Sixty-six timed essays from intermediate to advanced level ESL learners from a variety of language backgrounds were holistically scored by anonymous raters and these scores were compared with the following four measures of lexical diversity based on the essays:

a) lexical variation
b) error free lexical variation
c) percentage of lexical error
d) lexical density

The timed essay was written as a 35-minute portion of a two-hour test. The following title was chosen to give a culturally unbiased uniform product that would generate a sufficient amount of concrete vocabulary:
How will studying in the US help your country?

The essays varied in length from 119 to 378 words. Ten teachers graded the essays on a six-point scale adapted from the TOEFL Test of Written English. For each essay, the highest and lowest scores were dropped and the average score was based on the remaining eight grades. Average scores varied from 1.6 to 5.6 with a mean of 3.4 and a standard deviation of 0.8. Inter-rater reliability was high with $r = 0.93$.

The measures of lexical diversity were calculated as follows:
1) lexical variation including error (LV1)

$$\mathrm{LV1} = \frac{\text{lexical types per segment}}{\text{lexical tokens per segment}}$$

i) Lexical items were defined as full verbs, nouns, adjectives and adverbs with an adjectival base.
ii) The sum of lexical types treated inflected forms of the same word as the same item.
iii) Lexical variation tends to decrease as sample size increases so each essay was divided into 126 word segments, based on 1/3 length of the longest essay, 378 words. LV1 was then calculated for each segment (It is not clear from the paper how segments with less than 126 words were treated).

2) lexical variation excluding errors (LV2)

$$\mathrm{LV2} = \frac{\text{lexical types per segment} - \text{sum of lexical errors per segment}}{\text{lexical tokens per segment}}$$

i) Grammatical and syntactic errors were ignored.

3) percentage of lexical errors (%LE)

$$\%LE = \frac{\text{lexical errors}}{\text{lexical tokens}} \times 100\%$$

4) lexical density (LD)

$$LD = \frac{\text{lexical tokens}}{\text{total tokens}}$$

The correlation of holistic scores with lexical density was r = 0.23 and not significant. The lexical density values compared favourably with Ure's (1971) standard of at least a lexical density of 40% in native speaker texts. There were 10 essays with lexical density slightly less than 40%. In 36% of essays, lexical density was greater than 45%, in 73% of essays, it was greater than 42% and in 85% of essays, it was greater than 40%. The correlation of holistic scores with percentage of lexical error was r = -0.43 which is significant at $p < .01$. This suggests that score increases as error decreases. With lexical variation including errors, the correlation was r = 0.45, significant at $p < .01$, and with error excluded it was r = 0.57 at $p < .01$. The latter three results suggest that readers of essays are negatively affected by lexical errors.

Error and lexical variation together account better for holistic score than either error or lexical variation individually. This suggests that readers give higher scores to essays that use a variety of lexis correctly.

2.7.2 Comments

This paper has two important findings. The first is that both lexical diversity and lexical error have a considerable effect on holistic impressions of L2 writing. The second important finding is that together lexical diversity and lexical error account for holistic impressions better than either one can by itself. However, despite these positive findings, both lexical diversity and lexical error pose challenges to be incorporated into automatic assessment. This is because they involve judgment that is difficult for a computer to do.

An important consideration with these measures is how practical they are. One element of this is whether the calculation can be automated and done by computer. Unfortunately judging lexical errors is difficult even for experienced teachers and examiners. Some of the problems with lexical errors are:

- judging whether something is an error or not.

- judging whether an error is a lexical error or some other kind of error.
- judging to what degree complex compound errors are lexical errors.
- counting errors - a simple count of errors may treat all errors as equally serious.

Also, it may be useful to make a distinction between errors and mistakes. An error is associated with a lack of knowledge or a feature not yet acquired. On the other hand, a mistake involves something the learner already knows. If a learner uses a word one time and misspells it, it may be assumed that the learner has a gap in knowledge. However, if the learner uses a word a number of times, but only misspells it one time, one might assume that the learner knows how to spell it and has made a mistake much as a native speaker may do.

Accordingly, error analysis of an essay often needs careful consideration by an experienced professional and is likely to take as least as much time as a holistic evaluation of the essay. Therefore, it seems difficult to integrate lexical error measurement into automatic assessment.

A very simple variant of lexical density is the type-token ratio. This is simple to calculate automatically because it only demands some simple word counts. However, both lexical variation and lexical density include only lexical items in part of the calculation. This makes the calculation more complex since a decision needs to be made as to what constitutes a lexical item. An exhaustive list of lexical items for a program to check might be possible for task specific applications but cannot be compiled easily in more general cases. Nevertheless, it is still possible to estimate lexical items indirectly. First, a list of function words could be compiled and the number of function types and tokens for the essay calculated. Then an estimate of lexical types and tokens could be calculated by subtracting function types and tokens from total types and tokens. This is a relatively simple method of calculating a feature indirectly. It raises the possibility of measuring lexical error indirectly. Perhaps error could be estimated by somehow recognizing words that are correct. Another possible method of estimation would be to identify a feature that is associated with error.

This study shows that raters are influenced not only by lexical choices but also by lexical error. Measures that incorporate aspects of both lexical choice and lexical error are the best at predicting overall quality. However, error is difficult to calculate automatically. One challenge is to devise an alternative predictor of error that is simpler to calculate.

2.8 Ferris (1994)

2.8.1 Summary

Ferris investigates lexical and syntactic features of L2 essays written by learners of different proficiencies. Discriminant analysis is employed to see how well certain features can identify essays of the two groups and a stepwise multiple regression is used to see how well these features can predict holistic ratings.

Essays on the effects of culture shock written as a 35-minute portion of a university placement test were collected. One hundred and sixty students from four different L1 backgrounds took part, forty each from Arabic, Chinese Mandarin, Japanese and Spanish L1 backgrounds. The essays were rated from 1 to 10 by three independent judges and these ratings were summed to give a total holistic score ranging from 3 to 30. Learners were split into two proficiency levels: Group 1 consisted of 60 low level learners and Group 2 consisted of 100 high level learners. The average holistic scores for each group are shown in Table 2.8.

Table 2.8: Average holistic scores by proficiency

	Average score
Group 1	14.8
Group 2	22.9

The essays were analyzed for the occurrence of 62 quantitative lexical and syntactic features. These features were refined to 28 for further analysis as shown in Table 2.9.

Table 2.9: 28 quantitative, lexical and syntactic features in the analysis

1	Number of words	8	Impersonal pronouns	15	Coordination	22	Definite article reference
2	Words per sentence	9	Adverbials	16	Passives	23	Deictic reference
3	Word length	10	Special lexical classes	17	Complementation	24	Repetition
4	Present tense verbs	11	Relative clauses	18	Prepositional phrases	25	Comparatives
5	Past tense/ perfect aspect	12	Modals	19	Participials	26	Lexical inclusion
6	1st/2nd person pronouns	13	Negations	20	Nominal forms	27	Synonymy
7	3rd person pronouns	14	Stative forms	21	Coherence features	28	Reduced structures

A discriminant analysis was carried out to see how well the 28 features distinguished between the two proficiency groups and correlation coefficients of stepwise multiple regression were analyzed to see how well the same features predicted holistic scores. Table 2.10 shows standardized means and standard deviations (for an average length of 225 words) of occurrences of 18 features which showed a significant difference between the two groups.

Table 2.10: Means and standard deviations for 18 features by group

Feature	Group 1		Group 2	
	mean	sd	mean	sd
Number of words	187.43	25.63	251.38	32.71
Specific lexical classes	4.55	1.30	7.32	2.45
Complementation	4.52	1.76	7.52	2.12
Prepositional phrases	16.87	5.33	23.33	6.24
Synonymy / antonymy	1.82	.64	3.25	1.32
Nominal forms	4.13	.96	6.88	2.05
Stative forms	7.75	1.87	11.01	3.02
Impersonal pronouns	6.12	2.54	9.02	2.89
Passives	.93	.37	1.80	.86
Relative clauses	1.10	.24	1.94	.65
Deictic reference	12.75	3.18	16.84	4.13
Definite article reference	5.80	1.09	8.18	2.53
Coherence features	6.87	.78	8.77	1.96
Participials	.42	.11	.95	.43
Negations	2.07	.55	2.84	.62
Present tense	15.80	3.73	18.87	5.09
Adverbials	2.73	.88	3.70	1.21
1st/2nd person pronouns	12.98	3.23	16.14	4.57

Results of the discriminant analysis showed that:

- the 28 features divided the learners into two groups with 82% accuracy.
- higher level learners used more textual features than lower level learners.
- the data supports earlier research that a factor involving passives, nominalizations, conjuncts and prepositions was positively correlated with holistic scores. All were used with greater frequency by learners of higher proficiency.
- advanced learners have more lexical, syntactic tools available, use more cohesive devices and devices which show pragmatic sensitivity than less advanced learners.

Table 2.11: Results of a stepwise regression analysis

step/feature	Contribution to r^2 (%)
Number of words	37.6
Synonymy / antonymy	6.0
Word length factor	3.3
Passives	2.5
3rd person/impersonal pronouns	0.9

$F(1,157) = 11.92$; $p < .0001$; $r^2 = 0.503$

The results of the stepwise multiple regression in Table 2.11 show the features that were good predictors of holistic scores. Ferris concludes that since variety of lexical choices, syntactic constructions and cohesive devices seem to correlate with higher holistic scores, teaching should include more emphasis on these features and also student awareness of these features should be encouraged.

2.8.2 Comments

This analysis also clearly shows the correlation of essay length and other features with holistic ratings. However, there are some weaknesses in this kind of analysis where some features occur only rarely. The merits of discriminant analysis and multiple regression in automatic assessment are also considered.

One striking feature of this experiment was the influence of essay length on the multiple regression. It was the first factor to come out of a stepwise regression accounting for 37.6% of the variance out of 50.3% accounted for by the first five features. Despite this major influence, no mention is made of it by the author. Discussion and recommendations only involve the features that account for 6% or less of the variance.

This study highlights the dangers of using features that occur only infrequently. At this point, it is worth remembering that the occurrences in Table 2.10 are adjusted figures based on an average essay length of 225 words. This means that Group 1 occurrences are overstated by a factor of 225/187.43 and Group 2 occurrences understated by a factor of 225/251.38. For example, the real mean occurrence of participials for Group 1 is 0.42 x (187.43/225) = 0.35 and in Group 2 is 0.95 x (251.38/225) = 1.06. These means are based on count data which is discrete, i.e. taking only values 0, 1, 2 .... This means there is little scope for variation on the downside of these already low means. Scores for participials in Group 1 will be mostly 0 but will include many scores of one or more as well. Group 2 will have many scores of one, two or more but some of 0 as well. Therefore we are likely to see considerable overlap between the groups.

Low values may be less problematic for some kinds of feature than for others. Two possible types of feature are acquisitional and stylistic. Passives are an example of an acquisitional feature. Sometimes acquisition of a feature may give clues to proficiency. In that case, presence of such a feature could help in assessment. On the other hand, absence of an acquisitional feature from an essay does not tell us much about proficiency. This is because we do not know whether the student has not acquired the feature yet or just has not used it in this task. Stylistic features may also be useful in assessment if they are used to a different degree by students of different proficiencies. However, if these features occur rarely, they will be less useful because of the problem of reliability. If a feature occurs infrequently, the ranges of occurrence will be relatively high. If a feature is used twice in one essay and not at all in another essay, then in terms of reliability that is a huge difference. Reliability of variables is one of the biggest threats to validity of multiple regression analyses.

Regression and discriminant analyses assume that variables are normally distributed but rarely occurring features do not follow normal distributions. To illustrate this, consider again the case of participials for Group 1 as shown in Table 2.12.

Table 2.12: Statistics for participials in Group 1

number of students	60
mean number of participials (unadjusted measure)	0.35
standard deviation	0.09

In a normal distribution, 68% of points lie within +/- one standard deviation of the mean, 95% within two standard deviations and 99.7% within three standard deviations. But for participials, +/- 3 standard deviations gives us the range 0.08 to 0.62. Rather than 99.7% of values being in this range, incredibly none of the values are in this range and 100% are outside it. This is because this is discrete count data taking values 0,1,2,3....

Clearly, rarely occurring features can be problematic. In this study, nine of the eighteen significant features could be classified as rare with average values below five in either Group 1 or 2.

Multiple regression analysis is a useful approach since it allows various factors to be taken into account in determining scores. Any single lexical feature is unlikely to be able to predict holistic scores by itself. Therefore, one way to proceed is to consider several features together to strive toward a better predictor. Discriminant analysis allows possible predictor features to be identified and then multiple regression can help build a model to predict scores. However, caution is needed at this point. These features might only inform about scores in this case. If that is true, they are not actually predicting but describing. The real proof would be to try the model built from multiple regression on another set of essays to see how it performs.

In this study, the results of the regression show that five features including lexical and syntactic features account for over 50% of the variance of the holistic scores. Three of the features, essay length, synonymy and word length are lexical and account for 46.9%. These lexical features are probably applicable in other cases because of their general nature while the other two features, passives and 3rd person/impersonal pronouns seem more likely to be influenced by the context of this particular writing task.

This study develops a possible approach for assessing student compositions. A model using a number of features may be able to account for quality where one single feature may not. Essay length in conjunction with lexical and syntactic features seems to be able to account reasonably well for judgments of quality. The study again underlines the importance of essay length which is not only a good predictor of holistic scores but also more useful than rarely occurring features because of the reliability afforded by its magnitude. Infrequent occurrences of some syntactic and lexical features, while being useful indicators of acquisition, are problematic because of their rarity which may violate the normally distributed and reliability assumptions of the statistical models involved.

2.9 Laufer & Nation (1995)

2.9.1 Summary

This paper examines the vocabulary of written essays in terms of frequency in general English. Laufer & Nation describe the Lexical Frequency Profile and investigate its reliability and validity for assessing the vocabulary content of learner written productions.

The Lexical Frequency Profile (LFP) shows the percentage of words in a piece of writing belonging to different frequency levels. There are four frequency levels that can be used: the first 1000 most frequent words, the second 1000 most frequent words, the University Word List (UWL) (Nation, 1990) and any other words. Suppose an essay contains 200 word families and 150 belong to the first 1000 words, 20 to the second, 20 to the UWL and 10 are not in any list. Then the LFP of the essay is 75%-10%-10%-5%.

In practice, different LFP profiles may be constructed depending on the proficiency of the learners involved. For less proficient learners, three levels consisting of the 1000 most frequent words, the second 1000 most frequent words, and any other vocabulary may be suitable. However, for more advanced learners, three levels consisting of the second 1000 most frequent words, the UWL, and words not in any of these two lists nor in the 1000 most frequent words may be more appropriate. The authors argue that the LFP has advantages over other measures of lexical diversity which all have weaknesses.

An experiment was conducted to investigate the reliability and validity of the LFP. The subjects were 65 foreign learners of English split into three groups of varying proficiency. Group 1 was the lowest proficiency level and consisted of 22 learners studying in New Zealand. The other 43 learners were studying in the English department of an Israeli university. Group 2 was the middle proficiency level and consisted of twenty of these learners who were in their first semester. Group 3, the highest proficiency level, consisted of the other 23 learners who had finished two semesters.

Two essays of length 300-350 words were written by each group in the space of one week. The following essay was done by all subjects:
Should a government be allowed to limit the number of children a family can have? Discuss this idea considering basic human rights and the danger of population explosion.

For the second essay, subjects had a choice of three essay topics on other controversial issues. Subjects also took the active version of the Vocabulary Levels Test (VLT) (Laufer & Nation, 1999).

The completed essays were put into computer form and the following data processing was carried out:

- only the first 300 words were used.
- if a word was used incorrectly it was omitted.
- if a word was used correctly but misspelled, the error was corrected and counted as a word familiar to the student.
- a wrong derivative was not considered an error.
- proper nouns were omitted.

The LFP was calculated by recording the percentage of words in the four following categories: the 1000 most frequent words, the second 1000 most frequent words, the UWL, and words not in any of these three lists.

The following questions regarding validity were considered:

a. Is there a significant difference between LFPs of different language proficiency levels?
b. Does LFP correlate highly with scores on the active version of the Vocabulary Levels Test?

Mean percentages of words in the four categories are shown in Table 2.13.

Table 2.13: Mean percentages of word families at different frequency levels

	1st 1000		2nd1,000		UWL		Not in lists	
	Essay 1	Essay 2	Essay 1	Essay 2	Essay 1	Essay 2	Essay 1	Essay 2
Group 1	86.5	87.5	7.1	7.0	3.2	4.1	3.3	2.8
SD	3.8	5.3	2.0	2.3	1.8	2.5	2.3	1.8
Group 2	79.7	79.4	6.7	6.8	8.1	7.8	5.6	6.6
SD	5.3	4.5	1.7	2.2	2.3	2.3	3.5	3.3
Group 3	77.0	74.0	6.6	5.6	8.1	10.1	7.5	8.7
SD	6.1	5.9	2.6	2.5	3.2	2.9	2.9	3.5
F-test	19.35	33.1	0.29	1.89	24.86	27.40	10.46	22.74
p-value	.0001	.0001	.75	.16	.0001	.0001	.0001	.0001

In regard to the first question, an ANOVA revealed a significant difference between the groups in the percentage of the 1000 most frequent words. For the first essay, Group 1 was different to Groups 2 & 3, using more words of the highest frequency. For the second essay, all groups were significantly different, with Group 1 having the largest incidence of the highest frequency words and Group 3 the least. There was no significant difference between incidence of the second 1000 most frequent words. For UWL words, there was a significant difference for the first essay between Group 1 and the other groups with Group 1 using fewer words than the others. For the second essay, there was a significant difference between the three groups with Group 1 using the fewest words and Group 3 the most. The three groups were also significantly different in terms of use of *not in list* words for both compositions with Group 1 using the fewest and Group 3 the most.

To address the second question, LFPs were compared with composite scores on the active version of the VLT. The results of this comparison are shown in Table 2.14.

Table 2.14: Correlations between the LFP and the VLT

	1st 1,000		2nd 1,000		UWL		Not in lists	
	Essay 1	Essay 2	Essay 1	Essay 2	Essay 1	Essay 2	Essay 1	Essay 2
r	-.7	-.7	.01	.2	.7	.6	.6	.8
p	.0001	.0001	.9	.3	.0001	.001	.0002	.0001

Learners with higher scores on the VLT used more of the UWL and *not in list* words. There was a negative correlation between scores on the VLT and the number of 1000 most frequent words.

Two questions regarding reliability were considered as follows:

a. Do LFPs in two sets of essays by the same learners correlate highly with each other?
b. Do the percentage of words at each frequency level in the two sets of essays correlate highly?

A within-subject analysis was carried out on the two sets of essays. The analysis was done firstly for each word frequency level and then for proportions among the frequency levels. For Group 1, there were very few *not in lists* words and so they were added to UWL words before comparison. There were no significant differences for Groups 1 and 2 but Group 3 showed differences on the first 1000 words, UWL, and the proportions suggesting that the LFP might not be reliable for advanced learners. However, the first 1000 words include many function words and very basic lexical items and it may be argued they are not representative of a developed lexicon. When they are omitted from the analysis, there are no significant differences between the two essays.

Laufer & Nation conclude that the LFP is a reliable and valid measure of lexical use in writing. It provides stable results over pieces of writing by the same learner and discriminates between learners of different proficiencies. It correlates well with an independent measure of vocabulary knowledge. Another advantage is that the analysis can be done entirely by computer.

2.9.2 Comments

There are some obvious weaknesses in the analysis and interpretation of the LFP in this experiment. However, these problems should not detract from the usefulness of LFP as a research tool.

There seem to be some problems with the analysis which undermine some of the claims made about the reliability and validity of the LFP. The ANOVA on which the validity argument is based seems inappropriate given the nature of the data. One of the assumptions of ANOVA is that the groups are independent. This seems to be compromised by the profile percentage data which adds up to 100%. If there is a rise in one part of the LFP, there must be a fall somewhere else in response. It would have been safer to look for differences in one of the profile levels separately. The same can be said for the MANOVA in relation to the reliability question. This basic problem means that many of the claims remain unproven.

Various shortcomings of the LFP have also been found. This is partly because it has been taken up so readily by the research community and used in a lot of studies but also partly because the claims of this study were oversold. The following weaknesses in LFP have been found:

- Muncie (2002) finds that because the elements in the LFP are dependent on each other, they can change in unpredictable ways. For example, the ratio of above 2,000 words would not only change if an above 2,000-level word were added to an essay, but also if a below 2000-level word were removed.
- Laufer & Nation themselves report that LFPs for essays with less than 200 words may be unstable. However, many low level learner texts contain fewer than 200 words. At the other extreme, the LFP was found to be unreliable for advanced learners.
- Smith (2004) argues that the LFP may not be stable even for essays over 200 words. He also argues that LFP might be able to discriminate between groups of learners but is not sensitive enough to discriminate between individuals.
- Meara (2005) also raises doubts about the discriminatory properties of the LFP. Monte Carlo simulations suggest LFPs may differentiate between low learners and very advanced learners but the LFP may not be sensitive enough to discriminate reliably between groups with more modest differences in vocabulary size.
- Meara & Bell (2001) also argue that the low percentage of UWL and *not in list* words can be problematic especially in low level learners when the number may come close to zero.

The problems with this measure should not detract from its usefulness as a research tool. Its usefulness seems to stem from its difference to other measures of lexical diversity. Other measures are unitary measures in which a larger value usually involves improved performance.

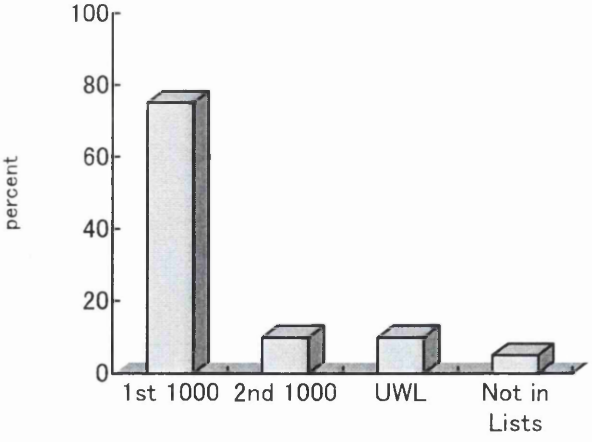


Figure 2.1: A graph of a 75%-10%-10%-5% LFP

Because the LFP as a profile has four elements, it gives more detailed information. It lends itself to graphical representation which may be an aid in interpretation. For example, a graph for an LFP of 75%-10%-10%-5% is shown in Figure 2.1.The LFP lends itself to vocabulary development research because more detailed changes may be spotted in profile or graph form than in a simple statistic. It has also been used in automatic scoring. Goodfellow, Jones & Lamy (2002) developed an automatic assessment system for L2 learners of French based on the LFP. They found that LFP values correlated significantly with scores awarded by human raters. They used the LFP profiles as a basis for feedback to learners. They believe it can form the basis of a system to help learners evaluate their own writing. Where it may not be so useful is in recognizing absolute differences between essays.

The LFP is a welcome addition to the range of measures of lexical diversity. In fact, it adds another dimension to this range because it evaluates diversity against a yardstick of the general frequency of words in English. However, the LFP needs to be used cautiously bearing its limitations in mind.

2.10 Meara & Bell (2001)

2.10.1 Summary

This paper by Meara & Bell (2001) describes P_Lex, another method to assess the lexical complexity of short L2 written essays by a comparison with frequency in general English. P_Lex is similar to the Lexical Frequency Profile but uses a more sophisticated mathematical methodology and the authors claim it works better with shorter texts than the LFP.

Most measures of lexical diversity or complexity depend only on the number and frequency of the words in the text but the nature of these words in terms of their frequency in English or their difficulty is not addressed. Meara & Bell call these *intrinsic* measures of lexical diversity. The counterparts to these measures are *extrinsic* measures which evaluate the type of words that are being produced. The LFP is one example of an extrinsic measure that categorizes the words learners use according to their frequency in general English.

However, Meara & Bell found the LFP to have some weaknesses. They note that a) it is not very reliable b) it has poor measurement properties c) it does not discriminate well between texts and d) low numbers of words in the lower frequency levels may lead to instability. Another major disadvantage of the LFP is that Laufer & Nation (1995) themselves found it unstable for shorter texts under 200 words.

P_Lex was developed as an alternative to the LFP with the aim of evaluating shorter L2 texts. The underlying assumption of the P_Lex process is that learners with large vocabularies are more likely to use infrequent words than learners with smaller vocabularies.

The method of P_Lex involves dividing an essay into ten-word segments and counting the number of *easy* and *difficult* words in each segment. *Easy* words are defined as words in the 1000 most frequent words plus proper nouns, numbers and geographical derivatives. Other words are classified as *difficult*. The number of these ten-word segments containing N *difficult* words is tallied. This distribution for different values of N tends to be skewed towards zero because the number of *difficult* words in each ten-word segment tends to be low. This distribution can be modeled using the Poisson distribution with mean lambda. Lambda values tend to vary from 0 to 4.5, but can be higher, with higher values indicating a higher use of infrequent words.

An experiment was performed to test the reliability and validity of P_Lex. Forty-nine L2 learners taking part in a university summer programme from a variety of L1 backgrounds ranging in proficiency from lower intermediate to advanced were involved. Each learner was asked to produce two pieces of written work of length 300 words within a week. The two essay topics were also used by Laufer & Nation (1995) as follows:

a) *Should a government be allowed to limit the number of children a family can have?*
b) *A person cannot be poor and happy.*

Each learner also took the active version of the Vocabulary Levels Test (VLT) (Laufer & Nation, 1999).

The shortest essay produced was of length 250 words and this was set as a standard and the first 250 words of each essay were used in the analysis. P_Lex was calculated for each essay for each learner. The reliability of P_Lex was then assessed by comparing scores on the two essays.

There was no significant difference between the mean P_Lex values for each essay. The P_Lex values for the two sets of essays correlated with r = 0.655. To test the discriminatory power of P_Lex, the learners were split into two proficiency groups based on the results of the VLT. T-tests showed that the P_Lex scores of these groups were different. P_Lex scores and scores on the VLT correlated significantly with r = 0.565 for the first essay and r = 0.339 for the second essay. To determine what length of essay was necessary for P_Lex to work properly, values were taken for essays of various lengths for learners of varying proficiency. P_Lex was found to be stable from about 120 words. Low-level learner essays seemed to stabilize at shorter essay lengths, from about 90 words. To reinforce this finding, the same reliability tests were carried out with only the first 150 words of each essay. The results were very similar to the earlier ones.

P_Lex produces data similar to the LFP but it is more reliable with shorter texts than the LFP particularly for low-level learners. However, there are problems with validation. Firstly, there are no tests of productive vocabulary with which to compare the data. Secondly, it is debatable whether difficulty can be defined purely in terms of frequency. Words may be unusual depending on the context of the writing task. Therefore it is suggested that P_Lex is used with a set of standardized tasks. Difficult words might be defined using the help of native speaker choices on the same tasks. Other applications for P_Lex are also suggested. One is in the evaluation of difficulty of examination texts. Another is the identification of students with abnormalities in writing, for example, an over-reliance on Romance cognates in their writing.

2.10.2 Comments

P_Lex is a useful new approach to measuring lexical diversity. In particular, it may be a better alternative to the LFP for evaluation of short low-level L2 texts. However, P_Lex may exhibit some of the same limitations as the LFP. Moreover, it is not clear that the P_Lex format is in fact superior to a simple ratio based on the same frequency levels.

One of the strengths of the LFP is that it charts vocabulary frequency over four bands. This makes it useful for detailed analyses. However, this is also a weakness when attempting to use it for comparison. In this sense, P_Lex is easier to use because it produces a single statistic. P_Lex values are therefore easier to use for comparisons of large numbers of essays or when comparing with other measures. For example, it is easier to correlate another measure with a unitary P_Lex value than a three value LFP. However, any measure that is based on two frequency bands may ultimately face a problem: If the learners doing a particular task do not vary sufficiently in different usage across words of the two bands then the measure may be unsuitable. This is equivalent to trying to differentiate students with a test that is either much too difficult or much too easy. With this in mind, P_Lex is probably more useful for lower level essays. LFP is probably more useful for higher level essays.

One unclear aspect is the so-called sophistication of P_Lex. Meara & Bell do not show that it is better than a simple ratio of difficult to easy words. P_Lex is certainly a more complex measure than a simple ratio but it is not clear that it is more sophisticated. The golden rule in statistical analysis is to only use a more complex analysis if it is superior to a simpler analysis. The reasons for this are twofold. Firstly, a more complex measure usually involves more assumptions which are often unrealistic and may limit its application. In this case, the fitting of a Poisson distribution involves an assumption that occurrences of rarer words are independent. This may be unrealistic. Secondly, a more complex measure is more difficult to interpret. In this case, it is quite easy to imagine possible limitations of a simple ratio based on frequency levels but not so easy to foresee the limitations of P_Lex.

Meara & Bell's categorization of intrinsic and extrinsic measures of lexical diversity is a useful one. There may well be advantages in using each type of measure. Extrinsic measures may be useful for comparative purposes and may be more sensitive and informative for particular tasks and contexts. However, this sensitivity seems to come at a price and care needs to be taken in selecting an appropriate task to elicit vocabulary at those levels. Also care needs to be taken that the choice of LFP or P_Lex is the best for the proficiency of the learners. The choice of LFP or P_Lex may also depend on the task and the learners involved.

In order to improve the performance of P_Lex, Meara & Bell suggest it be used on standardized tasks where words are evaluated against native speaker choices. It is an interesting idea which also reminds us to consider the validity of the benchmark data of any extrinsic measure. Both the LFP and P_Lex use the most frequent words in English. These are, in fact, the most frequent words across various contexts and genres of English. In reality, every combination of context and genre will have its own hierarchy of most likely used words. How well this set of likely words is represented by the general set of most frequent words may well determine how well these measures perform. For Japanese learners, the JACET word list (Ishikawa et al, 2003) might be a better alternative to base these measures on. The JACET list is based on words that Japanese learners are exposed to in their education system. It might also be that the 1000 word and 2000 word distinction is too arbitrary. Perhaps, with a standardized task, the easy words could include function words plus words commonly associated with that task. The difficult words could then be other words excluding proper nouns.

Assumptions for assessment are that there is some kind of progression in the measure that corresponds with increased proficiency. For the LFP, perhaps there is a general trend to the higher bands as proficiency increases. For P_Lex, the score should increase as proficiency increases. Some evidence of this sort may be a necessary first step in validating it for assessment purposes. With P_Lex, this validation might also involve an examination of the task itself. This would help to identify appropriate standardized tasks. There may be tasks that are inappropriate because they do not elicit enough difficult vocabulary even for high proficiency learners.

2.11 Meara, Jacobs & Rodgers (2002)

2.11.1 Summary

This paper introduces a technique called lexical signatures. Lexical signatures are distinctive patterns of lexical choices of L2 learners. Various properties of lexical signatures in essays of L2 learners of French are investigated and their application to evaluation of L2 writing is considered.

Lexical signatures aim to identify elements of lexical choice in the writing of L2 learners. Lexical choice is an element of writing style or genre. In the field of authorship attribution, statistical aspects of L1 lexical choice are used to identify a writer in solving problems of disputed authorship. This experiment aims to identify signatures of L2 learners which may be able to characterize learners in terms of proficiency.

Lexical signatures are investigated by taking small samples of words from a corpus of essays and looking at patterns of usage of these samples in each essay. In particular, unique patterns of these samples are of interest. In practice, these lexical signatures are derived in the following way:

- First, a small sample of words, say six, are sampled from the corpus. For example, the words *page*, *blanc*, *couleur*, *pas*, *office* and *comme*. These form a subset of words (*page*, *blanc*, *couleur*, *pas*, *office*, *comme*).
- Then, a binary representation of the subset is produced for each essay in the corpus. A one signifies presence of a word in the essay while zero signifies absence of the word. Accordingly, the binary representation (111000) of the (*page*, *blanc*, *couleur*, *pas*, *office*, *comme*) subset for a particular essay signifies that the essay contains the words *page*, *blanc*, and *couleur* but does not contain the words *pas*, *office*, or *comme*.
- Finally, the binary representations for each essay are compared. If an essay has a representation that is shared by no other essay, it is a unique representation or a unique lexical signature. The number of unique binary representations for each subset are counted.

In this experiment, lexical signatures were investigated in 59 essays written by L2 learners of

French. The task was a fifty-minute timed essay describing an advertisement for holidays in Scotland. The essays contained a total of 2272 different words. The distribution of word frequencies in essays showed that most words appeared in very few essays and just a few words appeared in many essays.

Three sets of fifty words each were selected, Set H, Set M and Set L. Set H contained high frequency words which appeared in between 32 and 50 essays, Set M contained middle frequency words appearing in 19 to 29 essays and Set L contained low frequency words appearing in 10 to 13 essays. For each set, 25 subsets of six words were selected randomly. The occurrence of each six-word subset was found for each essay and recorded as a lexical signature in binary form with one indicating presence of a word and zero indicating absence of a word in that essay. The number of distinct patterns and the number of unique patterns were tallied for each subset on each essay and then averaged for each subset over all essays. (Originally the authors planned for 64 essays which would allow for perfect distinctiveness in a six-digit subset since there are 64 permutations of a six-digit binary number. If every essay showed a different binary pattern for a subset there would be 64 distinctive patterns all of which would be unique.) The results of the analysis are shown in Table 2.15.

Table 2.15: Lexical signatures for subsets of high, medium and low frequency words

	mean distinct signatures	mean unique signatures	unique/distinct signature percentage
Set H	20.88	16.88	80.8
Set M	35.60	21.24	59.7
Set L	21.20	10.72	50.6

There was an average of 20.88 distinct signatures (out of a possible 59) for high frequency subsets, 35.6 for middle frequency subsets and 21.2 for low frequency subsets. There was an average of 16.88 unique signatures for high frequency subsets, 21.24 for middle frequency subsets and 10.72 for low frequency subsets. These averages masked considerable variability. For example, middle frequency subsets had an average of 35.6 distinct signatures but the highest number of distinct signatures produced by a subset was 41 distinct signatures while the lowest number produced was 32. The former can be called the best performing subset and the latter the worst performing subset. Similarly, for middle frequency subsets, the number of unique signatures varied from 29 to 16 with an average of 21.24.

Larger subsets of words showed, not surprisingly, both more distinct and unique signatures regardless of frequency of word as shown in Table 2.16.

Table 2.16: Distinct and unique signatures by subset size

	subset size (words)				
	six	seven	eight	nine	ten
Set H	20.88	28.88	41.76	45.80	48.88
Set M	35.60	43.56	49.72	54.40	55.12
Set L	21.20	26.20	31.04	36.32	39.68
	subset size (words)				
	six	seven	eight	nine	ten
Set H	16.88	23.84	32.28	37.84	42.00
Set M	21.24	31.80	42.28	50.48	51.80
Set L	10.72	15.64	20.48	26.68	30.40

For middle frequency words, an average of 35.6 distinct signatures for a subset of six words increased to an average of 43.56 signatures for a subset of seven words and then averages of 49.72, 54.40 and 55.12 for subsets of eight, nine and ten words respectively. Similarly, unique signatures increased as size of subset increased regardless of word frequency. For middle frequency words, an average of 21.24 unique signatures for a six-word subset increased to 31.8 for a seven-word subset and to 42.28, 50.48 and 51.8 for eight, nine and ten word subsets respectively.

The authors argue that an index of uniqueness based on lexical signatures may be useful in recognizing work of lower level L2 students. Lower level students might be expected to have a more limited vocabulary and show less unique signatures while higher level students may have more unique signatures.

2.11.2 Comments

Lexical signatures and the concept of uniqueness appear to offer a new dimension in the analysis of learner vocabulary. Looking at individual writer's style may also open the door to benefiting from the knowledge and methodology from the field of authorship attribution. Lexical signatures may have considerable potential in the evaluation of learner essays but the exact form and dynamic of a lexical signature approach needs more research.

Lexical signatures or uniqueness more generally seem to hold promise as a concept for analyzing learner writing. They offer a new dimension in analyzing learner writing in the form of style. Authorship attribution is the statistical analysis of L1 writer style with the aim of solving disputed cases of authorship. One feature of this field is the array of sophisticated statistical techniques that have been developed. Some of these techniques may have applications in exploring the writing style of L2 learners. It is also interesting to investigate how a measure of lexical uniqueness could work beside measures of lexical diversity and extrinsic lexical measures such as P_Lex and LFP.

It seems possible that lexical signatures could be helpful in assessing quality of essays. The underlying assumption is that less proficient learners have more limited lexical choices which may be picked up by a measure of uniqueness. There seem to be various ways one could measure uniqueness. For example, the simplest measure of uniqueness might be the number of words in an essay that are unique to that essay, i.e. not present in any other essay in the cohort. Lexical signatures offer a very interesting approach to uniqueness and possible enhanced properties compared to a simple measure of uniqueness. For example, lexical signatures investigate the uniqueness of co-occurrence relations of words. In this regard, this approach resembles that of latent semantic analysis but with a much smaller input of words.

Although this paper and an earlier paper, Meara, Rodgers & Jacobs (2000), lay out a case for uniqueness and its potential for assessing essays, there are many questions remaining about how this measure would work in practice. This experiment has clearly shown the power of lexical signatures to spot differences in essays. What remains to be shown, though, is that these patterns of lexical signatures do indeed relate to stylistic lexical choices rather than say arbitrary lexical choices. Also it remains to be shown exactly how lexical signatures could identify quality in essays. Another question relates to the complexity of lexical signatures. Where simpler measures of uniqueness exist, it is important to make sure that the complexity of lexical signatures has some justification in that it delivers more enhanced information than simpler measures of uniqueness. This may mean proving that lexical signatures are superior to simpler measures of uniqueness. There may also be technical concerns related to the complexity of lexical signatures. In general, a complex measure has less validity than a simple one because it is more difficult to understand how it works and interpret what results mean.

The lexical signature approach certainly warrants more research. However, given its complexity, results need to be interpreted with caution and in tandem with consideration of other simpler and more transparent measures of uniqueness.

2.12 Discussion

The experiments in this chapter have set out a case for automatic assessment for L2 learners. Page (1994) shows clearly that automatic assessment is not only a convenient and cheap alternative but also can improve reliability. This discussion section looks at two vital elements of an automatic scoring system: the features they involve and how they are put together. The field of authorship attribution is also considered as a potential source of both features and methodology.

2.12.1 Features

The most challenging aspect of this research is to identify features of essays that can best assess essay quality. There has been considerable research on the relationship of certain features with human ratings of essays. Two of the most studied features are essay length and measures of lexical diversity.

2.12.1.1 Essay length

Essay length has shown itself to be one of the best predictors of holistic scores in numerous studies. Both Larsen-Freeman & Strom (1977) and Ferris (1994) find essay length to have a major relationship with essay quality. The scoring systems developed by PEG and e-rater both involve direct or indirect influences of essay length. PEG employs a measure of the fourth root of essay length and e-rater involves various measures which are correlated with essay length.

Ferris (1994) found that out of 28 features, essay length was the one that most correlated with quality ratings. It contributed 37.6% of the variance in a stepwise regression analysis with the next most influential feature only contributing 6%. When dealing with low-level learners and timed tasks, the importance of essay length is likely to be emphasized. McNeill (2006), in a study of lexical features and quality in essays of Japanese college level EFL learners, found essay length to be the most influential indicator of holistic human ratings for a variety of written tasks both timed and non-timed. Essay length was measured as part of a set of indicators that included various measures of lexical diversity, error counts and T-unit counts.

McNeill also found evidence to support Page's assertion that essay length is more important in shorter essays. Longer essays correlated less with holistic scores than shorter essays and McNeill suggests a threshold of 400 words after which essay length becomes less important. Other studies back this up. Linnarud (1975), in a study of 36 essays written by Swedish learners of English, found that short essays tended to be rated as poor. However, the essay length and quality relationship did not hold for longer essays. Longer essays were often essays rated as average rather than essays rated as good. Page also reports a decreasing influence of essay length for longer essays which prompted the choice of *fourth root* of essay length as a feature.

Most of these studies involved timed tasks but Larsen-Freeman & Strom found that even with non-timed tasks, essay length was a strong predictor of proficiency.

There is contradictory evidence from other researchers though. For example, Perkins (1980), in a study of advanced learners from a variety of L1 backgrounds, found there was no significant difference between three proficiency groups in terms of essay length on a fifty-minute timed essay. However, in this experiment the learners were informed in the rubric that they would be scored on both essay quality and essay length. This could be a contributory factor as it seems to set this study apart from others where no specific mention of essay length influencing ratings was made to participants. The fact that these learners were advanced learners may also be significant.

The influence of essay length is likely to depend on the context and the learners. For lower level learners, it is likely to correspond positively with quality. For more advanced learners, its impact may be less. Also, it is more likely to impact on timed tasks rather than non-timed tasks.

2.12.1.2 Lexical diversity

Lexical diversity has been used widely in applied linguistics in a variety of contexts. It has several variants. There seem to be no naming conventions so sometimes the same formulation is used with different names and sometimes different formula are used with the same name. Although the naming of these measures may be confusing, measures of lexical diversity do provide a flexible way to analyze the lexical content of essays. Formula can be adjusted to fit the particular needs of the study as in Arnaud (1984). The flipside of this is that studies become difficult to compare. Of many measures of lexical diversity, the type token ratio (TTR) is perhaps the most often used. TTR is simply the number of different words (types) divided by the total number of words (tokens) as follows:

$$\text{TTR} = \text{types} / \text{tokens}$$

Other measures include lexical density, lexical variation, lexical rareness, and lexical originality. Lexical density (LD) is the number of lexical tokens divided by the total number of tokens as follows:

$$\text{LD} = \text{lexical tokens} / \text{tokens}$$

Lexical variation (LV) is the number of lexical types divided by the number of lexical tokens as follows:

LV = lexical types / lexical tokens

Lexical rareness (LR) used by Arnaud (1984) is the number of rare types divided by total types as follows:

LR = rare types / types

Lexical originality is the number of words in an essay not appearing in any other essays in the set of essays.

Laufer & Nation (1995) point out the following weaknesses of some of these lexical diversity measures:

- Lexical originality tends to be unreliable because it depends on the structure of the learner group.
- Lexical density is influenced by function words.
- Lexical sophistication may be constructed in several ways.
- Lexical variation is sensitive to essay length, dependent on the definition of a word, and unable to distinguish between words of different frequency levels.

Engber (1995) found that lexical variation measures adjusted for error were better at predicting essay quality than lexical diversity or error measures alone. Arnaud also proposed a more complex formula for lexical diversity which includes elements of lexical variation, lexical rarity and error.

One disadvantage measures of lexical diversity all share is they are affected by essay length. As more words are included, these measures tend to fall. Therefore, it is difficult to compare essays of differing lengths because, in effect, the longer essays will be penalized. Given that essay length is often correlated with quality, this is a major weakness of these measures in essay scoring. Even the output of a given individual can exhibit very different lexical diversity scores according to the size of the sample of output. In relation to TTR, this basic problem has long been recognized and various mathematical transformations of it have been proposed to alleviate the length effect. For example, Guiraud Index (Guiraud, 1960) is the number of types divided by the square root of the number of tokens as follows:

Guiraud Index = types / $\sqrt{}$tokens

However, the transformations have usually been found to also depend on essay length but perhaps to a lesser degree than TTR. For example, Malvern, Richards, Chipere & Durán (2004) found that Guiraud Index typically rises along with length for the first few hundred words of a text but then gradually decreases with increasing length after that.

Yule's K (Yule, 1944) which is often used in the field of authorship attribution is a measure that was developed to be independent of text length but research suggests it is only stable for texts of 1000 words or more. This makes it useful in authorship attribution where texts under consideration are often of book length but less useful for L2 essay scoring where essays may be only a few hundred words long.

One way to get around the problem of essay length is to standardize it by sampling a fixed number of words. Although this solves the basic mathematical problem, it is a less than perfect solution. It often means the shortest essay in a set becomes the standard and longer essays will be shortened and so much output wasted. It is also possible that drastic shortening of a longer essay may cause impairment in some way.

Malvern & Richards (1997) came up with an innovative alternative measure which they claim is independent of essay length particularly for short essays. They use the basic concept of TTR but they exploit the fact that all essays share the property of a curved graph when TTR is plotted against increasing essay length. By isolating a coefficient of this curve, a general measure for comparison emerges. They use a simplification of a formula developed by Sichel (1986) to produce a model for TTR. This simplified formula is as follows:

$$\text{TTR} = \frac{D}{N}\left[\left\{1+2\frac{N}{D}\right\}^{\frac{1}{2}}-1\right]$$

where N is number of tokens in the essay.

By using samples of writing, the parameter D can be estimated and provides a basis for calculating a measure of diversity. The calculation is complex but there is a computer package, VOCD, which produces a D estimate (Malvern et al., 2004). Using VOCD, D is estimated by taking many random samples of 35-50 words without replacement, calculating D for each one and then taking the average value of D as an estimate of D for the essay. Even though many small samples of words are used, this method has the advantage that it incorporates all of the words in the essay. The D estimate may be the best solution to the influence of essay length at the moment. However, it seems that the search for a perfect measure of lexical diversity may not be over yet. Recently, McCarthy & Jarvis (2007) showed that the D estimate is also affected by essay length in some situations. They found that the D estimate was still significantly correlated with essay length for essays over 400 words. However, many L2 essays fall into the range covered by the D estimate. Yu (2007) found that D scores based on essays written for MELAB tests (Michigan English Language Assessment Battery) correlated with human ratings significantly with r = 0.29. D worked even better for spoken tests with D calculated for spoken interview transcripts correlating with spoken test assessments with r = 0.48.

For low level learners, essay length is likely to fall in the range where the D estimate is not influenced by essay length so it offers a good alternative to predict lexical diversity. However, this practical way to control for essay length comes with a price. The calculation is complicated and needs a special computer package to handle it.

Meara & Bell (2001) refer to measures such as TTR, the D estimate or lexical variation as intrinsic measures of lexical diversity. These are mostly concerned with counts of vocabulary, either tokens or types. No attention is paid to the nature of the words themselves. Meara & Bell provide a good illustration of how these measures fail to identify important differences in the sophistication of vocabulary used. Each of the following sentences would score the same for an intrinsic measure of lexical diversity such as the TTR yet the latter two clearly involve more sophisticated vocabulary.

The man saw the woman.
The bishop observed the actress.
The magistrate sentenced the burglar.

Therefore they propose that extrinsic measures of lexical diversity such as LFP (Laufer and Nation, 1995) and P_Lex (Meara & Bell, 2001) might also be used. These measure lexical diversity in relation to outside measures of frequency in general English. These extrinsic measures would distinguish the examples above as having words from different frequency ranges. Measures such as lexical rareness (Arnaud, 1984) also qualify as extrinsic measures.

Daller, van Hout & Treffers-Daller (2003) propose using hybrid measures of lexical diversity which offer the same ease of calculation as the familiar intrinsic measures of lexical diversity but measured against frequency as in extrinsic measures. They experiment with two measures: advanced TTR (A_{TTR}) and Advanced Guiraud (A_G) as follows:

$$A_{TTR} = {}^{V_a}\!/_{N}$$

$$A_G = {}^{V_a}\!/_{\sqrt{N}}$$

where v_a is the number of advanced types and N is the total number of tokens.

Advanced types could be defined according the context. Daller, van Hout & Treffers-Daller were examining use of German and Turkish in bilinguals. For German, they defined advanced types as being outside 2000 basic words as defined by a well known word list. For Turkish, they depended on teacher judgments for a classification of vocabulary into basic and advanced.

It might be interesting to see if intrinsic measures and extrinsic measures could be used together. In addition, the lexical signature approach of Meara et al. (2002) adds another possible dimension to this model: the comparison of lexis produced by a learner with that produced by other learners in a group.

2.12.1.3 Other considerations

Selection of features depends on other considerations. One is validity. Validity will ultimately depend on an assessment system reliably predicting quality of essays. However, validity of features or methods can be a concern. Features can be selected according to statistical effectiveness at discriminating between candidates or with validity in mind. The PEG system introduces validity at an early stage by selecting features that are proxies for underlying traits in writing. Earlier versions of e-rater, on the other hand, seemed to be based on a more flexible approach. Initially, a large array of features were input into the system. The best eight to ten features were then selected for each model according to statistical efficiency. This left them open to criticism for the nature of these choices as shown by Sheehan's (2001) observation that many features were heavily influenced by essay length. The latest version of e-rater seems to be based on a much smaller and balanced set of features (Lee, Gentile & Kantor, 2008).

Using the IEA system or features such as LFP or lexical signatures offers validity based on a language model. IEA is based on an LSA based model of semantic relations and lexical acquisition. LFP and lexical signatures are based on the model of lexical learning according to frequency in general English.

Another consideration when selecting features is reliability. Reliability can be threatened by features that occur infrequently. Features that depend on frequent occurrences are likely to be more reliable. Essay length and measures of lexical diversity for instance depend on large counts which enhance their reliability. Examination of individual words or infrequent grammatical features will produce smaller counts which are less likely to be reliable. This was evident in the Ferris experiment where many of the features occurred only a few times at most in an essay.

2.12.2 Automatic scoring methods

The success of packages like PEG, e-rater and IEA show that automatic assessment of essays is a realistic objective. Despite the success of these packages, they are not suited to small scale practical applications in L2 classrooms. They all share one disadvantage. They require some sort of training mechanism external to the set of essays. PEG and e-rater systems require a set of essays to be rated by humans. IEA is slightly more flexible and requires a rated training set, a model answer or expert material. However, Landauer, Laham & Foltz (2003) conducted an experiment where essays were rated internally by relative distance from each other. This seems to be an approach worth investigating which might help realize an essay scoring system in a more ready to use format.

Some of the other papers developed ideas that could be incorporated in essay scoring. Ferris used discriminant analysis and regression analysis to identify features in a training set of essays. These features could then be used to rate other essays. Goodfellow, Jones & Lamy (2002) found the LFP to be a good way to automatically assess L2 learners of French for giving feedback about vocabulary usage. Meara, Rodgers & Jacobs (2000) proposed using a neural network with lexical signatures to classify essays as higher quality or lower quality.

Some of these scoring systems need to incorporate features. These need to be easily calculated. Some potentially useful features can be problematic. For example, Engber highlights the role of error in human ratings of essays. Unfortunately error is a difficult thing to handle efficiently in a simple computer package. Error requires skilled human judgment to determine the type of error or seriousness of error. The number of error free T-units was found by Larsen-Freeman & Strom to be a better discriminator of essay quality than essay length or average length of T-unit. However, both determining a T-unit and judging error are difficult to do automatically.

Some features are easy to measure such as counts. Others cannot easily be calculated but a good estimate can be made, for example, the number of sentences in an essay. Other features are more difficult to estimate such as error or a T-unit. One alternative approach is to find proxy measures as in the PEG study.

No matter how good any one feature or measure is, it is likely to be inadequate by itself for assessment purposes. To confidently predict essay quality, a range of measures may be necessary. PEG, e-rater and the approach of Ferris all showed some success by using models with numerous features.

2.12.3 Measures from authorship attribution

Authorship attribution has a long history and has developed a varied methodology. Some of this research may also have applications in L2 assessment research. Appreciating the similarities and differences between the two fields can help judge which features could be applied productively in L2 research.

There are some obvious differences between authorship attribution research and L2 assessment but there are also some similarities. Authorship attribution typically deals with long or multiple texts by the same author whereas in L2 assessment the texts are very short often only a few hundred words or fewer. Longer texts enable rare events to be investigated such as the rate of occurrence of particular words. This may not be possible with shorter texts. Moreover, some of the statistics developed for authorship attribution are designed for large sets of data. For example, Yule's K (Yule, 1944) is independent of text length but is probably only stable for texts longer than 1000 words or so. Context of writing is an important difference too. Authorship attribution usually cannot control for context. Samples from authors under consideration are likely to be from a variety of contexts and genres. Since context and genre can affect vocabulary choice, this can be a major problem for authorship attribution studies. In L2 assessment, very often learners are writing on a common theme. This makes it easier to make comparisons of vocabulary use.

Another major difference between the two fields is the underlying assumptions. In authorship attribution, a core assumption is that individual writers exhibit individual characteristics in their writings. These individual characteristics can be used to identify the writers. Different works by the same author should display similar characteristics while works by different authors should show differences in these characteristics. In L2 assessment, the assumption is that although learners may show individual characteristics, these will be outweighed by characteristics typical of their level of proficiency.

The two fields also have some things in common. Authorship attribution is typically about identifying the author of a document of unknown authorship from a set of possible candidates. It is therefore a type of classification problem. Assessment situations also often involve classification of an essay into one of a set of categories. The categories may be grades A to E, or a pass/fail distinction or placement into groups of various proficiencies. Classification often involves identifying similarities between members of a group and differences from members of other groups. Some measures may be reliable for individuals over different pieces of work but do not necessarily spot differences in L2 proficiency so clearly. This suggests that they are more use in authorship attribution than in L2 assessment.

The nature of the classification is different in one important respect between the two fields. In authorship attribution qualitative judgments usually do not need to be made, similarities and differences are important. However, in assessment, we are usually interested in qualitative differences. Therefore, features that offer differences according to proficiency are likely to be more useful in L2 assessment. For example, one possibility could be mean word length. If we assume that more proficient learners are more likely to use infrequent words, then we can expect mean word length to increase with proficiency. Frequent words are likely to be shorter than infrequent words. For example, the mean lengths of headwords in the 1K, 2K and UWL lists are 5.25, 6.57 and 7.29 respectively. Presence and absence of some features may also give clues to proficiency. Some features may be absent from a certain level because they have not been acquired yet. Conversely, the presence of a feature might suggest a certain proficiency. Of course, these features may not necessarily have a linear relationship with proficiency. A large number of occurrences may indicate an overuse of a feature.

Another aspect of authorship attribution that could be instructive to L2 research is the methodology. In authorship attribution, the methodology involves accumulating a body of evidence to support a theory. Each piece individually may not be conclusive but when seen together, there may be a compelling body of evidence. This helps explain the wide range of features and methods used. A description of some features and methods used in author attribution can be found in Holmes (1994).

2.13 Conclusion

This chapter has highlighted some of the main empirical work on methods and features that can help predict quality in learner essays. This work can help guide a search for an assessment system for low level learner essays as well as illustrate some problems and potential pitfalls.

Automatic assessment already is a reality in L1 contexts and in some L2 contexts such as the TOEFL Test of Written English. In this chapter, a few of the main systems have been considered. The PEG project and ETS's e-rater are based on multiple regression models using a variety of lexical, grammatical and discourse features. Many of the features used could be applied to L2 essays. However, it is not clear on what scale these techniques are able to operate. Ferris had some success with a relatively small set of 160 essays. However, this still involves more essays than many practical assessment situations. Another problem with these systems relates to a lack of flexibility incurred by the requirement of training samples.

The IEA system based on latent semantic analysis offers an alternative approach. This is based on semantic content rather than essay features. With L1 essays, this system seems to be effective because it informs about conceptual content of essays. Rating of L2 essays often depend less on conceptual structure than on language content so it is not clear how effective this type of system could be in L2. However, there is an argument that it might tell us something about the organization of the lexicon. In this case, it may complement an analysis based on essay features. The IEA research also hints at a possible solution to the training sample problem. In one experiment, essays were evaluated by their proximity to other essays in the set rather than to expert material. Reliability was relatively high if not as high as with expert material. It would be interesting to see if feature information as well as semantic distance could be incorporated within the same framework and if the essays could themselves provide a benchmark of quality. One challenge is to exploit some of this technology to make more flexible applications that can be used for smaller scale L2 classroom and research applications.

In terms of features, various measures of lexical diversity, both intrinsic and extrinsic have been considered. The consensus seems to be that there is no magic feature to account for essay quality. A great deal of research has been done with lexical diversity. More accurate measures such as Malvern & Richards' D estimate have been produced and have overcome some of the main technical problems such as the dependence on essay length. However, the relationship with essay quality is unclear and seems to depend on contextual considerations. There also needs to be more research on how different lexical diversity measures can complement each other. Extrinsic measures such as LFP and P_Lex add another dimension to lexical diversity by measuring lexical use against lexical frequency in general English. Measures such as unique lexical signatures also have the potential to inform about lexical usage relative to other members of the group. Lexical signatures also seem to provide information about semantic relations of small sets of words.

In the following chapters, the experimental work is presented. In the first of these, the lexical signature approach of Meara, Jacobs and Rodgers (2002) is investigated in more detail to determine whether it can help in essay assessment.

Chapter Three: Lexical Signatures

3.1 Introduction

This first experimental chapter describes two experiments exploring lexical choices of L2 learners using the lexical signature approach of Meara, Jacobs and Rodgers (2002). In the first experiment, a simple measure of uniqueness based on lexical signatures is created and applied to essays written by low level Japanese learners of English. In the second experiment, a more in-depth examination of some of the properties of lexical signatures is undertaken. In particular, the stability and reliability of the uniqueness measure is investigated. These experiments will help evaluate whether lexical signatures offer an alternative approach to automatic essay assessment either as a holistic method or as one element in an array of features.

3.2 Lexical signatures in low level learners

3.2.1 Introduction

Meara, Jacobs & Rodgers (2002) investigated lexical signatures, patterns of lexical choice, in essays produced by L2 learners of French. Inspiration came from the field of authorship attribution where lexical indicators of individual writing style have often helped in identifying authorship in disputed cases. Meara et al. investigated whether low level L2 learners might also make idiosyncratic lexical choices and whether, in turn, these choices might help identify their proficiency. Lower level learners can be expected to have a relatively small vocabulary which may manifest itself in a smaller range of individual lexical choices. If this is the case, lexical signatures could have a role in computer-aided assessment of L2 essays. In Meara, Rodgers & Jacobs (2000), the same authors proposed using lexical signatures in conjunction with neural network models to assign grades to essays according to quality. In Meara et al. (2002), lexical signatures using small subsets of words were used to investigate lexical choices of learners. These small subsets revealed a surprisingly large amount of variation in individual usage of words.

3.2.1.1 Aims of the experiment

In this study, a number of essays produced by low level Japanese learners of English as an L2 are analyzed using lexical signatures. The aims of the study are to:

1) see if Japanese L2 learners of English exhibit a similar amount of individuality in their lexical choices to the L2 learners of French in the original study.

2) create a simple index of uniqueness based on lexical signatures.

3) test the measurement reliability of this index of uniqueness.

3.2.2 Methodology

3.2.2.1 Participants and task

A set of 77 written essays was collected from a group of first-year Japanese university students. These students were all non-English majors and their English proficiency was relatively low. During their first year of study at university, these students have five compulsory 90-minute English classes a week. The students were asked to write a story based on a six-caption cartoon. The cartoon used is Cartoon Task 1 in Appendix 3.1. Students were given 25 minutes to complete the task. For the analysis, essays with fewer than 100 words were removed and 64 of the remaining essays were included.

A specially designed computer program, L-unique (see Appendix 1.1) was used to analyze the data. A single corpus was constructed from the 64 essays. This corpus was found to contain a total of 745 different unlemmatised words. Of the total 745 words, 371 (49.8%) appeared in only one essay each. Furthermore, 634 words (85.1%) appeared in fewer than ten of the 64 essays. There was only one word, *and*, that appeared in all 64 essays. Some sample essays can be seen in Appendix 3.3.

3.2.2.2 Measuring lexical signatures

For the analysis, thirty words were chosen that appeared in around half the essays. Words appearing in about half the essays had been shown in Meara et al. (2002) to be the most effective for producing unique signatures. In the original study, fifty words were selected, but given the smaller number of words in this corpus compared to the original (745 words in this study compared to 2272 in the original) there were a dearth of words that appeared in close to half of the essays. The number of essays in which these thirty words occurred varied from 23 to 41 essays. The words selected are shown in Table 3.1.

Table 3.1: Words selected for analysis

back boy came clothes day father for go good home I it my off one said sea so take that then there they this thought threw throw too took were

From this sample of thirty words, 25 subsets of six words were selected randomly. Meara et al. also used six-word subsets. Six words is an appealing choice to match the number of essays in this experiment. The number of different possible permutations of a six-word subset is 2^6 or 64. Matching the number of essays to the number of possible permutations of a six-word subset allows for the possibility for each essay to have a unique signature for a particular subset. This number of subsets necessitated each word being used multiple times. Some examples of subsets selected are shown in Table 3.2.

Table 3.2: Four examples of subsets

1	*came, my, see, clothes, threw, it*
2	*took, there, thought, home, back, came*
3	*good, it, were, off, this, then,*
4	*too, this, home, so, I, father*

For each subset, binary representations of the 64 essays were derived. For each word in a subset, a *1* indicated the word was present in that essay while a *0* indicated that the word was absent. So, for example, in the case of subset (*came, my, sea, clothes, threw, it*), a binary representation of (*100110*) would indicate that the words *came, clothes* and *threw* were present in that essay but that the words *my, sea* and *it* were absent. Table 3.3 shows the binary representations for each of 64 essays for the subset (*came, my, see, clothes, threw, it*).

Table 3.3: Binary representations of essays for (*came, my, sea, clothes, threw, it*)

000000	001000	010000	011000	100001	101100	110100	111011
000001	001001	010001	011001	100001	101101	110101	111011
000100	001001	010001	011010	100001	101111	110101	111011
000100	001010	010010	011011	100011	101111	110111	111011
000100	001011	010010	011011	100011	101111	111001	111011
000101	001100	010101	011100	100100	110001	111001	111100
000110	001101	011000	011101	100111	110001	111010	111101
001000	001101	011000	100000	100111	110001	111011	111101

From Table 3.3, we can see that one essay produced the representation 000000. This representation indicates this essay included none of the six words in the subset. On the other hand, no essays produced the representation 111111, so no essays contained all of the words in the subset. A six-digit binary subset can produce 64 combinations of representations. Amongst the 64 essays, there were 38 different representations of this subset which means that a further 26 possible representations did not occur. Of the 38 representations that did occur, 21 occurred in only one essay each. These representations that occur in only one essay were coined *unique lexical signatures* by Meara et al. For example, 000000 was a unique lexical signature as it appeared in only one essay. Similarly, 111100 was a unique lexical signature meaning that there was only one essay that included the words *came, my, sea,* and *clothes* but did not contain *threw* and *it*. The 21 unique lexical signatures for the subset (*came, my, sea, clothes, threw, it*) are shown in Table 3.4.

Table 3.4: Unique signatures for subset (*came, my, sea, clothes, threw, it*)

000000		010000			101100	110100	
000001			011001		101101		
			011010				
	001010					110111	
	001011						
000101	001100	010101	011100	100100			111100
000110			011101			111010	
			100000				

In the next stage of the analysis, the average number of different and unique representations in the 64 essays per subset was calculated for each of the 25 subsets.

Meara et al. noted that the number of different and unique signatures can be increased by increasing the subset size. While a six-word subset can potentially yield 64 different signatures, an eight-word subset affords 256 different signatures and a ten-word subset, 1024 signatures. The analysis was repeated for larger subsets of eight words and ten words.

3.2.2.3 Constructing an index of uniqueness

To create a simple index of uniqueness, the six-word subsets were used. The number of unique signatures was calculated for each essay. Since there were 25 subsets, the value of this uniqueness index could vary from zero to 25. A score of zero would mean that the essay had no unique lexical signatures for any of the 25 subsets, whereas a value of 25 would indicate that the essay had a unique lexical signature for every subset.

3.2.2.4 Measurement reliability

A final point worth investigating is the reliability of this estimate of uniqueness. This is a crucial point if we are thinking of using this process in measurement. Usually when we measure something, we do not need to worry about reliability of the measuring instrument itself. But because this index of uniqueness depends on sampled subsets, we need to find out how this value might vary with other possible sampled subsets. A simple way to do this is to do the calculations on different samplings and compare the results. In order to do this, the analysis of six-word subsets was run ten times and the uniqueness index scores for all essays on each iteration were calculated. The index of uniqueness scores for each iteration were then correlated against every other iteration to give 45 correlation coefficients.

3.2.3 Results

3.2.3.1 Lexical signatures

The average number of different signatures and unique signatures for each size of subset and the number of unique signatures for the best performing subset are shown in Table 3.5. The best performing subset is the subset that produces the most unique signatures.

Table 3.5: Statistics for subsets of various size

	subset size (words)		
	six	eight	ten
average number of different signatures	39.16	54.84	61.92
average number of unique signatures	22.00	46.64	60.04
best performance of unique signatures	30	56	64

A six-word subset, on average, uniquely identified 22 of the 64 essays. As in the original study, it can be seen that the number of unique signatures increased considerably as the number of words in the subset was increased. The average number of unique signatures increased from 22 for a six-word subset to 46.6 for an eight-word subset to more than 60 for a ten-word subset. Also, the best performing ten-word subset could uniquely identify every essay. The results for six-word subsets are compared with those of the original study by Meara et al. in Table 3.6.

Table 3.6: Comparison of six-word subset to the original study

	Set M (Meara et al)	This study
average number of different signatures	35.60	39.16
average number of unique signatures	21.24	22.00
ratio of unique to different signatures	.597	.562

The average number of different and unique signatures was slightly higher in this study than in the original, but that may reflect the fewer essays in the original study, 59 compared to 64 in this study. The ratio of unique to different signatures was slightly lower than the original study. These results suggest that this technique works as well for these learners as it did for the original learners of French despite the much smaller corpus of words.

3.2.3.2 Index of uniqueness

The numbers of unique signatures for each essay based on the six-word subsets are shown in Table 3.7.

Table 3.7: Summary of uniqueness index scores for 64 essays

	Uniqueness index
mean score	8.4
standard deviation	2.8
high score	15
low score	3

The highest score on the uniqueness index was 15 and the lowest score 3 with an average score of 8.4. This suggests that the uniqueness index based on six-word subsets gives a good range of values out of the possible range from zero to 25.

3.2.3.3 Analysis of measurement reliability

Table 3.8 shows the results of the reliability analysis.

Table 3.8: Average, highest and lowest reliability coefficients for ten iterations

	mean	high	low
reliability (r)	0.32	0.54	0.12

The average reliability between scores from two measurements using different subsets is only 0.32 with a range from 0.12 to 0.54. This seems very low. This shows that a different set of 25 six-word subsets from the sample of thirty words may lead to very different scores on the uniqueness index.

3.2.4 Discussion

On the whole, the results of this study reinforce the results of the original study. Even with very low level learners, a very small subset of words is capable of uniquely identifying most of the essays. Some ten-word subsets selected from words appearing in about half the essays can uniquely identify all the essays. An approach that with so few words can discriminate between 64 essays is obviously very powerful and has exciting possibilities if it can be applied to assessment of essays. This set of learners produced a much smaller corpus than the learners in the original study which suggests that the learners are of a lower proficiency. However, the amount of lexical variation was comparable over the two studies based on the number of different and unique six-word subsets.

A simple index of uniqueness was constructed by simply calculating the number of unique signatures for each essay out of a possible 25 subsets. Unfortunately, the measurement reliability for this index was very low. If this index is to be developed for measurement purposes, the measurement reliability needs to be improved dramatically. For most measurements, this type of reliability can be expected to approach one. For example, if you make the same calculation on a calculator a number of times, you would be very dissatisfied to find it give different answers. In fact, if you did get different answers, you would probably presume the error was due to the operator. In this experiment, the low reliability is a sampling problem. It occurs because there are a surprisingly large number of ways to select six words from thirty. The number of subsets sampled, 25, is too small to be representative of this large number of potential samples (see Appendix 3.4 for mathematical background to combinations). The number of different six-word subsets possible from a set of thirty words is 593775 which means a sampling rate in this experiment of 0.0042%. Clearly the number of subsets sampled needs to be increased dramatically. For example, an increase to a sampling rate of 1% would mean sampling 7143 subsets instead of the 25 sampled in this experiment.

3.2.5 Conclusion

This study supports the findings of Meara et al. (2002). It seems that even very low learners exhibit a lot of unique lexical choices as evidenced by occurrences of unique lexical signatures. A simple index of uniqueness seems to identify differences in lexical choices in essays. However, this index was not reliable over different sampled subsets because of sampling problems. Twenty-five subsets is a tiny sampling proportion so the number of subsets needs to be increased substantially. In the next experiment, the number of sampled subsets needs to be increased to improve reliability. In addition, reliability in the general sense needs to be tested by seeing if the measure provides a similar score for essays of similar quality. Finally, it needs to be ascertained whether the index of uniqueness can identify differences in essay quality.

3.3 Investigating an index of uniqueness

3.3.1 Introduction

This experiment investigates some of the questions raised by the previous experiment. In particular, it aims to develop a reliable index of uniqueness based on the lexical signature technique and to determine how this measure is associated with quality in written tasks.

One of the basic aims of this study is to establish reliability. There are two aspects to this. The first is measurement reliability. Firstly, the index generated for a given piece of written work must be consistent for different choices of parameter. This can be called measurement reliability. In the previous experiment, the index developed was found to vary considerably when only 25 random subsets were employed. This was due to the degree of sampling. Twenty five subsets was only a tiny proportion of all possible six-word subsets selected from thirty words. It is evident that a much larger number of subsets is necessary to provide consistency.

In addition, reliability in its more usual sense needs to be considered. Does the measure show consistency over different pieces of work by the same learner? This is vital if this uniqueness index is to be used to evaluate the lexical content of writing. This is equivalent to test-retest reliability and is easily tested by correlating values derived from two pieces of work from the same learners.

3.3.1.1 Aims of the experiment

The three basic aims of this study are to:

1) develop an index of uniqueness based on unique lexical signatures that has measurement reliability.

2) investigate the reliability of this index of uniqueness over two written tasks by the same learners.

3) explore the relationship between the index of uniqueness and quality assessments of the written essays.

3.3.2 Methodology

3.3.2.1 Participants and tasks

The participants in this study were the same first year Japanese university students as in the previous experiment. For this experiment, the same written essays from the previous experiment were used. In addition, the students completed a second similar written task towards the end of their first year with a three-month interval between tasks. Altogether, 77 students completed the first assignment and 76 students the second assignment. There were 74 students that completed both the first and the second assignment.

The two assignments both consisted of constructing a story from a six-frame cartoon. The cartoon strips used are Cartoon Task 1 in Appendix 3.1 and Cartoon Task 2 in Appendix 3.2. Students were given 25 minutes to complete each task. For students to be eligible to be included in the study, they had to satisfy the 100 word minimum on both assignments. There were 69 students that satisfied these conditions and 64 of these were selected at random to be part of the analysis. See Appendix 3.3 for sample essays for Cartoon Task 1 and Appendix 3.5 for sample essays for Cartoon Task 2. The analysis was carried out using the L-unique computer program (see Appendix 1.1).

3.3.2.2 Sampling of subsets

The basic lexical signature analysis is the same as in Meara et al. (2002) and the previous experiment. However, in the previous experiment, the number of unique signatures per essay out of 25 subsets was found to be unreliable. This seems to be because 25 represents only a small fraction of the possible combinations of six words selected from thirty. In the original study, 25 subsets were selected from fifty words. Table 3.9 shows the potential number of combinations of subsets of various sizes that can be selected from thirty or fifty words.

Table 3.9: Possible subset combinations by sample and subset size

Subset size	Number of words	
	thirty	fifty
six-word	593775	15890700
eight word	5852925	536878650
ten word	30045015	10272278170

This means the sampling rates in the original study and in the previous experiment are tiny. Meara et al. used 25 six-word subsets selected from fifty words. This represents a sampling rate of:

25/15,890,700 x 100% = 0.00016%

In the previous experiment, 25 six-word subsets were selected from thirty words. This represents a sampling rate of:

25/593775 x 100% = 0.0042%

The low sampling coverage in the previous experiment must contribute to the low measurement reliability. In order to improve this reliability, more subsets need to be sampled. If the number of subsets is increased to 1000, 10,000 or 100,000 subsets, then the sampling rate improves to 0.17%, 1.68% and 16.8% respectively as shown in Table 3.10. For considerations of processing time for the computer analysis, 10,000 subsets were decided on.

Table 3.10: Sampling rates by number of subsets for a six-word subset

Number of subsets	Sampling rate (%)
25	0.004
1000	0.17
10,000	1.68
100,000	16.8

Up to now, the choice of a six-word subset has been arbitrary and the choice of thirty words rather than the fifty words in the original study was necessitated by the relatively small corpus. However, it is clear that if a larger subset and/or fifty words is used, a lot more subsets will be needed to get the same sampling coverage. Table 3.11 shows how many subsets would need to be sampled to get the same coverage, 1.68%, as 10,000 six-word subsets sampled from thirty words.

Table 3.11: Number of subsets to achieve sampling rate of 1.68%

Subset size	Number of words	
	thirty	fifty
six-word	10,000	267,620
eight-word	98,570	9,041,734
ten-word	505,997	164,356,450

3.3.2.3 Quality assessments of essays

To address the third aim of gauging the usefulness of the uniqueness index in grading essays, the two sets of essays were graded by an experienced EFL instructor who specializes in teaching writing classes. The instructor was asked to put the essays into groups such that the essays within the group were judged to be similar in quality and where there were clear differences in quality of essays between the groups. The number of groups and the number of essays within a group were not specified. The essays from the two tasks were graded separately and the rater had no reference as to the candidates' identities. Results of the grading are shown in Table 3.12.

Table 3.12: Candidate grades for Tasks 1 & 2

		Task 2			
	Grade	1	2	3	Total
Task 1	A	9	13	0	22
	B	4	31	0	35
	C	0	2	5	7
	Total	13	46	5	64

For Task 1, the grading produced three groups A, B and C with Group A the best essays and Group C the worst. For Task 2, there were also three groups, 1, 2, and 3 with Group 1 the best essays and Group 3 the worst. The groups for each task are labeled differently since there is no reason to think they correspond with each other in terms of quality. The two shaded groups are of interest since one represents candidates whose essays scored high grades on both tasks and the other candidates whose essays scored low grades on both tasks. These two groups will be useful for comparison because they should be quite different in quality.

3.3.2.4 Lexical signature analysis

For the analysis, the thirty words that occurred in closest to half the essays were chosen. In fact, very few words appeared in exactly 32 essays: two in Task 1 (*off* and *one*) but none in Task 2. Therefore, thirty words were selected that occurred in closest to half the essays. In Task 1, the thirty selected words occurred in 23 to 41 essays, and the thirty selected words for Task 2 occurred in 24 to 40 essays. The words selected for each task are shown in Table 3.13.

Table 3.13: Thirty selected words for Tasks 1 & 2

Task 1	*back boy came clothes day didn't father for go good home I my off one said sea so take that then there they this thought threw throw too took were*
Task 2	*again are because car day for go good have idea in it made next one stop that their then there this though too trees very walk walking was were with*

3.3.2.5 Index of uniqueness

Six-word subsets were constructed randomly from these words and binary representations of each essay were examined for each subset. Any essay with a unique binary representation for a subset was awarded a point. This was done for 10,000 subsets and the number of points were totaled and then divided by 10,000 for convenience to give a value between 0 and 1 for a uniqueness index.

$$\text{Uniqueness index} = \frac{\text{number of unique signatures}}{10000}$$

3.3.3 Results

3.3.3.1 General test characteristics

Table 3.14: Task characteristics

	Task 1	Task 2
Number of essays	64	64
Total number of words	9044	9267
Number of different words	758	795
Average length of essay (words)	141.3	144.8
longest essay	230	250
shortest essay	100	101
Average number of different words per essay	63.2	64.7
highest	99	121
lowest	39	37
Average type-token ratio	0.45	0.45
highest	0.56	0.59
lowest	0.29	0.30

Some of the characteristics of the corpora derived from each task are shown in Table 3.14. For comparison of reliability, correlation coefficients (r) were calculated over the two tasks in terms of a) essay length, b) number of word types in the essay and c) the type-token ratio. These correlation coefficients are shown in Table 3.15.

Table 3.15: Reliability of basic features over the two tasks

	Essay length	Word types	Type-token ratio
Correlation (r)	0.53	0.62	0.50

All of these correlations were significant at the .001 level. From Table 3.14, it is clear that the characteristics of the two tasks were quite similar in terms of essay length and the number of word types generated. There is considerable internal variation in both tasks with regards to essay length and the number of word types per essay. The significant correlations between the two tasks in Table 3.15 suggest that students were likely to produce a similar amount of words on each of the tasks, a similar amount of word types and also have a similar type-token ratio over the two tasks.

3.3.3.2 Measurement reliability

The first aim of this experiment was to ensure the measurement reliability of unique lexical signatures. This reliability can be measured by doing the experiment two times using different sampled subsets and correlating the unique lexical signatures per essay each time. In the previous experiment using 25 subsets, the reliability averaged r = 0.32 over a number of trials. This time, 10,000 subsets were used and the reliability over two trials using different subsets was r = 0.995. This indicates a great improvement and suggests that this measurement of an index of uniqueness is reliable with 10,000 subsets.

3.3.3.3 Reliability over two tasks

Table 3.16: Unique lexical signatures and uniqueness index scores for both tasks

	Task 1	Task 2
Average unique signature per subset	21.6	22.8
Average uniqueness index score	0.34	0.36
highest	0.48	0.50
lowest	0.19	0.20
Correlation (r) (Task 1 & Task 2)	-0.10	

The second aim of the experiment was to investigate reliability over two written tasks by the same learners. In order to do this, uniqueness index scores of essays were compared for Task 1 and Task 2. The results for each task are compared in Table 3.16. It is clear that the average number of signatures per subset, the average uniqueness index score per essay and the highest and lowest uniqueness index score per essay are all quite similar for both tasks. However, the correlation coefficient (r) between uniqueness index scores for the two tasks is very low and negative. This suggests there is no obvious relationship between uniqueness index scores over the two writing tasks.

3.3.3.4 Uniqueness index scores and quality assessments

Even though there is no reliability in uniqueness index scores over tasks, the final aim, to see how uniqueness index scores relate to quality ratings of essays, will be briefly considered. For this question, the quality ratings were compared with uniqueness index scores for each task. Two groups of learners were isolated: those that were rated high on both tasks and those that were rated low on both tasks. There were nine learners that rated high on both tasks and five learners that were rated low on both tasks as indicated by the shaded portions of Table 3.12. If uniqueness index scores are a good predictor of essay quality, these two groups could be expected to be clearly distinguishable. The uniqueness index scores of these learners on both tasks can be seen in Figure 3.1. Points marked H indicate learners who produced essays which were rated high on both tasks, and points marked L indicate learners who produced essays which were rated low on both tasks.

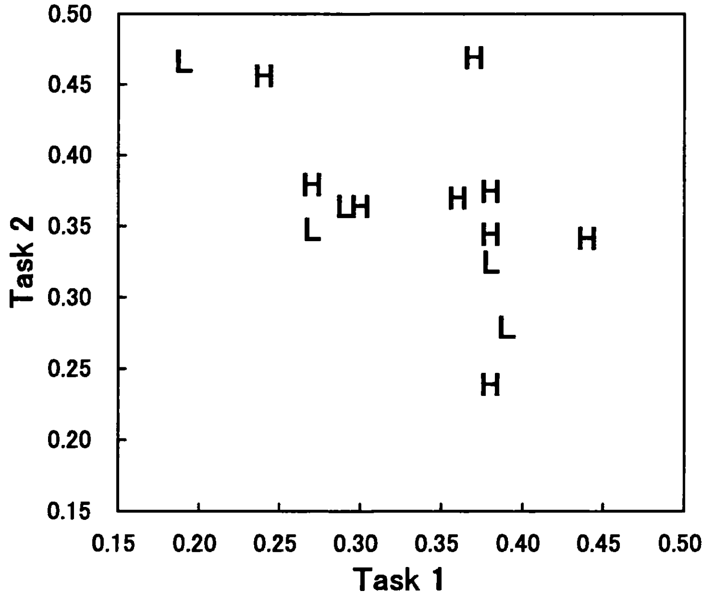


Figure 3.1: Uniqueness index scores for selected tasks

The graph shows a slight tendency for the lower rated students to have lower uniqueness index scores and higher rated students to have higher uniqueness index scores. There are three *high* candidates that seem to reliably score high on both tasks but there are some learners who score high on one essay but low on another. For example, in the top left hand corner of the graph, one *low* candidate scores high on Task 2 but low on Task 1. There is also one *low* rated learner who scores quite high on both tasks. However there is no clear delineation between *high* and *low* candidates. This means that it would be impossible to discriminate between the two sets of essays by unique lexical signatures alone. This is disappointing since these candidates should exhibit a clear difference in quality of essay.

The graphs also highlight the lack of reliability in the scores. If the scores showed perfect reliability, the points could be expected to lie along a 45° line coming from the intersection of the axes. This is clearly not the case. However, a few of the candidates do seem to have performed reliably in terms of their uniqueness index scores. For example, there are three high candidates with similar uniqueness index scores on each task.

3.3.4 Discussion

3.3.4.1 Problems of reliability

This experiment was successful in stabilizing the measurement reliability of uniqueness index scores but was ultimately unsuccessful in finding reliability on tasks written by the same learners and in finding a relationship between uniqueness index scores and quality ratings of essays. Such low reliability over two tasks cannot be accounted for by the time interval between the two tasks since essay length, the number of word types in the essay and lexical diversity all showed significant correlations of $r = 0.53$, $r = 0.62$ and $r = 0.50$ respectively.

Because this index of uniqueness does not seem to be reliable over tasks written by the same learners and does not show a strong relationship with quality, its use may be limited. However, it may be instructive to consider what might be the cause of these deficiencies.

One possible cause could be weaknesses of uniqueness as a concept. There are at least two obvious aspects to this. The first is that uniqueness is an unstable concept. It can easily change with slight changes in the environment. A lexical signature may have a unique representation for a certain essay in a set of essays but if one more similar essay is added to the set, this uniqueness is lost completely. Another weakness is that uniqueness is a skewed concept. In terms of reward, it is all or nothing. For example, a lexical signature is unique or not unique. There is no distinction between an essay that shares the same lexical signature with one other essay and an essay that shares a lexical signature with all other essays. These two aspects of uniqueness are likely to contribute to the lack of reliability over several tasks.

Another possible cause could be the symmetry of uniqueness. A unique representation could be the result of a learner using more sophisticated vocabulary than other learners, but it could also be because a learner uses vocabulary mistakenly. This may mean that high uniqueness scores could indicate good essays or poor essays.

Another possible cause could be the various parameters involved in this calculation. The effect of one parameter, the number of subsets used, was shown to exert a huge effect on measurement reliability. By increasing the number of subsets, the effect of this was alleviated. But other arbitrary decisions about parameters may also mask unreliability. One decision is in regard to the words used. In the analysis for each task, thirty frequently occurring words were sampled. Perhaps the uniqueness index score is affected by this sampling as well. How reliable would the measurement be if two different samples of thirty words were taken? This could not be tested easily since there were not many more words of a similar frequency. However for Task 1, there were two further words of a similar frequency that were not used in the initial analysis. To check the effect of small changes in the composition of the original thirty words, these two words were used to replace two words in the Task 1 analysis. The experiment for Task 1 was then repeated with 28 of the same words but with two words replaced and the scores compared with the main Task 1 analysis. When the same words were used, with different sets of 10,000 subsets, the reliability was r = .995. When two words were replaced this reliability fell to r = .84. This seems a large loss of reliability for such a minor change. This suggests that the actual thirty words sampled are also a considerable threat to reliability.

3.3.4.2 Considerations of word selection

The selection of words may not only cause problems related to sampling. There are other concerns.

One concern is the effect the number of sampled words in a particular essay may have on the uniqueness index score. Table 3.17 shows the number of the thirty sampled words that occurred in each essay for both tasks.

Table 3.17: Number of thirty words per essay

	Average per essay of 30 words	Highest	Lowest
Task 1	14.8	22	8
Task 2	14.6	23	8

For both tasks, there was a large variation in the number of the thirty sampled words in each essay. This variation was similar for both essays. Some essays included as few as eight of the thirty sampled words while others included as many as 23. This variation was despite the fact that these words all occurred in about half the essays. It is likely this number of words in a particular essay may affect the uniqueness index score.

A correlation analysis between the number of the thirty sampled words present in an essay and the uniqueness index score for that essay revealed a correlation of r = .39 for Task 1 and r = .62 for Task 2. These correlations seem large and so might indicate a problem. The small number of sampled words in some essays in the analysis may have another ramification on quality assessment. A method which assesses essays by the presence of only eight high frequency words may also encounter problems with validity.

The selection of words is an aspect that certainly warrants attention. In the original study, fifty words were chosen that occurred in about half the essays. In these experiments, the small corpus precluded using fifty words so only thirty were used. This turned out to be fortuitous in one respect since thirty words were easier to handle from a sampling point of view. Whether words occurring in half the essays are the best words to use in the analysis may be an area worth more investigation. These words are the most powerful in producing a range of lexical signatures and a high number of unique signatures. However, there may be an argument for using a different selection of words. One possibility might be words that are specific to the task.

Given the likely effect of word selection on the number of unique lexical signatures, the experiment was done one more time but this time using thirty words that appeared frequently in both tasks. These words are shown in Table 3.18.

Table 3.18: Thirty words that appeared frequently in both tasks

day for go good one that then there this thought too were father at by can do don't move of put want didn't had my take are have their with

The uniqueness index scores for each task were calculated using 10,000 six-word subsets of these words and the uniqueness scores for each task correlated with r = .11. Using words that appeared frequently in both tasks improved the reliability but only slightly. More detail of this analysis can be found in Appendix 3.6.

3.3.5 Conclusion

The results of the experiment in terms of lack of both reliability and a relationship with quality assessments mean that uniqueness index scores in this form show limited promise for assessment purposes. The complex calculation involves various parameters. These parameters may each introduce their own threats to measurement reliability making it difficult to develop this index of uniqueness in its present form. Further research could be done on the selection of words, and other parameters such as subset size or frequency of words selected. It seems that the concept of uniqueness itself may have poor properties on which to base a measurement instrument.

3.4 Conclusion

The two experiments in this chapter have investigated the lexical signature approach proposed by Meara, Jacobs and Rodgers (2002) with essays written by low level Japanese learners of English. The essays displayed a range of unique lexical signatures that was comparable to that found in the original study despite a much smaller corpus. This suggests that these low level learners also exhibit considerable individual lexical choice. However, harnessing these lexical signatures to produce a uniqueness index stable in measurement and reliable over tasks by the same learners proved difficult. Lexical signatures seemed to be particularly sensitive to small changes in the nature of subsets and words selected for analysis. In particular, small numbers of subsets led to considerable variation in scores due to insufficient sampling coverage. It was found that a large number of subsets were required to ensure measurement reliability. It is feared that other parameters such as the word selection may lead to similar variations in scores. Also preliminary comparisons with quality assessments found no evidence that uniqueness index scores were sensitive to variations in quality.

A lot of the problems faced in these experiments are likely to be due to the complexity of the lexical signature process and the underlying instability of uniqueness. In the next chapter, distinctiveness is examined. Distinctiveness enables a similar examination of the relative use of vocabulary across essays but is underpinned by a more robust concept yet a simpler and more transparent mode of calculation.

Chapter Four: Distinctiveness

4.1 Introduction

In the previous chapter, an index of uniqueness was developed and its reliability explored. This index based on lexical signatures was found not to be reliable over two similar tasks by the same learners. Several possible causes of this were considered. One possible cause was the small number of words involved in the analysis. Another cause may have been the properties of uniqueness. Firstly, uniqueness rewards unique occurrences but treats very rare occurrences in the same way as very common occurrences, i.e. it ignores them. This means there is a stark cut off point for reward. A word that appears in one essay is rewarded but one that appears in two is not. Secondly, uniqueness has unstable properties. This is a result of the previous characteristic since the addition of one similar essay can cause a score for another essay to drop from being very high to very low as occurrences change from being unique to rare.

In this chapter, distinctiveness is considered as an alternative to uniqueness. Distinctiveness subsumes the concept of uniqueness but considerably expands on it so that not only unique words but all words are considered. Each different word is rewarded according to how rare it is over a set of essays. A unique word, one that appears in only one essay in a set of essays, reaps the greatest reward followed by a word that appears in only two essays, followed by a word that appears in only three essays and so on down to a word that appears in all essays but one and reaps the least reward. This concept of distinctiveness seems superior to the concept of uniqueness for measurement purposes. Unlike the index of uniqueness which relied on a very small number of words in an essay, distinctiveness incorporates all of the different words in each essay. Moreover, because distinctiveness is based on arithmetic principles, it is a clear and easy to understand measure that has good mathematical properties. This makes distinctiveness less susceptible than uniqueness to sudden variation with minor changes to the essay corpus. Consequently, distinctiveness should be a more stable measurement than the index of uniqueness.

In this chapter, two experiments explore the properties of distinctiveness. In particular, its reliability over tasks written by the same learners and its relationship to essay quality are investigated. In the first experiment, to check for reliability, distinctiveness is calculated for two different but similar tasks written by the same learners. The distinctiveness scores are compared with holistic quality assessments. In the second experiment, a measure of distinctiveness adjusted for error is also investigated, but this time, on two essays written by learners on the same task.

4.2 Distinctiveness in L2 learner texts

4.2.1 Introduction

This experiment explores some of the properties of distinctiveness for a set of essays written by low level L2 learners. Firstly, a simple measure of distinctiveness is introduced. Secondly, distinctiveness scores are checked over two similar tasks written by the same learners in order to gauge reliability. Finally, the relationship between distinctiveness and essay quality is investigated by comparing distinctiveness scores of essays with quality evaluations.

4.2.1.1 Aims of the experiment

The three main aims of the experiment are to:

1) construct a measure of distinctiveness.

2) test the reliability of distinctiveness for low level L2 learners over two sets of written tasks.

3) see how this measure of distinctiveness relates to holistic evaluations of those tasks.

4.2.2 Methodology

4.2.2.1 Participants and tasks

The same sets of essays based on the cartoon tasks were used as in the Chapter 3.3 experiment (see Appendix 3.1 and 3.2 for cartoon tasks). Sixty-four candidates writing at least 100 words on both tasks were selected. Each set of 64 essays was analyzed separately using L-unique, a specially-designed computer program (see Appendix 1.1).

4.2.2.2 Calculating distinctiveness

A measure of distinctiveness was calculated for each candidate on both tasks. This measure is basically the sum of all the different word types in an essay divided by the number of essays they appear in. For example, if a word is unique and so appears in no other essays it would earn a credit of one. If a word appears in one other essay, it earns each of those essays a credit of 1/2. If a word occurs in 20 essays it earns each essay a credit of 1/20. However, this measure then needs to be adjusted for each essay by dividing by the number of word types in that essay. This is to adjust for essay length effect else longer essays with more word types will tend to get higher distinctiveness scores. This value is then multiplied by 100 to give a value that is easier to work with.

In this form, distinctiveness (DIS) can be represented by the following formula:

$$DIS = \frac{100}{v}\sum_{i}\frac{1}{\sum_{j} w_{ij}}$$

where v = the number of word types in an essay and w_{ij} = 1 if word i appears in essay j or 0 otherwise.

4.2.3 Results

4.2.3.1 Distinctiveness scores

Table 4.1 shows distinctiveness scores alongside some other lexical features for both Tasks 1 and 2, T1 & T2. Average scores (av.), highest (hi) and lowest (lo) values are shown. Also the correlation coefficient, r, between each value on Tasks 1 & 2 is given.

Table 4.1: Distinctiveness scores and lexical features for two tasks

	Distinctiveness			Essay length			Word types			TTR		
	Av.	Hi	Lo	Av.	Hi	Lo	Av.	Hi	Lo	Av.	Hi	Lo
T1	17.85	34.07	8.14	141	226	100	62	101	40	0.45	0.56	0.29
T2	18.18	32.76	7.79	145	234	103	63.4	113	38	0.45	0.59	0.30
r		0.12			0.50*			0.59*			0.50*	

*significant at the .001 level

The range of distinctiveness values were similar on the two tasks ranging from 8.14 to 34.07 with an average of 17.85 on Task 1 and from 7.79 to 32.76 with an average of 18.18 for Task 2. However, the correlation between values on the two tasks was only r = 0.12 which is not significant. There were quite strong correlations for all three of the lexical features over the two tasks. Essay length for Tasks 1 & 2 correlated with r = 0.50, word types with r = 0.59 while the TTR correlated with r = 0.50. All three of these were significant at the .001 level.

4.2.3.2 Eliminating inconsistent students

Although the correlation for essay length was significant, we might have expected it to be higher given the similarity of the tasks. We might expect the same students completing two very similar tasks under exactly the same conditions to produce essays of very similar length. Candidates who wrote two essays of very varied length may be suspect to some degree. Just from personal observation, some students were obviously less enthusiastic about completing the second task than they had been about the first. The premise of this experiment is to collect similar pieces of work from students. A considerable difference in length may be an indicator that a particular candidate has not produced two similar pieces. If that is the case then error will be introduced into the experiment. Therefore, the data was re-analyzed after eliminating candidates whose total number of words changed by more than 20% from Task 1 to Task 2.

Only candidates with similar length essays were considered. 48 of the 64 candidates wrote within +/- 20% of the total words in the first task for the second task. These 48 candidates were identified and a re-analysis was performed. The correlations of distinctiveness, essay length and the number of word types are shown in Table 4.2.

Table 4.2: Distinctiveness and other measures for revised analysis

	Distinctiveness	Essay length	Word types
correlation (r)	0.22	0.82*	0.74*

*significant at the .001 level

Table 4.2 shows that correlation for distinctiveness is improved by only selecting students who were relatively consistent in terms of amount of output. The distinctiveness measure has a correlation of r = 0.22 which is an improvement on the earlier figure but still low.

4.2.3.3 Distinctiveness and quality assessments

The ratings of each task for this reduced number of 48 students can be seen in Table 4.3.

Table 4.3: Holistic ratings of written tasks

		Task 2			
	Grade	1	2	3	Total
Task 1	A	7	11	0	18
	B	0	25	0	25
	C	0	1	4	5
	Total	7	37	4	48

For Task 1, Grade A essays were the best and Grade C essays were the worst. For Task 2, Grade 1 essays were the best and Grade 3 essays were the worst. Seven of the candidates had essays rated in the top group for both essays, Grade A for Task 1 and Grade 1 for Task 2. Similarly, four candidates had essays rated in the bottom group for both essays, Grade C for Task 1 and Grade 3 for Task 2. These are the shaded areas of Table 4.3. These candidates are likely to be of different proficiencies so it might be instructive to examine distinctiveness for these candidates. Since these are obviously the candidates whose essays varied the most in terms of quality, there should be a clear difference in distinctiveness scores if distinctiveness is to be a measure of essay quality.

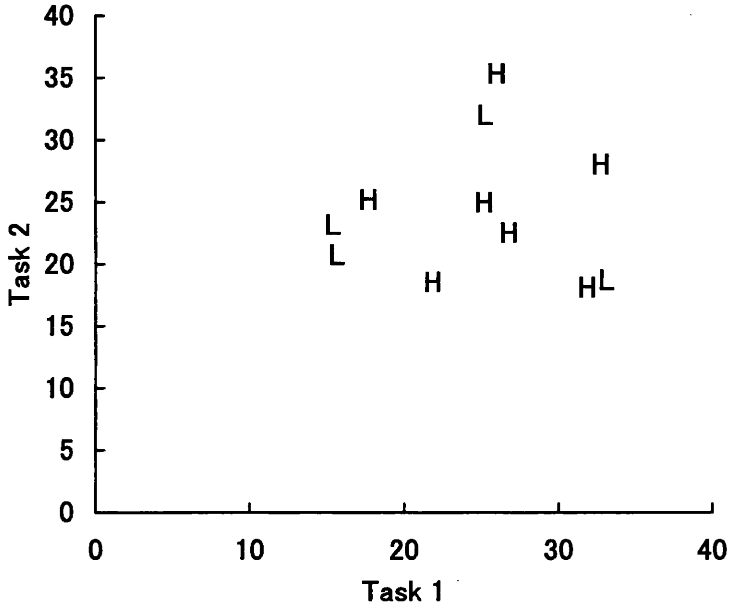


Figure 4.1: Distinctiveness for selected learners on both tasks

Figure 4.1 shows distinctiveness mapped for these 11 candidates on the two tasks. *High* (H) candidates are those whose essays were rated high on both tasks and *low* (L) candidates are those whose essays were rated low on both tasks. By drawing up a line from the x-axis, distinctiveness can distinguish two *low* candidates on Task 1 but by drawing a line from the y-axis none can be distinguished on Task 2. This graph also highlights the lack of reliability in the performance of some candidates. Some candidates scored very differently over the two tasks. One *low* candidate got a high score, over 30, on Task 1 but a very low score, under 20, for Task 2. However, there seem to be a cluster of *high* candidates who scored reliably on both tasks and sit close to an imaginary line at 45° from the origin. For comparison, Figure 4.2 and Figure 4.3 show essay length and the number of word types mapped for the same candidates. Figure 4.2 shows that essay length isolates three *low* candidates on Task 1 but only two on Task 2. Figure 4.3 shows that word types performs well on Task 1 isolating all four *low* candidates but only isolates two on Task 2. Although essay length and word types do not distinguish between *high* and *low* candidates perfectly, their performance is superior to distinctiveness. For both features, the proximity of the points to a 45° line illustrates the reliability of these features over the two tasks.

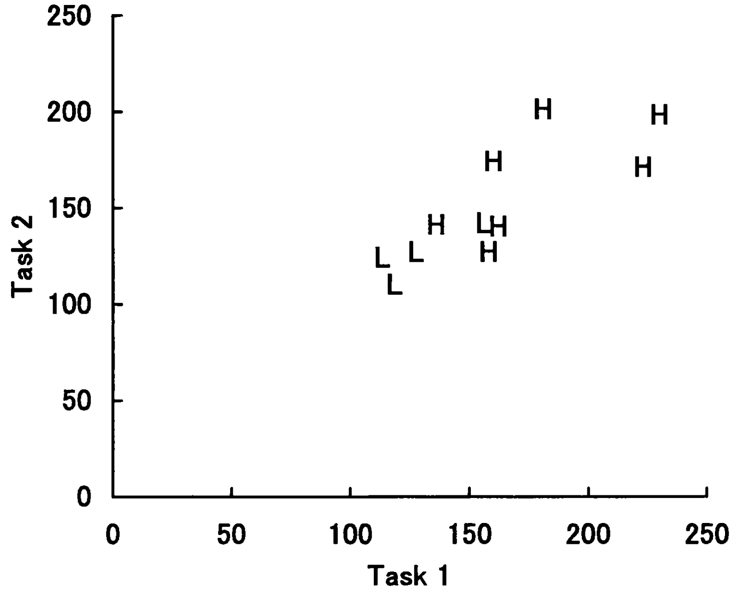


Figure 4.2: Essay length for selected learners on both tasks

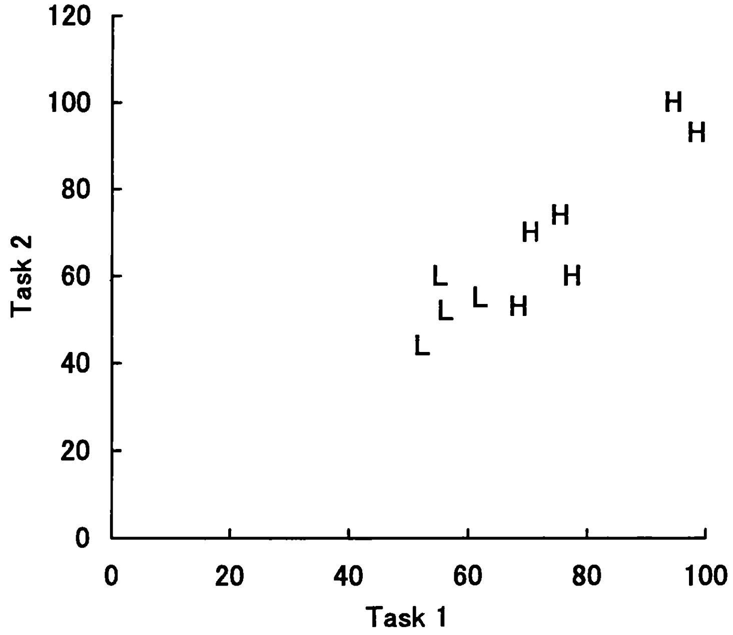


Figure 4.3: Number of word types for selected learners on both tasks

4.2.4 Discussion

The reliability for distinctiveness of r = 0.12 over the two tasks is disappointingly low. Even removing learners who appear to perform inconsistently over the two tasks only enabled us to improve the performance of reliability to r = 0.22. This is an improved performance but is still very low. Hatch & Lazaraton (1991) recommend an r value in the high 0.8's or 0.9's for test-retest reliability. It is unlikely that when using written tasks such reliability can be realized given the variability in content but to get close must be the target. The low reliability is highlighted by the superior reliability of other features such as essay length (r = 0.50), word types (r = 0.59) and TTR (r = 0.50).

The graphical comparison of *high* and *low* candidates suggests that distinctiveness is not a strong indicator of quality. There is some evidence from the same graphs that distinctiveness may be more reliable for higher quality essays than lower quality essays. A cluster of *high* candidates scored reliably over the two tasks whereas the *low* candidate scores seemed to be more variable. However, because the numbers of candidates are so small, this requires more investigation.

An assumption behind this study is that the two tasks are similar so that students should perform in a similar way over the two tasks. Evidence for this is the high correlations for the lexical features of essay length, number of word types and TTR. Also the performance of candidates from Table 4.3 suggests the tasks elicited a similar performance with 36 out of 48 candidates being awarded a similar grade over the two tasks. This represents agreement reliability of 75%. However, it is possible the scores for distinctiveness may have been influenced by the different contexts of these written tasks. It may be that distinctiveness is very sensitive to the actual vocabulary used in the tasks. If this is the case, care may have to be taken in the selection of tasks.

One other feature of note was that a large number of words unique to an essay were, in fact, spelling errors. Many of these were errors of words that occurred often among the essays. These errors are problematic because distinctiveness gives the most reward to unique words. This highlights a potential weakness of distinctiveness. The words that have the greatest input into the statistic are also the words that may be the least reliable because of their low occurrence. To solve the problem of unique spelling errors, unique words could be eliminated from the analysis altogether leaving words that appear in at least two essays in the set.

4.2.5 Conclusion

This experiment has failed to produce a reliable measure of distinctiveness over two similar tasks. This lack of reliability is contrasted with other lexical features performing relatively reliably. Also, there was no obvious relationship apparent between distinctiveness and essay quality. In an attempt to improve reliability, in the next experiment, we will try distinctiveness on two identical tasks. This should eliminate any variability in the vocabulary used in the essays. Also a measure without unique words will be included in the analysis to gauge the impact of spelling errors which may be distorting distinctiveness scores.

4.3 Testing reliability of a measure of distinctiveness on identical tasks

4.3.1 Introduction

In the previous experiment, a measure of distinctiveness was developed and its reliability explored over two written tasks by a group of low level L2 learners. Reliability was relatively low at r = 0.22. It is possible that the two different tasks used in that experiment may not be truly comparable since they involved quite different sets of vocabulary. An alternative way to test reliability is over two attempts at the same task. In this experiment, distinctiveness scores of learners are compared on two identical tasks. An additional measure of distinctiveness omitting words that occur only one time in an essay is also calculated. This is to guard against the analysis being affected by spelling errors.

4.3.1.1 Aims of the experiment

The aims of this experiment are to:

1) test the reliability of distinctiveness over two attempts on identical tasks by low level L2 learners.

2) compare the reliability of distinctiveness and distinctiveness without unique words.

4.3.2 Methodology

4.3.2.1 Participants and tasks

Thirty-six first-year Japanese undergraduate students of English completed the same written task twice at an interval of four weeks. The task was Cartoon Task 1 from Appendix 3.1. The students were given 25 minutes to complete the task.

4.3.2.2 Details of the analysis

The two sets of essays were analyzed separately using the L-unique computer program (see Appendix 1.1). Distinctiveness was calculated the same way as in the previous experiment. This time an additional measure was constructed for distinctiveness without unique words. The calculation was exactly the same as for distinctiveness but without including words that only occur in one essay in the corpus. The rationale for this was that an examination of the unique words in the previous study had shown that a high proportion were spelling errors of higher frequency words. Under distinctiveness, this rewards students for unique errors. The number of identical spelling errors occurring in more than one essay is likely to be small so this step may exclude a high proportion of the spelling errors.

4.3.3 Results

4.3.3.1 Distinctiveness scores

Table 4.4 shows distinctiveness scores and distinctiveness scores without unique words for both Tasks 1 & 2, T1 & T2. Average scores (av.), highest (hi) and lowest (lo) values are shown. Also the correlation coefficient, r, between each value on Tasks 1 & 2 is given.

Table 4.4: Distinctiveness scores with/without unique words

	Distinctiveness			Distinctiveness w/o unique words		
	Av.	Hi	Lo	Av.	Hi	Lo
T1	19.98	37.93	10.44	11.60	16.03	8.50
T2	18.46	31.09	7.59	11.44	15.38	7.59
r		0.61*			0.22	

*significant at the .001 level

Distinctiveness ranged from 10.44 to 37.93 with an average of 19.98 on Task 1 and varied from 7.59 to 31.09 with an average of 18.46 for Task 2. Distinctiveness without unique words ranged from 8.50 to 16.03 with an average of 11.60 for Task 1 and varied from 7.59 to 15.38 with an average of 11.44 on Task 2. The distinctiveness scores without unique words were considerably lower than distinctiveness. This suggests that there are a large number of unique words in the analysis. The distinctiveness scores on the two tasks correlated significantly at the .001 level with r = 0.61 but distinctiveness without unique words only correlated with r = 0.22. This shows the reliability of distinctiveness is considerably higher than for distinctiveness without unique words. The much lower scores for distinctiveness and the lower correlation suggest that unique words play a major role in the distinctiveness statistic. This is emphasized when one bears in mind that unique words also receive the most reward.

Table 4.5 shows values for essay length, word types and the type-token ratio for each task. Essay length for Tasks 1 & 2 correlated with r = 0.82 and word types with r = 0.83 which is significant at the .001 level. TTR correlated with r = 0.49 which is significant at the .01 level.

Essay length and the number of word types increased slightly over the two tasks. The average essay length, the highest essay length, the average word types and the highest word types were all higher in Task 2. This suggests that students may have performed slightly better on this task the second time around in terms of how much they wrote and the diversity of their vocabulary.

Table 4.5: Total words, word types and type-token ratio

	Essay length			Word types			TTR		
	Av.	Hi	Lo	Av.	Hi	Lo	Av.	Hi	Lo
T1	118.2	227	100	52.7	98	40	0.45	0.58	0.34
T2	128.8	288	100	59.1	120	35	0.46	0.57	0.35
r		0.82**			0.83**			0.49*	

**significant at the .001 level *significant at the .01 level

4.3.3.2 Consistency of learners

As noted in the previous experiment, this kind of experiment assumes that students will perform the two tasks in the same way. A lot of variation in the previous experiment was suspected as being performance error where student performance was inconsistent over the two tasks. However, the reliability of essay length in this experiment was much higher than in the previous experiment, r = 0.82 compared to r = 0.50 in the previous experiment. In fact, it is comparable to reliability in the previous experiment after unreliable candidate essays were removed. Therefore, it was deemed unnecessary to do a similar revision in this experiment.

4.3.4 Discussion

The first aim of this experiment was to test the reliability of distinctiveness on two identical tasks by the same learners. Using identical tasks succeeded in increasing the reliability considerably to r = 0.61. This was in line with other lexical features which displayed much more consistency over two identical tasks than two similar tasks with essay length correlating with r = 0.82 and word types with r = 0.83. This supports the idea that learners perform in a more similar manner over two identical tasks than over two different tasks. In fact, there is evidence from essay length and the number of word types that the learners performed slightly better in the second task. However, this is not reflected by either of the distinctiveness statistics where the average score on both was lower for Task 2 than for Task 1.

The higher reliability achieved with identical tasks rather than different tasks suggests that a distinctiveness measure may be very task specific. Scores for distinctiveness on one task might not correspond to those on a different task. This may mean that distinctiveness may be better used with standardized tasks so that results from different studies could be compared more meaningfully. The kind of task best selected could be an area for further research. The relationship of distinctiveness with essay quality also warrants further investigation.

The second aim of the experiment was to see if the elimination of unique words increased the reliability of the measurement by reducing the effect of rewarding spelling errors. Surprisingly, the reliability of distinctiveness without unique words was much lower than distinctiveness suggesting that removing unique words makes the measurement less reliable. This could mean that unique words are quite consistent over two identical tasks but the occurrence of other words relative to other essays is not as consistent. The scores also seemed to suggest that unique words make a considerable contribution to the distinctiveness scores. Average distinctiveness without unique words scores were over 40% lower than distinctiveness scores.

4.3.5 Conclusion

In this experiment, the reliability of distinctiveness was investigated over two identical tasks. Distinctiveness was reliable with r = 0.61 which was a considerable improvement over that for different tasks in the previous experiment. It was thought that removing unique words might improve performance but, in fact, without unique words, distinctiveness was much lower at r = 0.22. Also scores for distinctiveness without unique words were much lower. This suggests that unique words play a major role in the analysis and that these words may comprise a relatively reliable component of the statistic.

4.4 Discussion

4.4.1 A comparison of results of the two experiments

Table 4.6: Reliability for different tasks and same tasks

	DIS	DIS w/o unique words	Essay length	Word types
different tasks	0.22	-	0.50*	0.59*
same tasks	0.61*	0.22	0.82*	0.83*

*significant at the .001 level

Table 4.6 shows reliabilities of distinctiveness scores and other lexical features for learners on two different tasks and two identical tasks. The reliabilities are all noticeably higher for identical tasks than for different tasks. This is particularly noticeable for distinctiveness and suggests that this feature is particularly sensitive to differences in tasks. These differences also highlight one of the problems in testing the reliability of measures of writing proficiency. It is very difficult to achieve high levels of reliability without using identical tasks. It suggests that we might have to accept lower levels of reliability when using different tasks. However, if using repeated tasks, it must also be borne in mind that students may sometimes react negatively to repeating the same task which may also have a slight effect on the results.

This task specificity also hints at a lack of a strong relationship with quality. On the whole, we expect higher proficiency learners to perform well on any kind of written task. For a measure to have a strong relationship with quality, it needs to be as reliable as quality over different tasks. The lack of reliability over different tasks of many measures of written productions is reflected in only moderate relationships with essay quality.

4.4.2 Threats to reliability and validity

It is interesting to speculate on some of the problems with reliability. It is perhaps easy to see how fluctuations in learner performance can easily cause larger fluctuations in a value like distinctiveness. The distinctiveness score for a particular essay depends not only on the characteristics of that essay but also on the characteristics of all the essays in the set. The learners in this study seemed to perform slightly differently in terms of essay length and the number of word types. If part of this difference is performance error then the error manifested in a concept like distinctiveness or uniqueness can be expected to be much larger since the calculation of the measure for each candidate depends on the performance of all other candidates and thus incorporates, to some degree at least, the performance errors of all other candidates. This may also help explain why unique words seem a relatively stabilizing force on distinctiveness. Words which appear in more essays may be more at risk to this cumulative performance error than unique words. However, in another way this result is curious since there were a large number of spelling errors in the unique words. This may also mean that learners are consistent in the number of spelling errors they produce.

There are some challenges to the underlying assumption that distinctiveness indicates a more developed vocabulary while less distinctiveness indicates a less developed vocabulary. These should be labeled as threats to validity since they challenge the basis of what the distinctiveness statistic measures. Errors are one such threat as has been mentioned. Because distinctiveness rewards every word in an essay, it will also reward spelling errors. Because individual spelling error variants may often occur only once and since distinctiveness bestows the greatest reward on words occurring once, then spelling errors are likely to receive the highest reward and so be a significant threat to reliability and validity. One way to counter this may be to not include unique words. However, eliminating unique words in this study was found to lead to a fall in reliability. The effect of rewarding rare words could also be a weakness of distinctiveness. Because it emphasizes reward to words which occur rarely, there is more chance that it can be undermined by rogue words. This is another aspect of the fact that items that occur rarely are less reliable than frequently occurring items. In this sense, lexical signature analysis has an advantage over distinctiveness since all the words in that analysis are frequently occurring words.

Common errors may also be a threat. If the same error occurs in a number of essays, it will not be picked up by eliminating unique words. If a small number of learners produce the same error then they will be rewarded as if producing a relatively distinct word.

Another threat is inappropriate lexical choice. The assumption is that a distinctive word should be superior to a common word as a sophisticated synonym. However, some tasks call for particular or precise lexical items. In this case, using the inappropriate word may get more reward than the correct or precise alternative used by more learners. In Cartoon Task 1 (Appendix 3.1), 29 out of 36 learners used the appropriate word *stick*, but two learners used the word *rod* which seems less appropriate. The learners using the less appropriate word gained more credit than those using the more appropriate word. Similarly, four students used the word *see*. If in some cases, this was a misspelled variant of *sea* used by 20 learners, then more credit would have been received for this misspelled form than for the correctly spelled word.

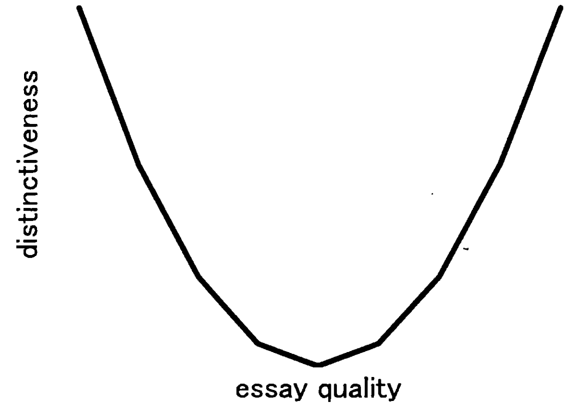


Figure 4.4: The relationship between distinctiveness and quality

These are all examples that point to the lack of linearity of distinctiveness with the construct of essay quality. This is a problem distinctiveness shares with uniqueness. Even if we expect more proficient learners to use more distinct vocabulary and so produce a higher distinctiveness score, we cannot necessarily say that a high distinctiveness score identifies a proficient learner. A learner that uses a lot of inappropriate vocabulary or makes a lot of errors may also score highly on distinctiveness. If we map uniqueness/distinctiveness against proficiency it might have a u-shaped relationship as shown by the graph in Figure 4.4.

Values of distinctiveness can be high at both low levels and high levels of proficiency. With this type of relationship, it is not clear whether a distinct word should be viewed in a positive light as a mark of a more developed vocabulary or whether it should be viewed with skepticism because it may be a sign of inappropriate usage or an error. It may be possible to solve this for distinctiveness. A distinctiveness profile may help determine whether distinctiveness is a positive or a negative indicator. An essay that is positively distinctive is likely to have unique words but also highly distinctive words shared by only one or a few other essays. It is likely to have a balanced distinctiveness profile. An essay with an unbalanced distinctiveness profile, for example, an essay with a lot of unique words but not many highly distinctive words may be more likely to have a number of idiosyncratic errors and be negatively distinctive. It may, however, be difficult to codify this type of analysis into a simple automatic algorithm.

A further threat could be an off-topic essay which would be likely to score very highly for distinctiveness because it may contain a different set of vocabulary to the rest of the on-topic essays.

4.4.3 Reward structure

There may be some argument for changing the reward structure of distinctiveness particularly given the influence of unique words and the susceptibility of these unique words to include error. Using a reciprocal of the number of occurring essays may be a rather top heavy system. For example, under this system, a unique word scores twice as much as a word that occurs in one essay. Figure 4.5 shows the reward system using a reciprocal of the number of essays a word occurs in as the weight. This strong weighting for unique words was highlighted in the study by distinctiveness scores being much higher than distinctiveness without unique words scores. This could be adapted, for example, by using a reciprocal of the square root value. This square root reciprocal weighting system is shown in Figure 4.6. The slope of the curve is less steep meaning that the impact of rarely occurring word will be less. There are an unlimited number of possibilities for different weighting systems which could be investigated.

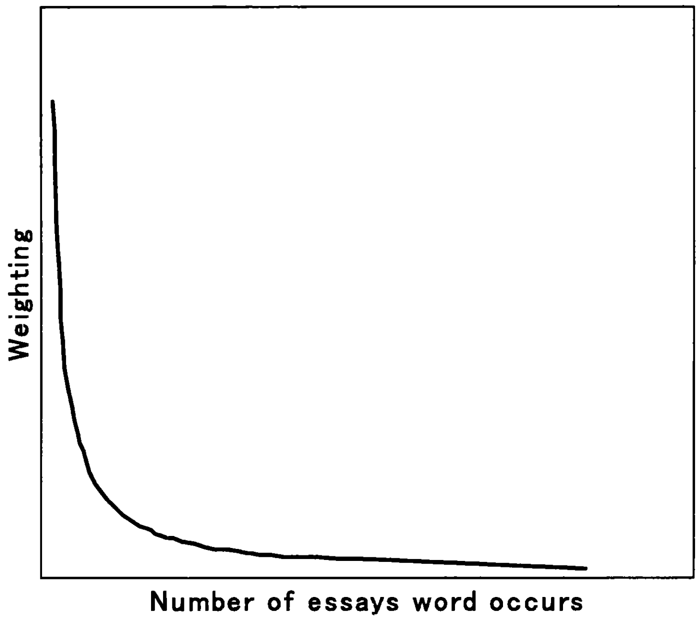


Figure 4.5: Distinctiveness weighting using reciprocal

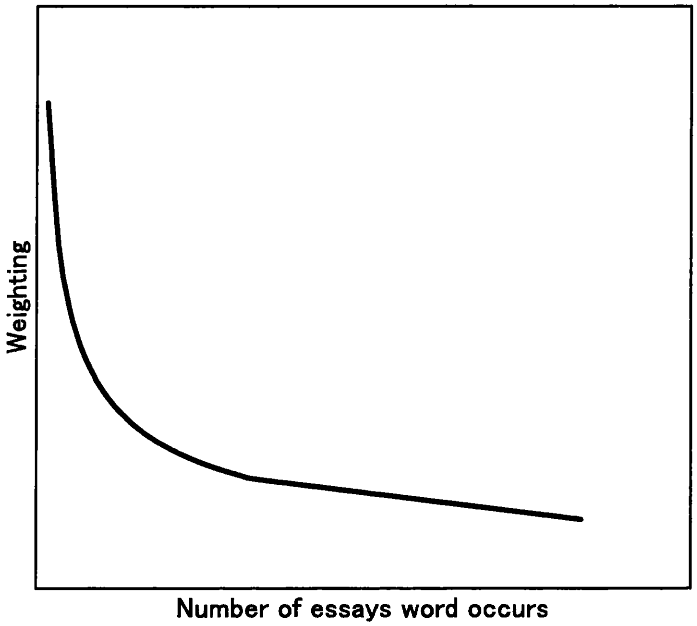


Figure 4.6: Distinctiveness weighting using reciprocal square root

4.5 Conclusion

This chapter looked at the concept of distinctiveness as an alternative to uniqueness. Distinctiveness is a measure of the different word types in an essay with each word type weighted inversely according to how many other essays in the set it occurs in. Distinctiveness has some superior mathematical properties to uniqueness and appears to be more stable. A simple measure of distinctiveness was calculated and tested for reliability over two written tasks by the same learners. Only a very low reliability estimate of r = 0.22 was achieved. Distinctiveness scores were also compared with holistic quality ratings. However, no strong relationship was found. It was noted that many unique words were spelling errors.

In a second experiment, the analysis was conducted on two identical tasks written by the same learners to control for vocabulary content. Reliability improved considerably to r = 0.61. This suggests that distinctiveness may be quite sensitive to changes in vocabulary and may be quite task specific. In the second experiment, an alternative measure of distinctiveness without unique words was also considered to counter the threat of many unique spelling errors. However, the reliability of this measure was much lower. Also, the scores for this measure were much lower than regular distinctiveness scores. This suggests that unique words play a major role in the distinctiveness analysis. There is also a suggestion that because measures such as distinctiveness and uniqueness depend on all the essays in a group, error may be intensified leading to a lack of reliability. Other weighting systems might also be considered to lessen the impact of unique words on the analysis.

The distinctiveness statistic has been shown to be reliable in some contexts. However, more research is necessary into the relationship of distinctiveness and essay quality. There does not seem to be a strong relationship with essay quality. However, more research using different weighting systems may be able to improve this relationship.

The concept of distinctiveness has not proved to be as robust as envisaged and seems to suffer from some of the same problems as uniqueness. It seems the assumption that an essay containing a lot of distinctive vocabulary will be an indicator of higher quality may have some major flaws. A greater understanding of the dynamic of distinctiveness and proficiency may help. It seems unlikely that there is a simple linear relationship between the distinctiveness of lexical items used and proficiency of the learner.

In the first two experimental chapters, two measures of vocabulary content in relation to other essays in the set have been considered. In the next chapter, the scope is broadened and a number of other essay features are considered as predictors of essay quality.

Chapter Five: Exploratory Studies

5.1 Introduction

This chapter reports on four experiments investigating properties of lexical features in L2 written work. The overall aim of the experiments is to identify features that may be useful in assessment of written tasks within larger assessment models. Two properties are particularly important, reliability over different written tasks and correlation with quality assessments. Some of the lexical features focused on have been effective in other fields or in other contexts of vocabulary assessment. The experiments in this chapter are small scale and are aimed at finding possible features to include in multivariate models rather than finding general truths about relationships between features and essay quality.

The first two experiments look at some features that have proven effective in the field of authorship attribution to see if there is a case for their use in assessment. In the first of these, two features, mean word length and mean sentence length, and two lexical statistics, Yule`s K and entropy, are examined. In the second experiment, the properties of some vocabulary distributions are considered. These are the word frequency distribution, the Zipf rank distribution and the word length distribution. The third experiment looks at the relative performance of two intrinsic measures of lexical diversity, lexical variation and the type-token ratio, in relation to low level L2 writing. The final experiment considers the relative performances of the two most common extrinsic measures of lexical diversity, the Lexical Frequency Profile and P_Lex, on some low level L2 essays.

5.2 Lexical features in L2 texts of varying proficiency

5.2.1 Introduction

This experiment looks at some features that have been used in authorship attribution studies and makes a case for their use in assessment. The experiment concentrates on two features of an essay, mean word length and mean sentence length, and two lexical statistics, Yule's K and entropy.

Both mean word length and mean sentence length have been used in authorship attribution. However, a stronger case could be made for their use in assessment than in regular authorship attribution cases. For regular authors, differences in mean word length are simply a function of individual style. In assessment, they may tell us something about the quality of the writing. In particular, the case for mean word length is strong. According to Zipf (1932), more frequent words are likely to be shorter than less frequent words. If we assume that more proficient learners are more likely to use infrequent words, then we might expect mean word length to increase with proficiency. In the case of mean sentence length, it might be argued that less proficient learners are likely to produce shorter sentences so that mean sentence length may be higher for more proficient learners. However, there may be a counterargument for lower level learners. Some low level learners might produce longer sentences because of a lack of skill in sentence construction and punctuation. For example, many low level learners produce long run-on sentences. An examination of learner essays may shed some light on this.

Text length can be a problem in authorship attribution studies as it is in assessment. Lexical statistics are often affected by text length but both entropy and Yule's K have been shown to be, in theory, constant with respect to length of text. These measures have been employed in authorship attribution where texts are often very long, but they are thought to be less stable with the shorter texts found in L2 assessment. In this experiment, we compare their performance against the D estimate (Malvern & Richards, 1997) which is widely recognized as the indicator of lexical diversity least likely to be affected by length for short essays.

5.2.1.1 Aims of the experiment

The aims of this experiment are as follows:

1) to see if mean word length and mean sentence length distinguish between essays written by learners of different proficiencies.

2) to see if Yule's K, entropy and the D estimate distinguish between essays written by learners of different proficiencies.

3) to examine the relationship of Yule's K, entropy and the D estimate with essay length in short L2 essays.

5.2.2 Methodology

5.2.2.1 Participants and tasks

Short essays from two groups of L2 learners of differing proficiencies were collected. Group 1 was made up of eleven third year Japanese university students majoring in English and Group 2 was made up of 24 first year Japanese university students not majoring in English. Each group wrote a short essay using a six-caption cartoon as a prompt. The cartoon used was Cartoon Task 1 in Appendix 3.1.

5.2.2.2 Analysis

The essays were input into a computer and the analysis was done automatically using a specially designed computer program called L-analyzer (see Appendix 1.1) except where acknowledged otherwise. Mean word length was estimated by dividing the total number of characters in an essay by the total number of spaces between words. An estimate of mean sentence length was calculated automatically by dividing the essay length in words by the number of sentence ending punctuation marks such as full stops, question marks and exclamation marks.

Entropy (Shannon, 1948) gives a measure of diversity of a text of any length with absolute diversity as 100 and absolute uniformity as 0. It can be described by the following formula:

$H = -100\sum p_i \log p_i / \log N$ where p_i is the probability of the appearance of *ith* word type in a text of length N words.

For entropy, we would expect a higher quality essay to be more diverse and so report a higher entropy value.

Yule's K (Yule, 1944) is thought to have an inverse relationship with diversity of vocabulary and is described by the following formula:

$K = 10^4(\sum r^2V_r - N)/N^2$ *(r = 1, 2,...)* where V_r is the number of types occurring r times in a text of length N words.

For Yule's K, we would expect a higher quality essay to contain a more diverse vocabulary and so report a lower Yule's K value.

Mean word length, mean sentence length, entropy and Yule's K were calculated for essays for each group along with the D estimate discussed earlier in Chapter 2.12.1.2. The D estimate was calculated using D_Tools software (Meara & Miralpeix, 2007a). The means of each feature or statistic for each group were then compared.

To test the effect of essay length on entropy, Yule's K and the D estimate, the longest essay from each of the two groups was taken. For Group 1, the longest essay was 258 words and for Group 2, it was 149 words. Entropy, Yule's K and the D estimate were then calculated for each essay for different lengths of both essays. The differing lengths were taken in intervals of ten words beginning at fifty words. Fifty words is the minimum number of words for D_Tools to make an estimate of D. For both essays, this meant taking the first fifty words and calculating the statistics, then taking the first sixty words and calculating the statistic, then seventy words and so on until the end of each essay.

5.2.3 Results

5.2.3.1 Mean word length and mean sentence length

Table 5.1 shows the results for mean word length (MWL) and mean sentence length (MSL) for each group.

Table 5.1: Mean scores for MWL and MSL for each group

Group	MWL		MSL	
	mean	sd	mean	sd
1	3.76	0.14	13.47	4.23
2	3.56	0.15	11.07	2.75

Group 1 essays, which we assume to be of higher quality, showed a higher average mean word length, and a higher mean sentence length than the Group 2 essays which we assume to be of lower quality. These averages seem to support the assumptions of relative quality. In addition, the two groups were significantly different in terms of mean word length (t = 3.73, p < .01) but the difference between the groups in terms of mean sentence length was not significant with t = 2.02. Both mean word length and mean sentence length distinguish between the two groups in general terms but are not perfect in discrimination. If we attempt to classify the essays to groups according to mean word length or mean sentence length, then 71% of subjects would be assigned to the correct group for both. However, the evidence overall suggests that mean word length may be the better of the two features to distinguish between essays according to quality.

5.2.3.2 Entropy, Yule's K and the D estimate

Table 5.2 shows the results for entropy, Yule's K and the D estimate for each group.

Table 5.2: Mean scores for entropy, Yule's K and the D estimate for each group

Group	Entropy		Yule's K		D estimate	
	mean	sd	mean	sd	mean	sd
1	78.79	2.66	220.62	48.08	40.78	9.12
2	75.25	2.66	327.27	68.30	24.28	5.82

For these statistics, the assumed higher quality Group 1 essays showed a more diverse entropy value, a lower Yule's K value and a higher D estimate than the assumed lower quality Group 2 essays. These averages all also suggest that Group 1 essays are of higher quality than Group 2 essays. In addition, the two groups were significantly different in terms of the D estimate (t =

6.49, p < .01), Yule's K (t = 4.66, p < .01), and entropy (t = 3.26, p < .01). Again, these measures distinguish between the two groups in general terms but they are not perfect in discrimination. If we attempt to classify essays the correct groups according to each statistic, then the D estimate assigns 89% of the essays to the correct group, Yule's K assigns 83% of the essays to the correct group while entropy assigns 78% of essays to the correct group. For comparison, essay length assigns 83% of the essays to the correct group and the number of word types assigns 89% of the essays to the correct group.

These results suggest that all three measures, entropy, Yule's K and the D estimate may be useful for discriminating between essays according to quality but suggest that the D estimate may be the most sensitive.

The relationship between essay length and the three lexical diversity measures was also investigated by considering the longest essay from each group. Figure 5.1 shows entropy plotted against essay length for these two essays.

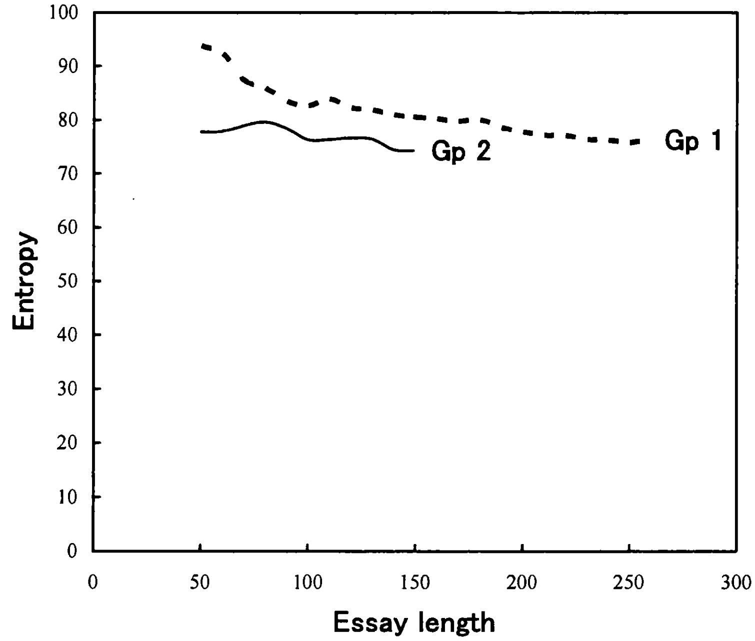


Figure 5.1: Entropy for increasing essay length for two essays

The graph seems to show that entropy is stable even at very low values but the essay from Group 1 shows a slight but steady fall in the value with increasing essay length. The essay from Group 2 also seems to follow this pattern but the fall is less clear.

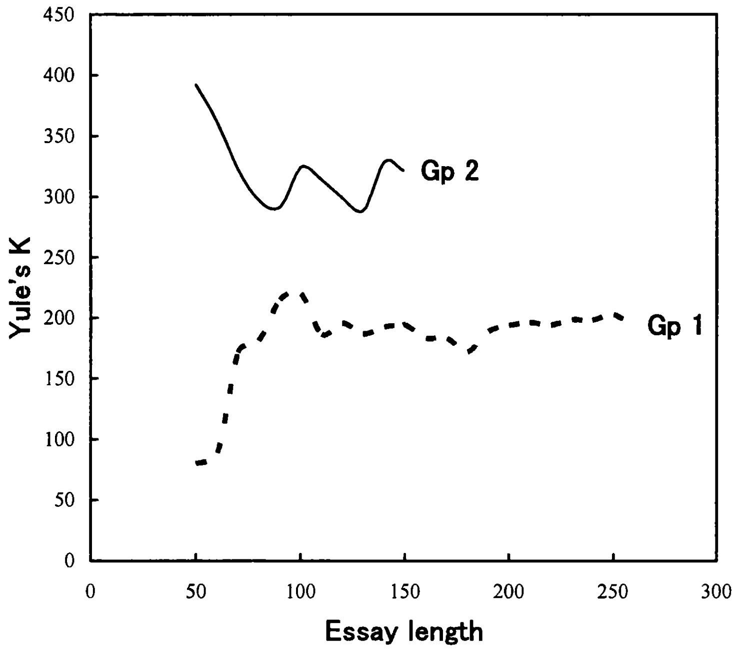


Figure 5.2: Yule's K for increasing essay length for two essays

Figure 5.2 shows Yule's K plotted against essay length for the same two essays. Yule's K seems to be very unstable for short essays less than 100 words or so. However, it seems to level out after about 100 words for one essay but there is still considerable variation in the other essay for the whole essay length. This suggests that for some essays at least, Yule's K may stabilize at quite short lengths. The graph also seems to suggest that after 100 words or so any variation consists of fluctuations rather than a relationship with essay length.

Figure 5.3 shows the D estimate plotted against essay length for the same two essays. The behaviour of the D estimate is very similar to that of Yule's K. For less than about 80 words, the D estimate is very unstable and then levels out with mild fluctuations. The D estimate seems to stabilize at a slightly shorter essay length than Yule's K and, in one essay at least, seems to show less fluctuation.

The results suggest that entropy is affected by essay length but relatively stable for short texts. On the other hand, the D estimate and Yule's K seem to be very unstable for short essays under about 80 words in the case of the D estimate and under about 100 words in the case of Yule's K. Both seem independent of length after this but are affected by fluctuations.

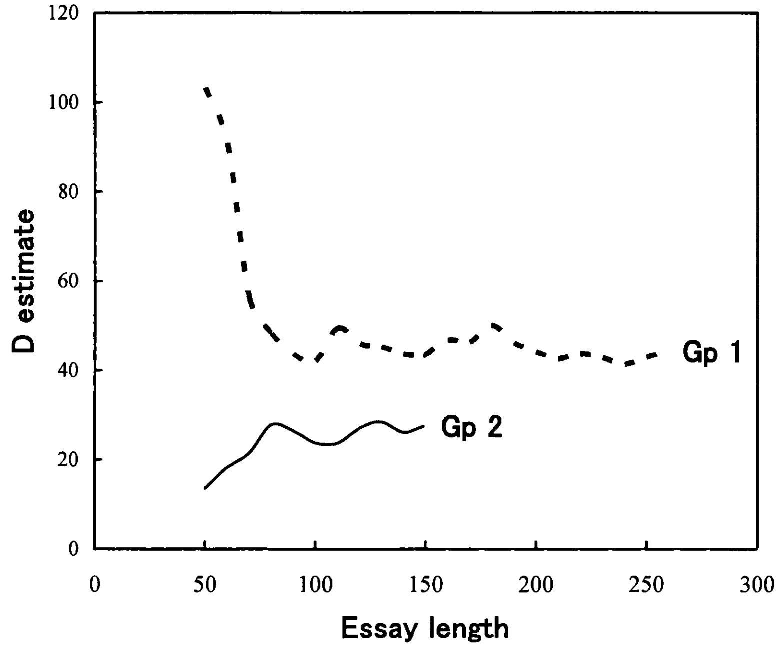


Figure 5.3: The D estimate for increasing essay length for two essays

5.2.4 Discussion

In this experiment, both mean word length and mean sentence length showed a tendency to increase with proficiency. However, only mean word length showed a significant difference between the two groups suggesting that it may be a more likely candidate for assessment. This supports the idea that learners of a higher proficiency are likely to use more infrequent words that will affect the average mean word length. In this set of essays, mean word length by itself was able to assign 71% of essays to the correct group. It is not clear how sensitive this measure is likely to be in reality even if the basic relationship is a sound one. A clear difference in mean word lengths can be seen between the 1K, 2K and UWL frequency levels. However, an essay, even of a proficient learner, will include a lot of high frequency function words. These are likely to be used often by learners of any proficiency. Also, in a task like this, there are likely to be a lot of content words specified by the context. This means that the number of words actually available to increase the mean word count may be limited. It might be worth checking whether eliminating function words and perhaps context-specific words improves the performance of the feature.

The measures of lexical diversity all showed a significant difference between the two groups.

The D estimate was able to assign 89% of essays to the correct group and Yule's K was able to assign 83% of the essays to the correct group. Entropy was less able to discriminate only assigning 78% of essays to the correct group. Yule's K and entropy are used with long texts where they are thought to be independent of essay length. The comparisons of the two essays for varying lengths of essays highlights the problems of these two measures for short essays. Entropy shows a slight but steady decrease in value with increasing essay length. Yule's K shows itself to be very unstable at less than 100 words but stabilizes after that with fluctuations. The D estimate is much steadier than Yule's K but shows a similar pattern. At lengths under 80 words, it shows considerable instability but levels out afterwards but still shows slight fluctuations. For the D estimate, the fluctuations are probably a result of the calculation which involves sampling. This may give a slightly different D score if calculated twice. The evidence seems to support claims that the D estimate is independent of essay length effects but the claim that the D estimate is stable from fifty words seems to be contradicted by these examples. For these two essays, the measure seems to stabilize from about eighty words.

It is interesting to reflect on the problem of relationships with essay length. If a feature has a positive relationship with essay length, any correlation with essay quality may be caused by an underlying positive relationship with essay length. In those cases, standardizing the feature for essay length may help identify whether a relationship actually exists. A positive relationship with essay length is often seen with simple raw counts of features. Correlation with quality may disappear when standardized for essay length. On the other hand, if a feature shows a correlation with quality but has an inverse relationship with essay length, this may mean that this feature is correlating with quality despite being negatively affected by length. The correlation we see may be an underestimate of the true amount. Standardizing for essay length helps us to find the true extent of the correlation. This also means that some measures show little correlation with essay quality because any correlation is being annulled by essay length effects. This may be the case with measures of lexical diversity that often show weak or no correlation with proficiency in raw form, but when the negative influence of essay length is removed, the correlation may be significant. In this experiment, entropy correlates with proficiency even before adjusting for length. The inverse relationship with essay length suggests that this correlation may be underestimated. If essay length was standardized to the length of the shortest essay, a stronger correlation may emerge. However, with low-level learners, the shortest essay can be 100 words or less. It is not clear how stable the entropy value would be at that level.

5.2.5 Conclusion

If we assume that the essays in Group 1 are of higher quality than those in Group 2, the results of this experiment support the view that mean word length may be an indicator of essay quality, with higher values predicting higher quality and lower values predicting lower quality. There is some evidence that mean sentence length is also longer for better essays but this result was not significant so we must conclude that mean sentence length may not be such a good indicator. The measures of Yule's K, entropy and the D estimate all distinguished between essays of the different groups. Analysis based on two essays suggests that the D estimate and Yule's K may be unstable at low levels under about 80 words and 100 words respectively. At longer lengths, both the D estimate and Yule's K seemed to stabilize and be unaffected by essay length but subject to slight fluctuations. Entropy showed a slight inverse relationship with essay length but no evidence of instability at low levels. This suggests that standardizing entropy for essay length may improve performance.

5.3 Word distribution properties of L1 and L2 texts

5.3.1 Introduction

This experiment concentrates on some vocabulary distributions. Zipf (1932) identified a number of general properties of word distributions. One of these properties is the distinctive shape of word frequency distributions. These frequency distributions usually have many words occurring once and progressively fewer words occurring more than once and very few words occurring many times. Another property is observed when words are ranked according to their frequency. Zipf showed that occurrences of these ranked words decrease by factors approximating the power law. A third property involves the inverse relationship of word lengths with frequency. High frequency words tend to be short while infrequent words tend to be longer. In order to investigate application of word distributions to evaluation of L2 essays, various Zipfian distributions of learner and native speaker texts are examined.

5.3.1.1 Aims of the experiment

The aims of this experiment are to see if there are differences between essays of L2 learners and native speaker writing samples in terms of:

1) word frequency distributions.

2) ranked word frequency distributions.

3) word length distributions.

5.3.2 Methodology

5.3.2.1 Participants and tasks

Eleven essays produced by high beginner L2 learners of English were compared with five native speaker written samples. The learner essays were reviews of movies, books, music or restaurants. These essays varied in length from 269 to 418 words Learners wrote their essays in their own time and were free to use whatever resources they wished. Learners submitted their essays electronically via email. Native speaker samples on matching subjects were also collected (see Appendix 5.1 for sample learner essays and native speaker sample). To control for length effects, analysis was performed on samples of text of equal length. The length of the shortest essay, 269 words, was used as a standard and the first 269 words of each essay were used in the analysis. Samples of the same length were taken from similar reviews written by native speakers.

Comparisons of learners with native speakers may not be ideal for L2 assessment purposes but there are some advantages. Firstly, they maximize the chances of finding differences related to quality. Secondly, comparisons can be made without getting L2 learner essays rated. The dangers need to be acknowledged too. It is not necessarily the case that the same features that differentiate between native samples and learner essays will also discriminate between learner essays of differing quality. The idea that features will develop in a linear fashion from low level learners through high level learners to native speakers may be naïve. For example, Larsen-Freeman & Strom (1977) showed that even within learners, there was no linear relationship between the number of errors and essay quality.

5.3.2.2 Analysis

A specially designed computer program, L-distribution (see Appendix 1.1) was used to analyze the essays. For word frequency distributions, the occurrence of each word type in the essay was calculated and then occurrence totals collated, so that the number of words occurring n times in an essay could be calculated from n = 1 to the maximum occurrence of a word in any essay.

For the ranked word distribution, the word types were ranked according to occurrence in the essay from one to v where v is the number of types in the essay. Rank one was the word occurring the most in the essay and rank v was the word occurring least. The rank of these words was graphed from 1 to v against the occurrence of that word.

For the word length distribution, the number of words of length n was calculated from n = 0 to s where s was the length of the longest word in the essay. This was then graphed for n = 0 to s against the number of tokens of that length.

5.3.3 Results

5.3.3.1 Word frequency distribution

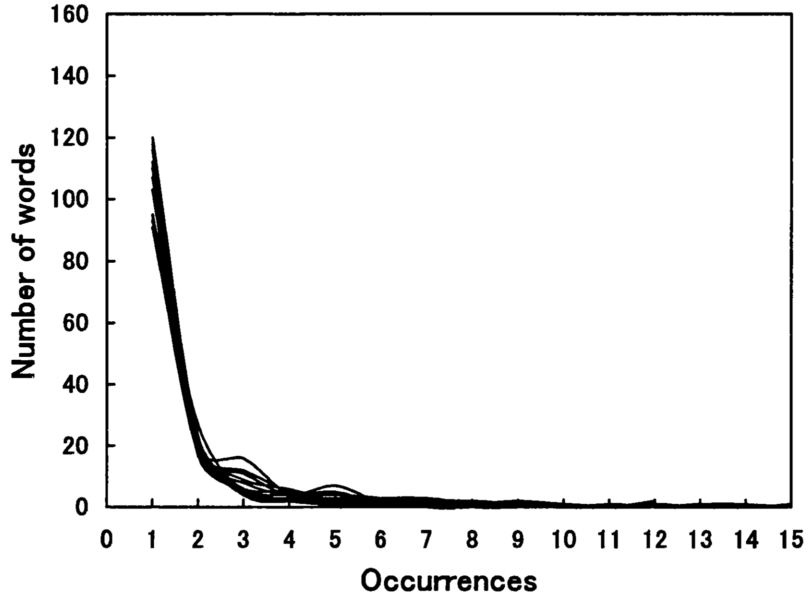


Figure 5.4: Word frequency distribution for learner essays

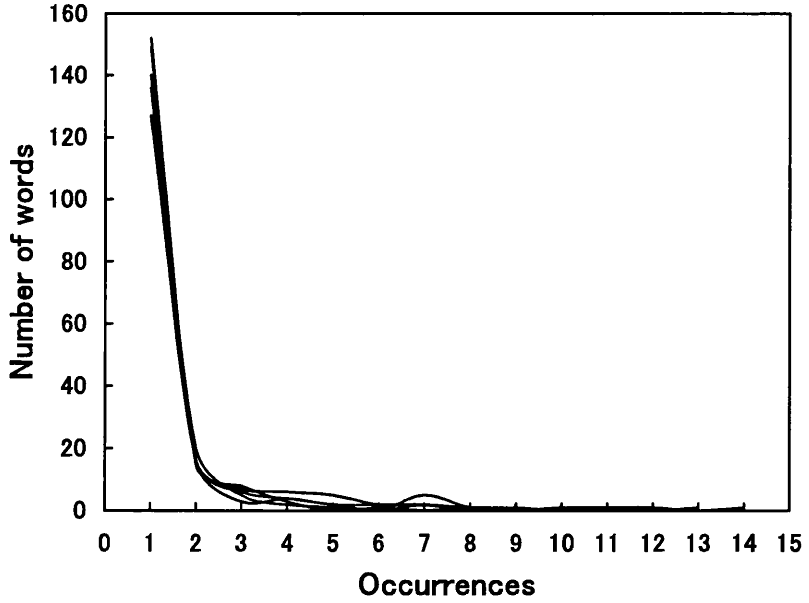


Figure 5.5: Word frequency distribution for native samples

Figure 5.4 shows the number of words occurring n times for each of the eleven learner essays while Figure 5.5 shows the same for the native samples. The distributions show a typical shape with a large number of words occurring once (hapax legomena) and then progressively smaller numbers of words occurring at higher frequencies. The graphs highlight the problem of using distributions for comparisons of large numbers of texts. It is difficult to see any discernable difference between the distributions once there are more than a few cases involved. There does, however, seem to be one clear difference between the learner essays and the native samples. The learner essays seem to have fewer hapax legomena than the native samples.

In the learner essays, the number of hapax legomena peaks at around 120 words but for the natives, the number goes as high as 150. This is more clearly shown in Figure 5.6. In fact, there is not only a significant difference between the two groups of natives and learners but also an absolute difference with all the native samples having more hapax legomena than all of the learner essays. The fewest hapax legomena amongst the native samples is 127 (47.2%) and the most in the learner essays is 120 (44.6%). The difference in occurrence of hapax legomena stands in contrast to the occurrence of hapax dislegomena (words appearing twice and twice only in a text) which shows no obvious difference between native speakers and L2 learners.

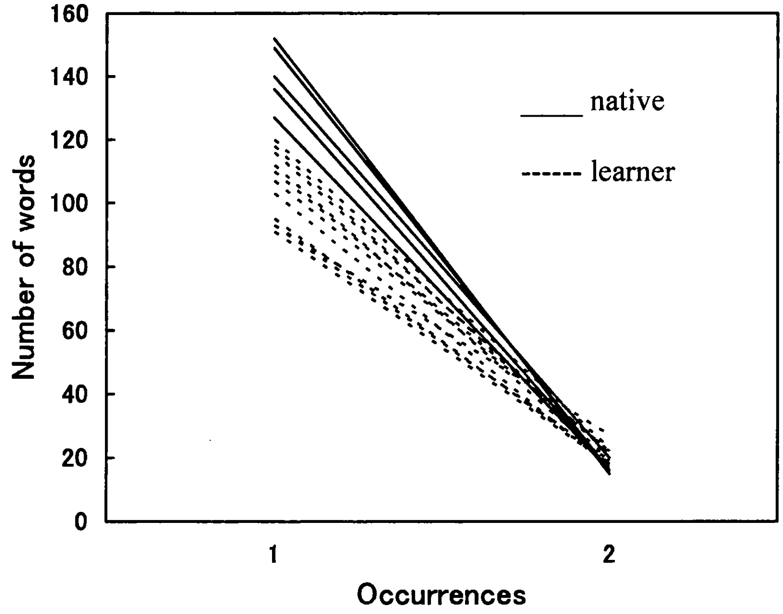


Figure 5.6: Hapax legomena for native and learner essays

This suggests that hapax legomena may be a feature that can help predict whether an essay is written by a learner or a native.

5.3.3.2 Word rank distributions

For each text, words were ranked according to their frequency in that text. Rank one indicated the most frequently occurring word. Figure 5.7 shows distributions of these Zipf ranked words versus frequency at ranks from one to fifteen. Although there are some differently shaped distributions, there are no clear differences in distributions between native samples and learner essays.

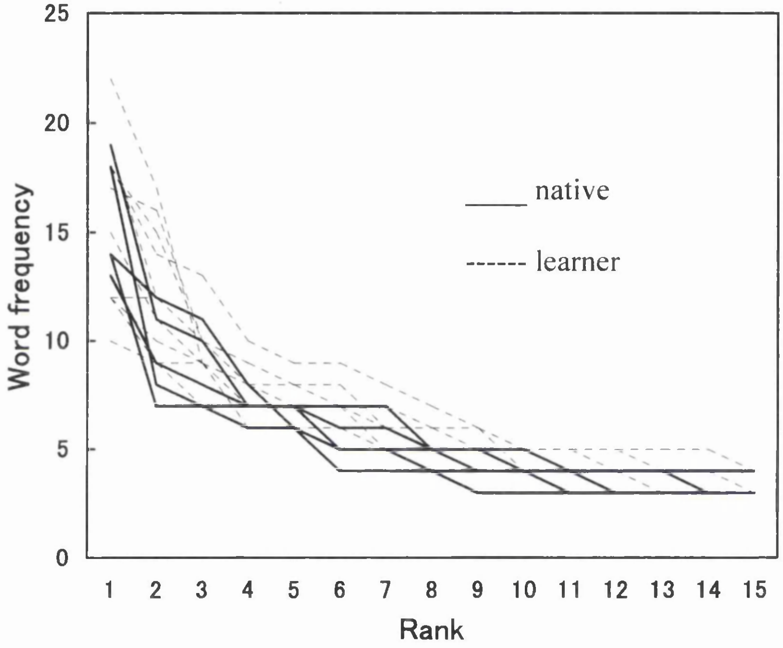
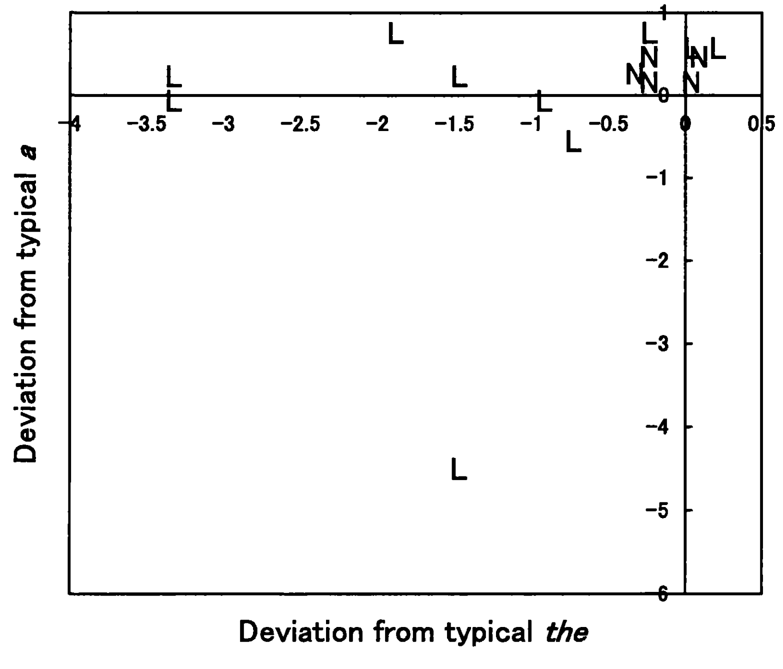


Figure 5.7: Zipf ranked words 1 to 15

Table 5.3 shows the actual words that are ranked one to five for each essay or sample. Here trends seem more obvious. In the table, N1 to N5 are the native samples and L1 to L11 are the learner essays. *The* is the top-ranked word for all the native samples but the top-ranked word in only three out of eleven learner essays. In fact, in four learner essays, *the* does not appear in the five top ranked words. Also, *a* appears in the top three words in all native samples but in the top five words of only four of eleven learner essays. According to the British National Corpus (BNC) (Leech, Rayson & Wilson, 2001), *the* and *a* account for 6.44% and 2.20% respectively of all words in written English. Figure 5.8 shows deviations from the average occurrence of *the* and *a* for all texts. All the native samples and some learner essays have only slight deviations from the expected values but many learner essays have quite large deviations.

Table 5.3: Top five Zipf ranked words for each text

Essay	Ranked word				
	1	2	3	4	5
N1	the	a, is		Pete's, that	
N2	the	a	and	of, to	
N3	the	in	a, he, is, of, to		
N4	the	is	a	and, in	
N5	the	in	a	and	his
L1	is	he	in	move	Leon
L2	her	she	in	is, Josie	
L3	she	in	and	a	album, of
L4	the	is	and	Disney, songs, you	
L5	is	the	and	of	a, he
L6	they	the	and	in, music, them	
L7	and	in, the		is	she, to
L8	the	and	he	in	a, to
L9	they	and, the		have, this, to	
L10	of	was	in, it, to		
L11	the	is	of	spaghetti	a, there, to



Note: One learner deviation is so great that it is not covered by the graph

Figure 5.8: Deviations from typical occurrence of *a* and *the*

As with word frequency distributions, it is difficult to assess differences in patterns between native samples and learner essays by considering the ranked word graphs. However, some features show potential to help differentiate between them such as the top-ranked words, or the relative occurrence of frequent words such as *a* and *the*.

5.3.3.3 Word length distributions

Figure 5.9 shows the distribution of word lengths for each essay or sample.

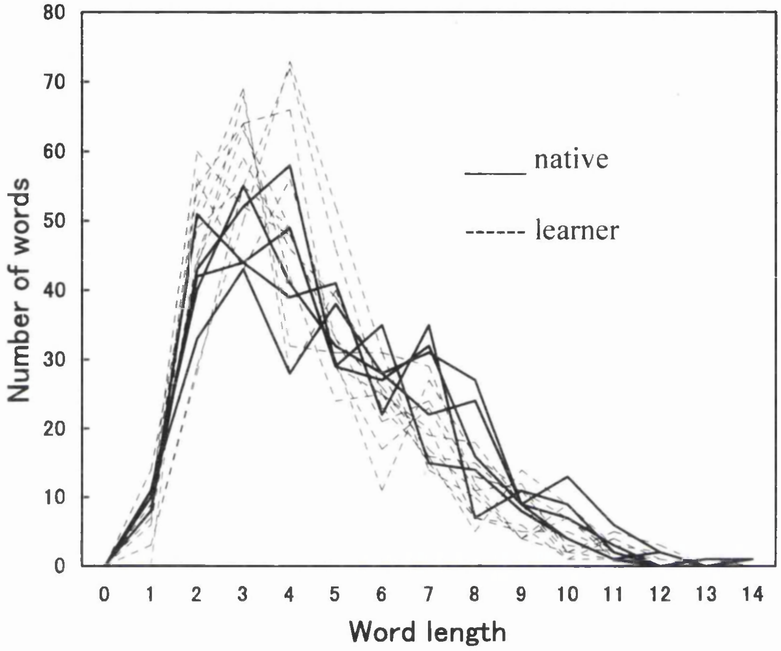


Figure 5.9: Distribution of word lengths for native and learner essays

Although some learner essays have higher numbers of words at the length 3 and 4 level and some native samples have higher numbers of words at the length 7, 8 and 9 level, the distributions do not show any obvious patterns of differences. However, Table 5.4 shows that the mean word lengths for native and learner groups are significantly different. This seems to support the evidence from the previous experiment that better essays are likely to have a higher mean word length. Although, the groups of native samples and learner essays are significantly different in relation to mean word length, they are not exclusively different. Some learner essays have higher means than one of the native samples. Mean word length was also calculated for only hapax legomena and showed the same pattern.

Table 5.4: Mean word length for native and learner texts

	Native samples	Learner essays
Mean word length	4.75	4.40
sd	0.26	0.14
high	5.24	4.53
low	4.48	4.08

5.3.4 Discussion

The results of this experiment seem to suggest that while graphs are a good visual way of appreciating distributional features for small numbers of essays, they are difficult to compare with larger numbers of essays. However, in this experiment, it has been shown that various distributional features show some patterns of difference between native samples and learner essays. There was consistently a higher proportion of hapax legomena in native samples than in learner essays. It may be that the proportion of these words only occurring once in an essay could be an indicator of the vocabulary available to the writer.

For Zipf ranked words, *the* and *a* were ranked highly in all native samples. In all samples, *the* was the top ranked word and *a* was in the top three ranked words. However, *the* was only top ranked in three out of eleven learner essays and in some essays did not rank in the top five ranked words. The relative occurrence of *the* and *a* in native samples seemed to be consistent with the relative occurrence of these words in general English according to the BNC. On the other hand, in some learner essays, the occurrence of these words was quite different from what we find in general English. These two words may be useful because of their high frequency in English. This should make them more reliable indicators than other less frequent words. Relative occurrence of words has been used extensively in author attribution studies. For example, Mosteller & Wallace (1984) used relative occurrences of various synonyms to make a prediction of authorship for *The Federalist Papers*. In learner essays, few words can be used for this kind of analysis because of reliability concerns related to low frequency items. *The* and *a* may be exceptions. Another possibility is relative occurrences of classes of words that can be counted automatically, for example, content words, function words, conjunctions, pronouns etc. The proportions of articles, pronouns and conjunctions have all been shown to correlate moderately with quality assessments by Evola, Mamer & Lentz (1980).

As the results of the previous experiment suggested, there seems to be the potential for using mean word length to differentiate between essays according to quality. However, the scale of the differences found here were not as conclusive as we might have hoped for. It might be that mean word length is not a sensitive enough measure given the small amount of words available in most learner essays.

As mentioned in the introduction, one needs to be cautious about generalizing results from comparisons of native and learner texts to differences between essays from learners of different proficiencies. It is not clear how effective these features would be for evaluating differences in quality within learner essays. For example, the incidence of *the* and *a* would probably be less effective in discriminating between learner essays than in identifying native speaker essays. However, these comparisons may have highlighted some features for further investigation. The properties of Zipf word distributions may be developed further for assessment. Meara & Miralpeix (2007b) have developed a computer tool based on a Zipf distribution for predicting productive vocabulary size given a sample of writing.

5.3.5 Conclusion

This very simplistic analysis of various Zipfian word distributions suggests that there are various useful properties of word distributions that may be used to aid classification of essays into native and non-native essays. In this limited study, the proportion of hapax legomena, the rank and relative proportion of *a* and *the*, and mean word length all showed promise for distinguishing between native and learner essays. It may be that some of these features may be useful in determining proficiency of learner essays. Hapax legomena is a strong possibility. It is quite possible that the proportion of hapax legomena increases as essay quality increases and so the number of hapax legomena may act as a proxy measure for vocabulary size. In the case of relative proportion of *the* and *a*, it seems less likely that it would be useful for determining essay quality at lower levels of proficiency. However, the relative proportions of some easily countable classes warrants further research.

5.4 The type-token ratio and lexical variation

5.4.1 Introduction

This experiment investigates the performance of two lexical diversity ratios. The first is the type-token ratio (TTR) and the second is lexical variation (LV). The TTR is the number of types in the essay divided by the number of tokens while lexical variation includes only lexical items as follows:

TTR = types / tokens LV = lexical types / lexical tokens

If lexical words can be equated with content words then the TTR includes content and function words while lexical variation includes only content words. Calculating lexical variation is not as straightforward as calculating TTR since decisions have to be made about what constitutes a lexical item. This is a relatively simple task for a human but not so simple to do automatically. An exhaustive list of lexical items for a computer program to check cannot be compiled easily but it is still possible to do indirectly by compiling a list of function words and calculating the number of function types and tokens for an essay and then getting an estimate of lexical types and tokens by subtracting function types and tokens from total types and tokens.

5.4.1.1 Aims of the experiment

The aims of this experiment are to:

1) test the reliability of TTR and lexical variation over two tasks written by the same low level L2 learners.
2) test how TTR and lexical variation predict quality of L2 learner essays.

5.4.2 Methodology

5.4.2.1 Participants and tasks

Two short essays were written by 14 low level L2 learners of English based on the cartoon tasks in Appendix 3.1 and 3.2. These essays were holistically rated into two groups, *Good* and *Poor*. Eighteen were graded *Good* and ten *Poor*.

5.4.2.2 Analysis

The essays were analyzed by a specially designed computer program,L-analyzer (see Appendix 1.1), and the two lexical statistics were calculated. TTR can be calculated directly using a simple count of the essay length in words and the number of different word types. To make a decision about lexical words is difficult so estimates of lexical types and tokens were calculated indirectly using a list of common function words. These function words used were taken from the most commonly occurring words in the BNC (Leech, Rayson & Wilson, 2001). See Appendix 5.2 for a full list of function words used in the analysis.

5.4.3 Results

5.4.3.1 Reliability over two tasks

The essays varied in length from 103 to 218 words with an average of 149 words. The values for TTR and lexical variation are shown in Table 5.5. Lexical variation was higher than TTR in all but two texts (26 out of 28 cases). Lexical variation ranged from 0.32 to 0.65 with an average of 0.51. TTR varied from 0.34 to 0.55 with an average of 0.45. Both values had a similar average and range over the two tasks.

Table 5.5: Mean, maximum and minimum values for TTR and LV

	TTR			LV		
	av.	hi	lo	av.	hi	lo
Task 1	0.45	0.53	0.34	0.52	0.64	0.36
Task 2	0.44	0.55	0.35	0.51	0.65	0.32
r		0.79**			0.68*	

**significant at the .001 level *significant at the .01 level

The greatest difference of lexical variation over TTR was LV = 0.64 to TTR = 0.48. The greatest difference of TTR over lexical variation was TTR = 0.35 to LV = 0.32. The two cases of TTR higher than lexical variation both occurred in essays rated *Poor*. It is not surprising that values of lexical variation tend to be higher than for TTR since the words that occur most frequently are likely to be function words. Function words are more likely to occur multiple times than content words. Thus when they are subtracted they affect the denominator of the ratio more than the numerator making the LV larger than the TTR. For example, one of the essays contained 125 tokens and 66 types. The word that occurred most often in the essay was a function word, *the*, occurring 13 times. Just taking this one function word out increases the ratio because we lose one type but 13 tokens as follows:

TTR = 66/125 = 0.53 TTR(without *the*) = 65/112 = 0.58

Where a lexical variation measure is larger than a TTR measure, the most frequently occurring words are likely to be content words rather than function words. This may indicate over repetition of nouns rather than use of pronouns. This is likely to be a feature of a poor quality essay.

In regard to reliability over the two tasks, the scores for both statistics correlated significantly over the two tasks. TTR was more reliable than lexical variation with r = 0.79 compared with r = 0.68 for lexical variation.

5.4.3.2 Relationship to quality assessments

To investigate the predictive performance of these two measures we can compare them against the holistic ratings. If we envisage the holistic ratings as a continuum where the top 18 are the essays rated *Good* and the last 10 rated *Poor* then we can chart the success of lexical variation and TTR in predicting quality as in Table 5.6.

Table 5.6: Ratings of essays ordered by lexical score

Holistic ratings

GGGGGGGGGGGGGGGGGGPPPPPPPPPP

Lexical variation

GGGGGGPGGGGPGPGPGGPGGGPGPPPP

TTR

GGGGGGGPGPPGGPGGGGGPPPGPGPPG

Both lexical measures correctly categorized 71% of the essays as being good or poor. However, taking order of values into account, lexical variation corresponded slightly more than TTR with holistic ratings of the essay. This representation shows that in the case of lexical variation, the six highest values corresponded to essays rated as *Good.* The seventh highest value of lexical variation was rated *Poor* but the next four were also rated as *Good.* The lowest four values of lexical variation were rated *Poor.* In the case of TTR, the highest seven values corresponded to essays rated *Good* but also the lowest value of TTR was rated *Good.* High values of both lexical variation and TTR seemed to indicate *Good* rated texts. However, low values of lexical variation better indicated *Poor* texts than low values of TTR (highlighted by the 2 *Poor* cases where lexical variation was lower than TTR). Middling values of TTR and lexical variation seemed to be little use in defining whether an essay was *Good* or *Poor.*

5.4.4 Discussion

Although this is a very small scale study, it produced some interesting observations. Firstly, it suggests that TTR may be more reliable than lexical variation over two tasks by the same learners. This may be because the use of function words is more consistent than the use of content words. The improved reliability may also be because simply more words are involved. This may be another consideration when deciding whether to use only content words or to include function words. Many measures of lexical diversity use a smaller class of words rather than just types or tokens. Lexical variation uses lexical types and lexical tokens. Selection of content words rather than function words is also an issue with some other features. For example, in the first experiment of this chapter, it was speculated that mean word length might be improved by focusing on content words because function words may contain noise. In some cases, content words may focus on the words where learners are likely to be most different. However, a counterargument may be that if the function words are not completely noise, they may help in the analysis by providing more words.

Secondly, it shows that even two very similar measurements may have different domains of usefulness when it comes to assessment. The results of this experiment suggest that high values of both TTR and lexical variation may be a strong indicator of a high quality essay, whereas a low value of lexical variation may, in some cases, be a better predictor of a low quality essay than a low value of TTR. Middling values of either feature may be less useful in predicting quality.

Both measures do surprisingly well in predicting quality with these essays given their dependence on essay length. Because many low level learner essays tend not to vary much in length, there may be many occasions where they can be used in their raw state. Controlling for length is a way to improve performance and may be necessary in cases where there is considerable variation in length of essay. Then TTR or LV could be used on a sample of words so that TTR(n) and LV(n) would indicate that each measure was based on a sample of n words. However, one concern might be that if length is controlled for, then the shortest essay is usually set as the standard. Where the shortest essay is very short, this may be a problem. In an automatic analysis, it might not always be best to standardize to the shortest length if it is very short. It might be best to decide a minimum length under which shorter essays are treated in a different way.

5.4.5 Conclusion

This experiment investigated the reliability of lexical variation and TTR for two sets of essays written by L2 learners. Both measures scored relatively highly for reliability. Lexical variation had a correlation of r = 0.68 and TTR a correlation of r = 0.79 over the two tasks. This study also investigated the relationship of these measures with essay quality. Both measures performed relatively well at predicting quality with a 71% decision agreement with the holistic ratings. However, there is a suggestion that these measures operate more reliably in different domains. Both high values of lexical variation and TTR seem to indicate good quality essays while low values of lexical variation may be better at predicting poor essays than low values of TTR. This is a very small study so the results are highly speculative and should be viewed with caution but could form the basis for more research.

5.5 LFP and P_Lex

5.5.1 Introduction

This brief experiment investigates the relative performances of P_Lex (Meara & Bell, 2001) and the LFP (Laufer & Nation, 1995) for two sets of L2 essays produced by low-level learners.

These measures are referred to as extrinsic measures of lexical diversity by Meara & Bell (2001). They measure lexical diversity against a benchmark of frequency in general English which may give a guide to the difficulty or sophistication of the vocabulary used. These may be more useful in gauging quality of an essay than lexical diversity measures based only on types and tokens which do not inform about the quality of the lexical items. However, because these measures depend on benchmark measures external to the essays, it is important that they correspond to the learners and to the task. If low level learners do not produce a range of vocabulary that can be identified by these measures, then they may not be effective. Similarly, if a task does not elicit a wide enough range of vocabulary, the measures may not be effective. This experiment investigates the performance of the two measures for low level L2 learners of English.

5.5.1.1 Aims of the experiment

The aims of the experiment are to:

1) test the reliability of LFP over two tasks written by the same low level L2 learners.

2) test the reliability of P_Lex over the same two tasks.

5.5.2 Methodology

5.5.2.1 Participants and tasks

The two sets of essays based on the cartoon tasks from the previous experiment were used (see Appendix 3.1 & 3.2 for tasks).

5.5.2.2 Analysis

The RANGE computer program (Heatley, Nation & Coxhead, 2002) was used to calculate the LFP for each essay on each task. As part of the preparation of files for input into the program, spelling error words were deleted.

Table 5.7: ***Not in lists*** **words for Tasks 1 & 2**

	Not in lists words
Task 1	beach (2), ok, passion, pants (2), smart, boring, amazing, bog*, sharks, smily, cute, hobby, jacket
Task 2	crazy, carpenter, amazing, ammonia, pole, restroom, zone, impatient (2), smart, stair, toilet (2), ok

*probably a spelling error of *dog*

Range produces output in four categories: *1K, 2K, UWL* and *not in lists*. Laufer & Nation (1995) suggest using three categories of LFP for low level learners. These are *1K*, *2K* and *any other* *words*. This means amalgamating the categories of *UWL* and *not in lists* categories to get the *any other words* category. This decision was backed up by the analysis which showed a total of only eight words from the fourteen candidates from the *UWL* for Task 1 and a total of only two for Task 2. There was a total of 15 *not in lists* words for Task 1 and 14 for Task 2 but these include words repeated in several essays. The *not in lists* words are shown in Table 5.7.

To calculate P_Lex, the P_Lex computer program (Meara, 2007) was used. The ratio of *difficult* words to *easy* words was also calculated. *Difficult* and *easy* words were defined in the same way as for P_Lex. *Easy* words were words in the 1K category, and proper nouns while *difficult* words were words beyond this category.

5.5.3 Results

5.5.3.1 LFP

The LFPs for all learners are shown in Table 5.8. The profiles all showed a high proportion of words from the 1K category. For Task 1, the percentage of words in the 1K category ranged from 83.3% to 96.6% with an average of 90.3%. The percentage of words in the 2K category varied from 2.9% to 13.3% with an average of 7.1. The percentage of *any other words* varied from 0% to 4.6% with an average of 2.6%.

Table 5.8: LFPs for 14 learners on two tasks

Student	High/Low	Task 1			Task 2		
		1k	2k	other	1k	2k	other
1	H	88.2	8.8	3	91	7.5	1.5
2	H	92.6	5.3	2.1	95.6	3.3	1.1
3	H	93.3	5	1.7	96.4	2.4	1.2
4	H	95.7	2.9	1.5	90.8	7.9	1.3
5	H	96.6	3.5	0	94.3	2.9	2.9
6	H	88.6	8	3.4	87.5	8.3	4.2
7	H	87.8	10.8	1.4	93.6	4.8	1.6
8	H	94.1	5.9	0	93.1	4.2	2.8
9	H	84.6	10.8	4.6	94.4	1.9	3.7
10	L	90.4	5.8	3.9	95.3	4.7	0
11	L	87.8	8.2	4.1	95.1	2.4	2.4
12	L	88.5	7.7	3.8	94.4	5.6	0
13	L	92.7	3.6	3.6	92.9	5.4	1.8
14	L	83.3	13.3	3.3	95.6	4.4	0

For Task 2, the percentage of words in the 1K category seemed a little higher than for Task 1. Percentages ranged from 87.6% to 96.4% with an average of 93.6%. The percentage of words in the 2K category varied from 1.9% to 8.3% with an average of 4.7%. The percentage of *any* *other words* varied from 0% to 4.2% with an average of 1.8%.

To test the reliability of LFP over the two tasks, the proportions of words in each category were compared independently over the two tasks and the results are shown in Table 5.9.

Table 5.9: Reliability statistics for LFP over Task 1 & 2

	1K words	2K words	other words
reliability	-.07	-.08	-.11

The LFP percentages by category on the two tasks show no obvious correlation. This suggests that the LFP may not be suitable for low level learners on this type of task.

5.5.3.2 P_Lex

P_Lex scores and the ratio of difficult to easy words for each task are shown in Table 5.10. The first impression is that P_Lex scores were low. In fact all essays had a P_Lex score less than 1 and some were close to zero. The mean P_Lex score for Task 1 was 0.38 and for Task 2 was 0.22. There was no correlation between the two sets of learner essays for P_Lex (r = -0.02). The values for the ratio of difficult to easy words were also very low with a mean of 0.09 in Task 1 and 0.04 in Task 2. There was no evidence of correlation between the two tasks either (r = -0.09). However, TTR showed a correlation of r = 0.79 over the two tasks with a mean TTR of 0.45 in Task 1 and 0.44 in Task 2.

Table 5.10: P_Lex, and difficult/easy words for Tasks 1 & 2 (T1 & T2)

	P_Lex		Difficult/Easy	
	T1	T2	T1	T2
Mean	0.38	0.22	0.09	0.04
s.d	0.22	0.18	0.06	0.02
high	0.88	0.71	0.19	0.08
low	0.02	0.02	0.02	0.02
r		-0.02		-0.09

5.5.4 Discussion

The results of this experiment raise doubts about the appropriateness of both the LFP and P_Lex for very low level learners or for this type of task. Whether the poor results are due to the proficiency of the learners or the task involved is not clear. It is probably due to both to a certain degree. Many of the essays contained no words from the upper two levels of the LFP. In fact, most of the words came from the lowest level. The same seems to be true for P_Lex. P_Lex scores were very low because of a relative shortage of difficult words. Meara & Bell (2001) report ranges of P_Lex from 0 to 4.5 and above. However, the highest value in this study was still less than one. A comparison of various learner and native speaker texts suggests that the higher values in the range stated by Meara & Bell range may only be found in specialist L1 texts, for example, an academic paper displayed a P_Lex score of 3.5 and a newspaper sports report displayed a P_Lex score of 2.5.

The lack of reliability over these two similar tasks may be due to the context sensitivity of P_Lex or it may be due to the inherent unreliability associated with low totals. It could be that in this case, the differing contexts within the two cartoons may call for different sets of vocabulary which may be at different frequency levels. There is some evidence for this with mean P_Lex scores on Task 1 being higher than on Task 2. However, another method of assessing reliability, the split halves method, suggests that the lack of reliability is not only due to the differing contexts. Since many of these essays were quite short, a split halves analysis was not possible. However, a split half analysis of some of the longer essays found cases where the two halves produced quite different P_Lex scores.

However, the poor performance on this experiment should not detract from P_Lex and LFP as measures. Meara & Bell (2001) and Laufer & Nation (1995) have shown them to be effective with higher proficiency learners and different tasks. Meara & Bell have noted the context specific effect of P_Lex and argue that this characteristic could be a strength making it useful for evaluating the difficulty of tasks. The problem of poor tasks could be controlled for by pre-testing tasks to ensure they involve sufficient vocabulary from each level. The low scores in this experiment may be a warning that this type of task may not elicit enough of a range of vocabulary for assessing learners.

In contrast to the poor performance of LFP and P_Lex, measures of intrinsic lexical diversity seemed to perform well on these tasks in the previous chapter. Both TTR and lexical variation showed no difference in means over the two tasks and scores were highly correlated across the two tasks.

5.5.5 Conclusion

In this experiment, the performance of two extrinsic lexical diversity measures, LFP and P_Lex was compared on two written tasks by low level learners. Both measures seemed inappropriate for this type of learner who produced words overwhelmingly in the first 1000 words. LFP was highly skewed to the lower 1K word category. Also P_Lex scores were very low. Neither measure showed reliability over the two tasks which was probably because the small numbers of words outside the first 1000 words were too few in number to be measured reliably. The lack of diversity may not be entirely a function of the learners. It is possible that the task does not encourage a diversity of vocabulary. Although not detracting form the usefulness of these measures which is well-documented elsewhere, this experiment shows the importance of using them with appropriate learners and tasks.

5.6 Conclusion

In this chapter, four exploratory studies have revealed some features that may be applied to predicting essay quality in L2 learners. However, these were all very small scale studies and so the results are far from conclusive but may identify areas for further research. In the first experiment, mean word length was found to correlate with essay quality. Higher mean word lengths seemed to indicate better essays and lower values seemed to indicate poorer quality essays. Given Zipf's rule that more frequent words tend to be shorter, mean word length seems to be an intuitive approach to measuring vocabulary sophistication. One concern, however, is whether this measure is sensitive enough to illustrate differences given the small number of content words in a typical learner essay. In addition, this study showed that measures such as Yule's K and entropy which were thought to be unstable for short texts performed surprisingly well on short L2 essays.

In the second experiment, an analysis of various word distributions revealed some features that may be able to help discriminate between native and learner essays. The proportion of hapax legomena and the relative occurrence of the words *the* and *a* seemed most effective. Whether these features could be used with L2 essay assessment is not clear. The case for the proportion of hapax legomena as a proxy measure for vocabulary size may be appealing but the relative occurrence of *the* and *a* seems less applicable to purely L2 contexts. However, relative occurrences of other features or word classes may be worth investigation.

In the third experiment, lexical variation and the TTR were shown to be relatively reliable over repeated tasks by the same learners. They also both seemed to be good at discriminating between good and poor rated essays. The results also suggested that lexical variation might be better at indicating poor quality essays. This showed that in some cases, concentrating on content words may give more precise information. The flipside is that by eliminating words, reliability may be compromised.

In the final study, LFP and P_Lex were found not to be reliable over two simple writing tasks by the same learners. This suggests that we need to be very careful when selecting measures to ensure that they are appropriate for learners and tasks. In particular, measures which depend on external criteria such as frequency bands need to be calibrated to fit the learners and the tasks. Comparing these results with those in the previous experiments suggests that intrinsic measures perform better with essays of lower level learners both in terms of reliability and predicting essay quality.

This chapter concludes the first part of the experimental work. In these three chapters, various features have been considered for a role in automatic assessment. Many show some promise but none seems powerful enough to be considered as a single measure of essay quality. In the next three chapters, the focus of the experimental work shifts to finding systems where complementary features can be used together. In the next chapter, a simple two-dimensional model of quantity versus content incorporating essay length and lexical diversity is considered.

Chapter Six: A quantity/content model

6.1 Introduction

In the previous chapters, various features and their relationship with essay quality were examined. Although many features showed a strong association with essay quality, none was strong enough to be able to entirely account for it. In the next three chapters, the emphasis is on finding a system to account for essay quality which incorporates various features.

This chapter explores a very simple possible model for essay assessment. Plenty of studies have shown that both essay length and lexical diversity can be strongly correlated with essay ratings. However, despite significant correlations, these features are by themselves inadequate both empirically and theoretically to account for essay ratings. In this experiment, we consider whether together they may be able to better account for essay ratings. To this end, we investigate a simple two-dimensional quantity/content model of essay assessment. Quantity is represented by essay length and content by lexical diversity.

6.1.1 Aims of the experiment

The aims of this experiment are to see:

1) if a two-dimensional quantity/content model employing essay length and lexical diversity can predict assessment of these essays better than a single dimension of either quantity or content.

2) which measure of lexical diversity works best as the content dimension alongside the quantity dimension of essay length for this set of essays.

6.2 Methodology

6.2.1 Participants and tasks

A set of 34 essays written in a timed format was collected. The participants were third year English majors at a Japanese university. The task involved writing an essay in thirty minutes given the following prompt (See Appendix 6.1 for sample essays):

Watching television is bad for children. Do you agree or disagree with this statement? Use specific reasons and examples to support your answer.

6.2.2 Essay ratings

The essays were assessed by a native speaker judge as one of five categories: *good, above average, average, below average,* and *poor*. The number of essays for each rating is shown in Table 6.1.

Table 6.1: The number of essays for each rating category

Rating	*good*	*above average*	*average*	*below average*	*poor*
No. of essays	4	5	18	6	1

6.2.3 Analysis

6.2.3.1 Selection of quantity and content measures

A simple two-dimensional model of quantity and content was used. The two dimensions were given equal weight. Quantity was represented by essay length in words and content was represented by lexical diversity. In addition, several lexical diversity measures were evaluated. There were two considerations in selecting these measures. One was that they have been shown to correlate with essay ratings in other studies. Another was that they were easy to calculate and could be embedded in a simple computer program. This second condition precluded using a measure such as LFP because the profile output is difficult for a computer program to process and use with other measures. It also precludes using P_Lex because it needs a special computer program to calculate it. The same was also true for Malvern & Richards' D estimate. However, it was included in this analysis for comparison in its role as the most commonly used measure of lexical diversity. The measures considered in this analysis were as follows:

1) TTR(100), the number of word types in a hundred word sample

2) Guiraud Index

3) Yule's K

4) the D estimate

5) Hapax(100), the proportion of hapax legomena in a hundred word sample

6) an Advanced Guiraud estimate

The first four measures are all alternatives to standard TTR and attempt to get around the effect of essay length. TTR of a fixed sample is in effect the number of word types in a fixed sample. The fixed sample length compensates for length effects but the drawback of this measure is that it does not use all the word types for many of the longer essays. There are also practical problems associated with essays that contain less than one hundred words. Guiraud Index (GI) is a variant of the TTR and is the number of word types divided by the square root of the number of word tokens as follows:

$$GI = {}^{v}\!\big/\!\sqrt{N}$$

where v is the number of word types and N is the number of word tokens. Guiraud Index is also affected by essay length but in a different way to TTR.

Yule's K was found in Chapter Five to be remarkably robust even for short essays. However, values for essays shorter than one hundred words may be unpredictable. The D estimate is also included for comparison but it should be noted that results in Chapter Five suggested that values of the D estimate for essays below 80 words may be unreliable.

Results of experiments in Chapter Five also suggested that the proportion of hapax legomena may be a good discriminator between learner and native essays. It is included here to see how it performs with differences between learners. In Chapter Five, the comparisons were made on essay samples of equal length. As this measure may also be affected by length, it was calculated for a hundred word sample. The final measure, Advanced Guiraud is considered as a measure of extrinsic lexical diversity. This was proposed by Daller, van Hout & Treffers-Daller (2003). Advanced Guiraud (A_G) is the number of advanced word types divided by the square root of the total number of tokens as follows:

$$A_G = {}^{v_a}\!\!\Big/\!{\sqrt{N}}$$

where v_a is the number of advanced word types and N is the total number of word tokens. As with Guiraud Index, Advanced Guiraud may be affected by essay length but to a lesser extent than TTR.

In Chapter Five, LFP and P_Lex were found to be problematic with some low level learners because the learners failed to produce enough advanced words. The students in this experiment are of a higher proficiency and so we might expect more advanced types to be used. Since P_Lex and LFP are difficult to use in an automated system, Advanced Guiraud may be an alternative extrinsic measure of lexical diversity to include in the analysis. However, in this analysis, an estimate of Advanced Guiraud is used. This involves an estimate of advanced word types calculated by subtracting frequent types from total types. A concern is that the measure may be contaminated by error variants of frequent types. For this study, frequent types were defined as word types from a list based on the first 1000 words of the JACET word list (Ishikawa et al., 2003).

6.2.3.2 Sampled features

The effect of essay length on TTR is well documented. As essay length increases, TTR tends to decrease. This makes it problematic for comparisons of essays of varying length. There is also a suggestion that the proportion of hapax legomena may decrease with essay length. Therefore, both of these features were calculated for a hundred word sample from each essay. However, six essays were shorter than one hundred words. These essays were of length 95, 95, 86, 79, 68 and 50 words. These six essays were treated differently for these two sampled measures. To make up the shortfall of words, additional words were sampled from each essay and added to the essay to bring the word count to a hundred. Therefore, the word tokens were increased to a hundred but the word types were not increased. This procedure was incorporated into the computer program. Only TTR and hapax legomena were calculated for sampled values. The notation TTR(100) and Hapax(100) is used to show these measures are based on a sample of 100 words. All other features were calculated using raw data.

Guiraud measures have been found to be affected by essay length. Malvern et al. (2004) found that Guiraud Index rises with increasing essay length for the first few hundred words but then decreases steadily with increasing essay length. Given the range of essay length in this set of essays, these Guiraud measures are likely to be rising against essay length. Features with a positive relationship with essay length are easier to incorporate into an analysis than features with a negative relationship. Therefore, the Guiraud measures are not controlled for essay length but the relationship with essay length needs to be taken into account in interpretation.

6.2.3.3 Calculation of quantity and content dimensions

For each essay, the measures of lexical diversity were calculated along with essay length. To ensure an equal weighting and easy comparison between essay length and the measure of lexical diversity, the values for each measure were standardized over the 34 essays by transforming into z scores. Z scores are calculated by subtracting the mean from each value and then dividing by the standard deviation. Standardization ensures that the mean of each measure over the essays is zero with a standard deviation of one.

6.2.3.4 Comparison of content variables

The two-dimensional model using z scores is graphed for each measure of lexical diversity. The use of z scores is quite conducive to easy interpretation. On a simple graph, four quadrants can be seen as shown in Figure 6.1. Quantity as represented by essay length is graphed on the x-axis while content represented by a measure of lexical diversity is graphed on the y-axis. Quadrants A and B are of particular interest. Quadrant A indicates a domain where essays are above average on both dimensions. If this quantity/content model is sound then highly rated essays should appear in this domain. Conversely, Quadrant B indicates a domain where essays are below average on both dimensions. We expect essays in this domain to be rated low. Quadrant C shows essays that are above average on one dimension, in this case essay length, but below average on the second dimension, lexical diversity. Quadrant D shows essays that are above average on the second dimension, lexical diversity, but below average on essay length. The quality of essays appearing in Quadrants C and D is likely to be more difficult to predict. In this graphical representation, essay points close to the origin will also be close to other quadrants and potentially unreliable. Therefore, we would hope that essays assessed high or low would be clearly defined in the respective quadrants not close to the origin or either axis.

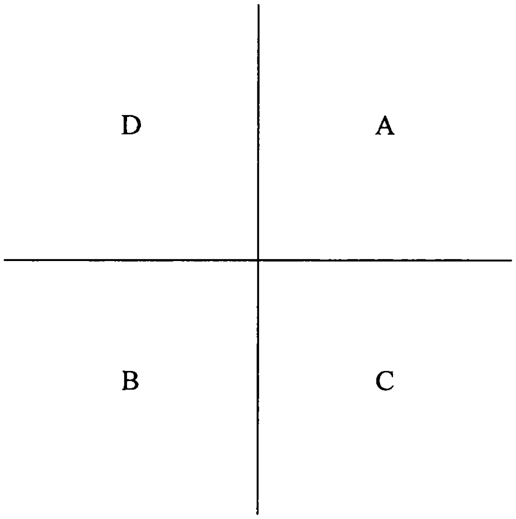


Figure 6.1: Four quadrants in a standardized variable graph

Yule's K is a special case of lexical diversity because its values have an inverse relationship with diversity. Lower values of Yule's K indicate higher lexical diversity and higher values indicate lower lexical diversity. To counter this and align Yule's K with the other variables, a negative value of the measure is graphed.

6.3 Results

6.3.1 Basic characteristics

The values of the various base measures before standardization are shown in Table 6.2.

Table 6.2: Essay characteristics of 34 learner essays

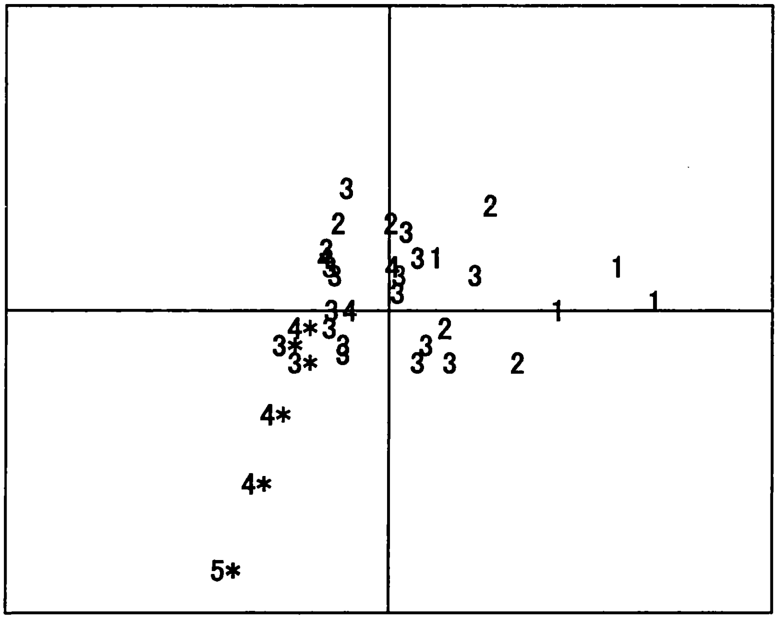
	length	TTR (100)	Guiraud	Yule	D	Hapax(100)	AG
Mean	146.3	64	7.07	121.7	75.85	46.59	5.20
SD	56.6	9	0.90	34.6	25.30	8.78	0.67
High	328	78	9.13	192	153.03	65	6.62
Low	50	34	4.95	58.7	38.25	25	3.54

All the measures showed a wide range of values. For example, essay length ranged from 50 words to 328 words.

6.3.2 Quantity/content graphs

6.3.2.1 Sample word types TTR(100)

Figure 6.2 shows a quantity/content graph using TTR(100) as the content dimension. Essays rated *good* are indicated by 1, *above average* by 2, *average* by 3, *below average* by 4 and *poor* by 5.



* essays with less than 100 words

Figure 6.2: Quantity v content using TTR(100) as content

In this graph, Quadrant A contains three of the four most highly rated essays but only two of the *above average* essays. It also contains four average essays but also one *below average* essay. However, these essays are not clearly defined. In Quadrant B, the one *poor* essay is well isolated as are two *below average* essays. However, this clear definition is deceptive. TTR(100) was based on a hundred word sample but these clearly defined essays all contained less than a hundred words. The topping up of tokens may have emphasized the dearth of types in these essays so the poor performance of these essays probably owes itself as much to the shortage of words as to the content dimension. Overall, the two-dimensional model seems inferior to a one dimensional model using only essay length which would clearly identify some *good* and *above* *average* essays at the top end of the essay rating spectrum.

6.3.2.2 Guiraud Index

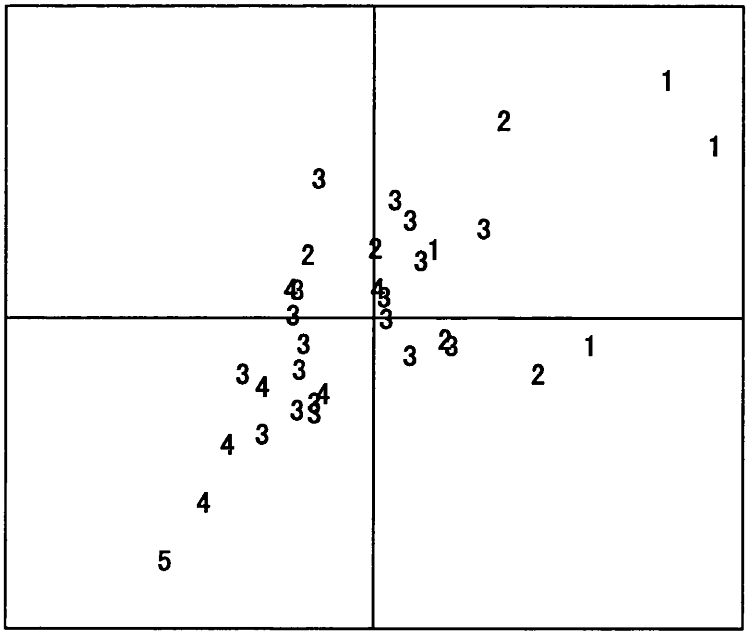


Figure 6.3: Quantity v content using Guiraud Index as content

Figure 6.3 shows Guiraud Index as the content dimension. This seems to perform better than TTR(100) with two *good* essays and one *above average* essay well defined in Quadrant A. Similarly, the *poor* essay and two *below average* essays are well-defined in Quadrant B. Unlike TTR(100), Guiraud Index isolates these lower end essays by itself. This suggests that a two-dimensional model based on Guiraud Index would be able to clearly isolate essays at the top and bottom of the rating spectrum. The model seems to indicate essays at the bottom end of the scale more clearly than essay length by itself although essay length may still be a superior discriminator at the top end.

6.3.2.3 Yule's K

Figure 6.4 shows Yule's K as the content element. As with TTR(100), essays are not very well defined at the top of the range but the measure seems better at the bottom end distinguishing the *poor* essay and some *below average* essays. However, it does not seem to be an obvious improvement over essay length by itself.

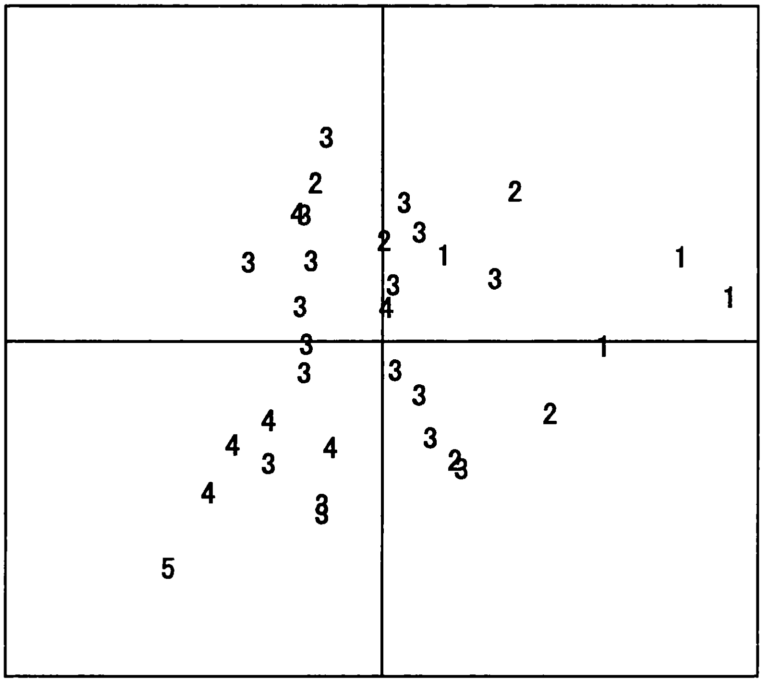
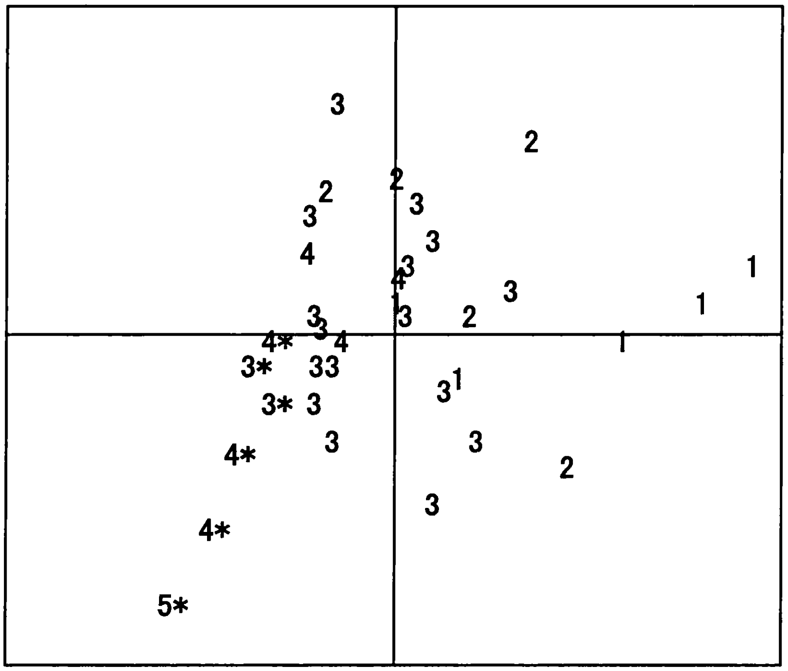


Figure 6.4: Quantity v content using Yule's K as content

6.3.2.4 Hapax(100)



* essays with less than 100 words

Figure 6.5: Quantity v content using Hapax(100) as content

Figure 6.5 shows Hapax(100), the proportion of hapax legomena in a 100 word sample, as the content element. As with TTR(100), values at the bottom end may be emphasized by the short length of the essays rather than the lexical diversity measure. At the top end, the performance is inferior to essay length by itself.

6.3.2.5 The D estimate

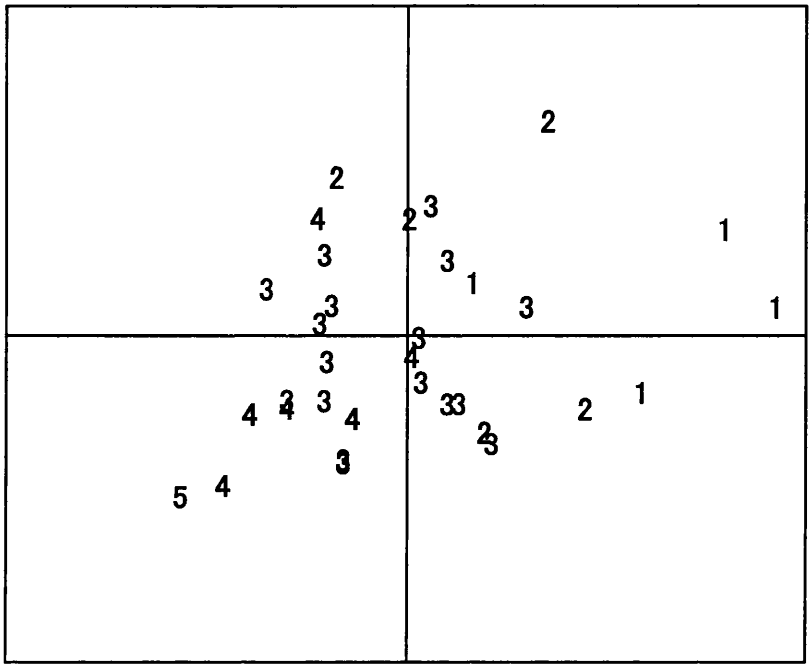


Figure 6.6: Quantity v content using the D estimate as content

Figure 6.6 shows that the D estimate isolates essays at the bottom end of the scale well with several *poor* and *below average* essays clearly defined. In Quadrant A, the D estimate does not define highly rated essays so clearly. D seems the most conservative estimator in that no major misclassifications occur. No *below average* or *poor* essays appear at all in Quadrant A and no *good* or *above average* essays appear in Quadrant B. However, it does not highlight essays at the top end of the scale as well as Guiraud Index.

6.3.2.6 Advanced Guiraud

Figure 6.7 shows Advanced Guiraud estimate as the content dimension. Advanced Guiraud seems to perform well at the extreme of both ends of the scale clearly identifying two *good* essays, one *poor* and one *below average* essay. However, other essays are less clearly identified.

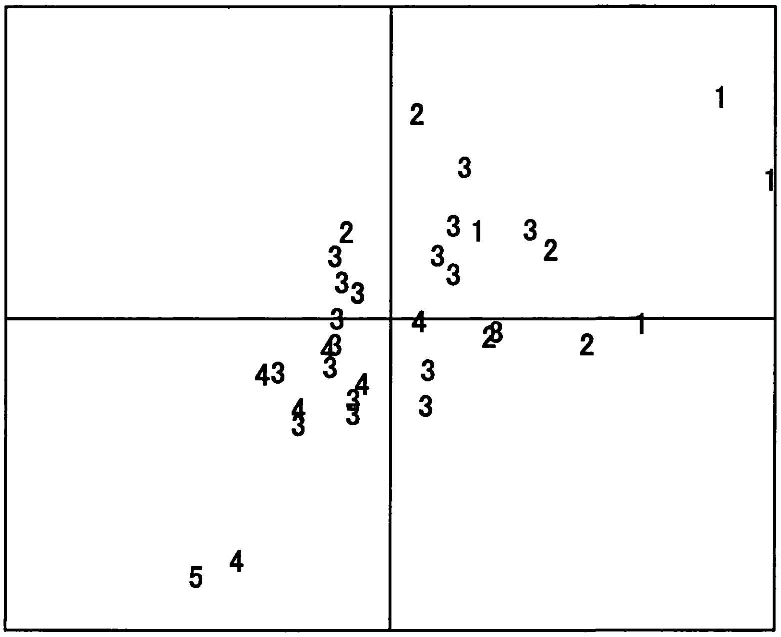


Figure 6.7: Quantity v content using an estimate of AG as content

The graphical evidence suggests that for these students on this task, essay length was a very strong predictor of overall essay ratings. Certainly, the content dimension of the model seems to take second place to the quantity dimension. This means that perhaps we need to rethink the equal weighting of the model to one that offers a greater contribution to quantity. Guiraud Index and the Advanced Guiraud estimate seem to perform better than the other lexical measures in terms of clearly defining a small number of high quality or low quality essays. However, the fact that these may both still depend on length may be accounting for their performance to some extent.

6.3.3 Correlation analysis

Table 6.3: Correlation of essay ratings, essay length and lexical diversity

	length	TTR(100)	Guiraud	-Yule	D	Hapax(100)	AG	ratings
length		0.359	0.682**	0.254	0.202	0.230	0.646**	0.786**
TTR(100)			0.811**	0.842**	0.802**	0.935**	0.696**	0.468*
Guiraud				0.781**	0.783**	0.783**	0.846**	0.611**
-Yule					0.931**	0.823**	0.636**	0.382
D						0.851**	0.595**	0.337
Hapax(100)							0.579**	0.341
AG								0.660**

**significant at the .001 level *significant at the .01 level

Looking at graphs can give us a general idea. Correlation analysis may help give some more objective evidence. Table 6.3 shows correlation values between the features and essay ratings. The correlations with essay ratings underline the impact of essay length which has the highest correlation of r = .786. The next two highest are Advanced Guiraud and Guiraud Index with r = .660 and r = .611 respectively. These very high correlations may reflect an indirect influence of essay length. The other measures are broadly similar with sample types performing the best, r = 0.468. This contradicts the graph results somewhat but alerts us to an important aspect of many assessment situations which is decision accuracy. The identification of essays at the top of the range and the bottom of the range may be more important than essays in the middle because there is a chance with these essays to make a larger misclassification error.

The strong correlations of Guiraud Index and Advanced Guiraud with essay length, r = .682 and r = .646 respectively, support the idea that they are still considerably affected by length of essay. Generally high correlations between the different measures of lexical diversity suggest that they are all measuring broadly the same thing. Very high correlations between pairs of these measures reflect similarities in properties. The Guiraud Index and Advanced Guiraud, r = .846, are both counts divided by the same denominator and also both depend on essay length. The correlation between Yule's K and the D estimate is particularly high, r = .931, perhaps reflecting the fact that they both represent parameters of word distributions. The correlation between TTR(100) and Hapax(100) is high, r = .935, probably because they were both measured on the same sample of words.

6.3.4 Partial correlation analysis

Essay length seems to be the best predictor of essay ratings with this set of essays. We are interested in the measure of lexical diversity that in conjunction with essay length can best improve on this prediction. One way to think of this is which measure can add something to a one-dimensional essay length model. Table 6.4 shows the partial correlations of the measures of lexical diversity with essay ratings when essay length is controlled for. These partial correlations tell us how much each measure correlates with essay ratings if the influence of the length effect is taken out. The calculations were done with PASW software (PASW, 2009).

Table 6.4: Partial correlations with essay ratings controlled for essay length

	TTR(100)	Guiraud	-Yule	D	Hapax(100)	AG
r	.323	.166	.304	.294	.267	.323

The interesting values are for Guiraud Index and Advanced Guiraud. Although both of these have relatively strong correlations with essay ratings, they are both considerably weakened when controlled for essay length. This again underlines the effect of essay length on these measures. Guiraud Index has by far the lowest partial correlation with essay ratings. The other measures are all about the same but TTR(100) and Advanced Guiraud are the strongest. However, excluding Guiraud Index, the measures vary in a short range from r = .267 to r = .323 so it is probably unwise to make any conclusions on which is best since the values are quite similar.

6.3.5 Multiple regression analysis

A multiple regression analysis was run to see how well each measure of lexical diversity performed alongside essay length in predicting essay ratings. Each measure was used with essay length in a two independent variable model with essay ratings as the dependent variable. Table 6.5 shows the r values and r^2 values for each measure of lexical diversity in a two variable model with essay length. The calculations were done with PASW software (PASW, 2009).

Table 6.5: R and r^2 values for two variable model with essay length

	TTR(100)	Guiraud	-Yule	D	Hapax(100)	AG	length only
r	.811	.792	.808	.806	.803	.811	.786
r^2	.657	.628	.653	.650	.644	.657	.617

The regression analysis highlights that essay length alone accounts for 61.7% of the variance. Another dimension of lexical diversity adds only a slight improvement to the model with TTR(100) and Advanced Guiraud performing best and accounting for a further 4% of the variance.

6.3.6 Results conclusion

The results suggest that lexical diversity together with essay length can more accurately predict essay ratings than either feature alone with this set of essays. However, essay length is a very strong predictor as a single dimension. Accordingly, in a two-dimensional model, essay length seems to be the dominant dimension with a greater weighting than lexical diversity.

It is less clear which measure of lexical diversity is the best to use in the two-dimensional model. The results suggest that TTR(100), Yule's K, Hapax(100) and Advanced Guiraud perform similarly well in correlation and regression analyses. They all perform as well as the standard measure of lexical diversity, the D estimate. Guiraud Index seems to perform less well. The graphical evidence suggests that Advanced Guiraud and Guiraud Index are good at clearly identifying a few essays at the top and bottom end of the scale. TTR(100), Hapax(100), and Yule's K seem less able to distinguish essays clearly. However, this may be partly due to the fact that the graph model gives the two dimensions an equal weight. The regression analysis suggests that essay length deserves a greater weight. Because the two Guiraud measures are affected by essay length, their inclusion as the content dimension in effect boosts the essay length weighting to a value closer to that estimated by the regression analysis.

6.4 Discussion

6.4.1 Overview

The results of this study while being very interesting are still inconclusive. However, there are some observations to be made. One is that with this set of essays, essay length proved to be overwhelmingly the dominant predictor of essay ratings. Lexical diversity had a relatively minor role to play. However, there are likely to be situations where the role of essay length is unlikely to be so great. Research suggests that higher level learners who we expect to write more are less likely to be easily discriminated by essay length. Also, non-timed tasks may show less dependence on essay length and more on lexical diversity.

6.4.2 Advanced Guiraud

Despite having reservations about the accuracy of an automated Advanced Guiraud estimate, it seemed to perform relatively well as a complementary measure to essay length. It was able to clearly identify two highly rated essays and low rated essays. It is worth considering how the four clearly defined essays could be interpreted. The two at the top of the scale have a high quantity of words and also contain a higher ratio of advanced word types than other essays. The two essays identified at the bottom of the scale contain few words and a low ratio of advanced word types compared to other essays.

One reservation about Advanced Guiraud concerned its relationship with essay length. Even though Advanced Guiraud had a significant correlation with essay length, it also had the highest partial correlation with essay ratings when controlled for essay length. This suggests that it is measuring another dimension of vocabulary use. Another reservation about this estimate of Advanced Guiraud centered on whether spelling errors conflated with advanced types would create a major effect. The clear identification of low rated essays suggests this may not have been a major problem. However, this weakness needs to be considered if used with other tasks involving other learners. It may be that in this case, the number of errors was relatively small compared with advanced types. In a situation where there are more spelling errors and/or fewer advanced types, this kind of estimate could be seriously compromised. A superior method of estimating advanced types may still be necessary.

6.4.3 Classifying essays

This experiment has looked at the relationship between essay length and lexical diversity but it does not actually give us a mechanism to use the information to split the essays into groups. A split into three groups could be done by isolating essays at the top end of the scale to be part of an above average group and at the bottom end of the scale to be part of a below average group. This would leave essays in the middle range to form an average group. For example, because z scores have an additive property, z scores weighted for the contribution by each dimension to total variance could be used. We could then add the weighted z scores for quantity (essay length) and content (lexical diversity) and rank the scores. A ranking was calculated using the z scores on the two dimensions weighted for contribution of total variance with essay length as the quantity dimension and Advanced Guiraud as the content dimension. These weighted z scores for each dimension were simply added together. Essays were arranged in descending weighted score order from left to right with the number representing the essay rating as follows:

1112231332332343323343333343433445

The three essays with the highest weighted z scores were three essays that were assessed as being *good* essays. The other essay rated *good* was ranked seventh best according to weighted z scores. At the bottom, the lowest ranked essay was rated *poor* and the next two lowest ranked essays were rated *below average.* The ranking seems to progress broadly in line with the ratings but there are some anomalies. An *average* rated essay ranks sixth best, a *below average* essay ranks fifteenth best while an *above average* essay ranks only eighteenth best.

6.4.4 Linearity

A comparison of graph data and correlation ratings suggests that linear methods such as correlation and regression may not always be the best for recognizing the power of some features. Although both sample types and Advanced Guiraud showed similar correlations, Advanced Guiraud seemed better at isolating essays at the top and bottom of the scale. One problem of linearity is that sometimes more of a feature is not necessarily better. Often simple presence or absence is important. Another aspect of the limits of linearity is highlighted in a study by Jarvis, Grant, Bikowski & Ferris (2003). In a study involving clustering of essays according to various essay features, they found that highly rated essays had a variety of profiles.

Although many features showed an overall positive correlation with essay ratings, an individual feature often had relations with other features that would override this general relationship in certain situations. They noted two such scenarios, complementarity and compensation. Complementarity refers to a situation where the high presence of one feature may often be accompanied by a low presence of another. For example, in one cluster of highly rated essays, the use of many nouns was accompanied by a relatively low occurrence of pronouns. In another cluster, the opposite pattern was noted. This was despite both features having an overall positive relationship with essay ratings. Compensation refers to where learners who are not good at producing one feature make up for it in another way. For example, they found some learners with low values for clausal embedding, might make up for this in other ways, for example, by writing longer essays with higher lexical diversity values. This type of pattern may come as no surprise to teachers involved with student writing, yet standard statistical analyses have no way of recognizing this kind of relationship.

6.5 Conclusion

This chapter has focused on a simple two-dimensional quantity/content model for essay evaluation. The dimension of quantity was represented by essay length and content by various measures of lexical diversity. The two features together appeared to account for quality better than either feature by itself. For this set of essays, essay length was found to be the dominant dimension accounting for over 60% of the variance in essay ratings. The addition of a lexical diversity measure only accounted for a further 4% of variance. However, the relative dominance of the dimensions is likely to vary by task and learners. For example, in the experiment in Chapter 5.1, lexical diversity by itself accounted for quality better than essay length.

The number of word types in a fixed sample (TTR(100)), Advanced Guiraud, Yule's K, the D estimate and the number of hapax legomena in a fixed sample (Hapax(100)) all accounted for a similar amount of the variance. Advanced Guiraud and Guiraud Index both showed a strong correlation with essay length which reflects their dependence on essay length. However, when essay length was controlled, Guiraud Index lost most of its association with quality while Advanced Guiraud remained strongly associated to essay quality. There was also a very strong association between the D estimate and Yule's K which was much higher than between other pairs of lexical diversity measures which suggests that these two features may be measuring the same construct.

Graph evidence showed slightly different patterns to the correlation and regression evidence.

Whereas the correlation evidence suggested that the measures of lexical diversity had a similar correlation with quality when complementing essay length, graphs showed that some features more clearly identified extreme high quality and low quality essays. In particular, Advanced Guiraud seemed to be the most effective complement to essay length for clearly identifying a small number of highly rated essays and low rated essays.

In the next chapter, this two-dimensional model is extended to more complex models that can incorporate more than two features.

Chapter Seven: Automatic scoring algorithms

7.1 Introduction

In the previous chapter, we considered a simple two-dimensional model for assessing essays. In this chapter, we look at two complex automatic algorithms that can incorporate more than two features. The first uses cluster analysis of lexical features of essays to help classify essays as higher quality or lower quality. The second uses a Bayesian analysis of the actual words produced in the essays. In this algorithm, a number of basic lexical features are used to guide the classification in place of a training set. Both algorithms used the same essay data for comparative purposes.

7.2 A clustering algorithm

7.2.1 Introduction

In this experiment, a clustering algorithm is used to classify a set of essays into two groups according to quality. This algorithm uses a small set of lexical features to cluster essays according to similarity. This small set of features is refined from a larger set through a principal component analysis (PCA). The advantage of features selected by PCA is they account for as much of the variance in the essays as possible, and also they are likely to be independent or close to independent of each other. Two initial clustering points are chosen: one that indicates a likely high quality space and another that indicates a likely low quality space.

7.2.1.1 Aims of the experiment

The main aim of this experiment is to compare the performance of a clustering algorithm with the results of human quality assessments. A secondary aim is to evaluate lexical features for their usefulness in automatic assessment.

7.2.2 Methodology

7.2.2.1 Participants and tasks

One hundred essays written by first year Japanese university students were collected. The written task was based on Cartoon Task 1 in Appendix 3.1. Students were given 25 minutes to complete the task (see Appendix 3.3 for sample essays).

7.2.2.2 Holistic assessments

These one hundred essays were classified by two native judges into two groups of fifty each according to quality. One group was a higher quality group and the other a lower quality group.

This type of assessment system was decided on for a number of reasons. Firstly, it mirrored the kind of assessment situation that the two judges often experience in their professional lives. Students entering university in Japan are typically streamed according to ability into several English classes by means of a placement test. This usually involves splitting a group of students into a number of classes of a given size. Secondly, the judges are used to making these decisions by relative proficiency rather than by absolute proficiency as measured by some outside criteria. In addition, using outside criteria has the disadvantage of requiring considerable training of raters.

7.2.2.3 Analysis

The analysis involved a number of stages. This analysis was carried out using the L-analyzer and L-cluster programs (see Appendix 1.1). The main stages of the analysis were as follows:

1) selecting features for the analysis

2) calculating these features automatically

3) identifying a set of principal components that were independent and accounted for most of the variance in the data

4) finding features in the original set which approximated these principal components

5) clustering the essay data using this reduced set of features to find two groups of high quality and low quality essays

Each of these stages is now described in more detail.

7.2.2.3.1 Selection of features

The first stage of the experiment was to identify a set of lexical features to input into the analysis. This set consisted of simple features that have been associated with essay quality in previous research. The initial input features were as follows:

1) Essay length in words
2) TTR(100), the number of word types in a 100 word sample
3) Hapax(100), the number of hapax legomena in a 100 word sample
4) An estimate of Advanced Guiraud
5) A distance measure of occurrences of *the* and *a* to typical occurrence in English
6) An estimate of mean sentence length (the number of words divided by the number of sentence ending punctuation)
7) An estimate of mean clause length estimate (the number of words divided by the number of commas and sentence ending punctuation)
8) Mean word length

9) Entropy
10) Yule's K
11) An estimate of lexical error

Features were chosen that have been shown to be associated with quality assessments. Essay length has been shown by many including Larsen-Freeman & Strom (1977) to be positively related to written quality. McNeill (2006) found length to be the feature most correlated with holistic assessments in essays written by Japanese college learners.

The number of word types in a fixed sample is in effect the TTR of a fixed sample. Raw TTR has been shown to be affected by essay length, but TTR of a fixed sample is controlled for this effect. In addition, in earlier experiments in Chapter Three and Four, the number of word types was the most reliable measure over similar tasks written by the same learners. In Chapter Six, a sampled word type measure had the strongest correlation with essay ratings alongside Advanced Guiraud from a set of lexical diversity measures.

The number of hapax legomena in a fixed sample of words was shown in Chapter Five to discriminate well between learner essays and native samples. It was also found in Chapter Six to correlate with essay ratings.

In Chapter Six, Advanced Guiraud was found to be strongly correlated with essay ratings even when essay length was controlled. Advanced Guiraud is a measure of extrinsic lexical diversity and works on the assumption that less frequent words are more likely to be used by more proficient learners. Since LFP and P_Lex are difficult to embed into a computer program, it may be a useful alternative for automatic assessment.

A distance measure of the relative occurrence of *the* and *a* was shown in Chapter Five to have promise in discriminating between native and learner essays. It is not clear how effective it is at discriminating between learner essays of different quality. However, Evola et al. (1980) found that the proportion of articles in essays was significantly correlated with quality assessments.

Mean sentence length and mean clause length may help identify essays written by higher proficiency learners if learners are expected to produce longer clauses and sentences as proficiency increases. Studies in Chapter Five found some differences in mean sentence length relating to proficiency.

Zipf (1932) showed that frequent words tend to be shorter in length than more infrequent words. Accordingly, essays of more proficient learners might be expected to exhibit a higher mean word length. Mean word length was also shown in Chapter Five to vary according to proficiency of learners. Also, native samples showed a higher mean word length than learner essays.

Although statistics like entropy and Yule's K may be less suited to short essays, results in Chapter Five still showed significant differences between learner essays of differing proficiencies. Also, in Chapter Six, Yule's K was found to correlate with essay ratings as strongly as other measures of lexical diversity.

Lexical error has been shown in various studies to be related to judgments of L2 essay quality, for example, Engber (1995).

It is also worth mentioning why some measures have not been included. Although the results of experiments in Chapter Three and Four suggest that distinctiveness and uniqueness have correlations with essay quality in some situations, they were not included here because of concerns over their linearity. Linearity is an assumption of the PCA process.

7.2.2.3.2 Calculation of features

Most of these features are easily calculated automatically. However, some are less easy to calculate. In particular, lexical error is a problematic concept to include in an automatic application. For instance, there are various difficulties in judging error such as different types of error and different degrees of seriousness of error. This makes error judgment difficult even for humans. Error in this analysis was predicted by a small subset of error. To calculate this estimate, a minimal amount of human intervention was required. However, if effective, this intervention should be required less in future due to a simple 'learning process' built into the algorithm. This error estimation process involved checking all words in the essay against a list based on the first 1000 words of the JACET word list and against a list of proper nouns. Any words found outside these lists were flagged for human checking. Any judged to be non-words were tallied for an estimate of lexical error. Any words cleared at this final stage as being legitimate words could be added to a user dictionary for the program to learn more acceptable words. It might also be possible to include common errors in a common error dictionary to also speed processing and lessen human intervention in the future.

Advanced Guiraud involves tallying advanced types. The JACET 1000 word list was used as a basis for calculating this measure. Any words not in this list and also not appearing in a list of errors and proper nouns were considered advanced types.

Both mean clause length and mean sentence length cannot be calculated accurately in a simple program so an estimator was used for each. The total number of words in an essay divided by the sum of total sentence ending punctuation counts was used for mean sentence length. Similarly, the total number of commas was added to the denominator for the previous calculation to give an estimate for mean clause length.

7.2.2.3.3 Principal component analysis

The next step of the analysis was to use a principal component analysis (PCA) to identify a smaller set of features to use in the cluster analysis. Principal component analysis is a statistical technique which realigns multivariate data to provide a new set of variables which are ordered in terms of variance and are independent of each other. An assumption of PCA is that the input variables are linear in their relationship with the dependent variable. In this case, the assumption means that the input features have a linear relationship with essay quality.

Consideration of this assumption affected the form of some of the selected features. For example, a previous study showed that the co-occurrence of articles *the* and *a* may be a reliable discriminator between essays written by native speakers and those written by L2 learners. According to the British National Corpus, the ratio of occurrence of *the* to *a* in general English is about 2.9 (Leech, Rayson & Wilson, 2001). A simple ratio of occurrences of these two words is unlikely to be linear in its relationship to written quality. A value close to 2.9 may be desirable with values distant less desirable. Therefore, for this experiment a measure of distance from a typical occurrence in English was calculated. A low value for this feature would indicate similarity to norms of English (a positive sign) whereas distance from this norm would be a more negative sign. A feature in this form is linear in its relationship to essay quality albeit inversely.

The set of features was calculated for each essay and the standardized z scores were subject to a PCA. These z scores were used to prevent features with large variances dominating the analysis. The first six principal components (PCs) produced by this analysis accounted for over 91% of the total variance. A scree chart showing the variance accounted for by each PC can be seen in

Figure 7.1.

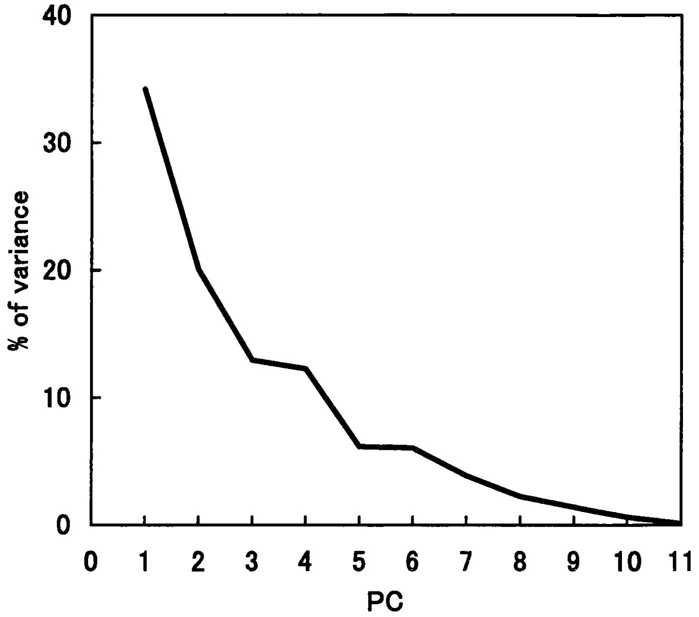


Figure 7.1: Scree chart of PCs

The first principal component (PC1) accounts for the largest share of the variance, in this case about 34%. The second PC accounts for about 20% of the variance, PC3 for about 13%, PC4 for about 12%, PC5 and PC6 for about 6% each. The remaining seven PCs account for a very small proportion of the variance from about 4% for PC7 down to a minute percentage for PC11. The variance accounted for by each PC and the cumulative variance is shown in Table 7.1. The first two PCs together accounted for 54.3% of the variance in the data, the first three PCs for 67.2% of the variance and the first six PCs for 91.7% of the variance in the data.

Table 7.1: Percentage of variance accounted for by the PCs

PC	1	2	3	4	5	6	7	8	9	10	11
variance %	34.2	20.1	12.9	12.3	6.2	6.1	3.9	2.2	1.4	0.6	0.1
cumulative variance %	34.2	54.3	67.2	79.5	85.6	91.7	95.6	97.8	99.2	99.9	100

7.2.2.3.4 Comparison of features with principal components

Because PCA produces independent principal components (PCs), each identified PC should reflect a different strand of variance within the data. It is possible, but by no means certain, that these could correspond to different aspects of knowledge that manifest in essay quality. The independence is a useful property. Many lexical features are often suspected of being predictors of essay quality simply because of their own relationship to essay length. Also, many lexical diversity statistics measure broadly the same construct but do seem to have differences. Because principal components are independent, we do not need to worry about using features together that may be dependent on each other. It also means that it is not necessary to control features for essay length for PCA. However, controlling the features for length in this case is useful because we match the most influential PCs back to the original features in the next stage of the analysis.

If the original z score of features are mapped against each PC, original features that correspond closely to each PC can be found. This will give us a set of independent features to use to predict essay quality. This matching of original features to principal components is a technique recommended by Jolliffe in his handbook on principal component analysis (Jolliffe, 2002). For example, the first PC which accounted for over 34% of the variance was highly correlated with the feature essay length. This matching process also eliminates features measuring similar properties. For example, PC3 correlated closely with both mean sentence length and mean clause length. Mean sentence length correlated better with PC3 than mean clause length so mean sentence length was selected and mean clause length discarded. The graphs of the first six PCs with matched features in z score form can be seen in Appendix 7.1. The six selected features for each of these six PCs were as follows:

PC1 Essay length
PC2 TTR(100), the number of word types in a 100 word sample
PC3 An estimate of mean sentence length
PC4 An estimate of lexical error
PC5 Mean word length
PC6 Hapax(100), the number of hapax legomena in a 100 word sample

7.2.2.3.5 Clustering

Using this set of six features, a clustering was carried out. Clustering was initiated by using high values and low values of features. These high values are likely to be indicative of high quality essays and low values are likely to be indicative of lower quality essays. Feature 4 was opposite in its orientation. A low value of lexical error is likely to be indicative of a high quality essay while a high value is likely to be indicative of a low quality essay. Again the standardized z scores for features were used for clustering. Cluster locations were estimated for a high quality cluster and a low quality cluster. The high quality cluster was built around a point in six-dimensional space based on the following parameters:

- Mean essay length + 1 standard deviation

- Mean TTR(100) + 1 standard deviation
- Mean sentence length +1 standard deviation
- Mean lexical error -1 standard deviation
- Mean word length + 1 standard deviation
- Mean Hapax(100) + 1 standard deviation

In a similar way, the initial cluster point for the lower cluster was set as:

- Mean essay length - 1 standard deviation
- Mean TTR(100) - 1 standard deviation
- Mean sentence length -1 standard deviation
- Mean lexical error +1 standard deviation
- Mean word length - 1 standard deviation
- Mean Hapax(100) - 1 standard deviation

Table 7.2: Features and weightings used in the clustering

feature	essay length	TTR(100)	mean sentence length	lexical error	mean word length	Hapax(100)
weighting	1	20/34	13/34	12/34	6/34	6/34

Each feature was weighted in accordance with the percentage of variance accounted for by each corresponding PC. The six features received the weighting shown in Table 7.2. Essays were progressively added to each cluster according to relative distance to the midpoint of each existing cluster. At first, each cluster is represented by the initial cluster point. For each essay, the Euclidean distance (see Appendix 7.2) in six-dimensional space from the low cluster point was divided by the distance to the high cluster point. The essay with the lowest value is relatively close to the low cluster point and the essay with the highest value is relatively close to the high cluster point. These two essays were added to the respective closest cluster point to form a cluster of two points, an essay point and the initial cluster point. The midpoints of each cluster were calculated. The midpoint of a cluster was calculated as the average in six dimensional space of each essay or point in the cluster. The distance from each remaining essay to each cluster was then calculated again and the procedure was repeated until each group contained fifty essays. After many iterations, two clusters of fifty essays each were formed.

7.2.3 Results

The results of the experiment were compared to ratings of two native speaker judges and the decision agreements and Kappa statistics (see Appendix 7.3) for agreement adjusted for chance agreement are shown in Table 7.3.

Table 7.3: Decision agreement (DA) and Kappa for clustering algorithm

	Rater 1		Rater 2	
	DA	Kappa	DA	Kappa
Clustering	.78	.56	.76	.52
Rater 1	-	-	.72	.44

The results of this clustering algorithm agreed with Rater 1 in 78 cases out of a 100 and with Rater 2 in 76 cases out of a 100. The Kappa statistic corrected for chance agreement is r = 0.56 for the clustering algorithm and Rater 1 and 0.52 for the clustering algorithm and Rater 2. The two raters agreed with each other in 72 cases out of a 100 which corresponds to a Kappa reliability of 0.44. Therefore the clustering algorithm agreed with each human rater more than the human raters agreed with each other.

According to this analysis, the lexical features that seem to correspond with the most influential PCs are essay length, TTR(100), mean sentence length, error estimate, mean word length and Hapax(100). This order reflects the variance explained by the corresponding matched PC.

7.2.4 Discussion

7.2.4.1 Reliability

The clustering algorithm in this experiment has achieved a reliability with human ratings which more than matches the reliability between the two raters themselves. Decision agreement with raters is 78% with Rater 1 and 76% with Rater 2. However, decision agreements, as they stand, may not be accurate indicators of inter-rater reliability. Consider a case of 100 random allocations into two groups of fifty. If this was done twice, we would expect 50% agreement by chance. The Kappa statistic can be used to translate rater agreements to inter-rater reliabilities by eliminating this chance agreement element. After correcting for chance the reliability rating is quite reduced at r = 0.56 with Rater 1 and r = 0.52 with Rater 2 but still superior to reliability between human raters, r = 0.44.

One of the reasons for this low value is the requirement that fifty essays be classified into each group. Although this is a realistic situation in practice, it may cause problems in classification. It is unlikely that the hundred essays naturally fall into two sets of fifty essays according to proficiency. This may be problematic for raters. Probably, very high quality and very low quality essays may be relatively easy to allocate but essays that are borderline are likely to prove more difficult.

A good deal of this lost reliability may be due to the two raters dealing with borderline essays in different ways. Therefore, to judge the performance of the automatic assessment it might be instructive to look at the essays the human raters agreed on and see how these essays were classified by the clustering algorithm. This is shown in Table 7.4.

Table 7.4: Automatic treatment of essays agreed on by raters

		Raters	
		High	Low
Clustering	High	32	5
	Low	4	31

The two human raters agreed on 36 high quality essays and 36 low quality essays. Of these 36 high quality essays, 32 were also classified as high quality by the clustering algorithm. Conversely, this means that four essays that both raters identified as high quality were rated low quality by the automatic method. Out of 36 essays rated low by both human raters, 31 were also rated low by the algorithm but five were rated high. These cases may be of interest and may warrant further attention. It could be that these essays are being mis-rated by the clustering algorithm. Some caution is necessary, however. It could also be that these are, in fact, borderline essays that both raters just happened to rate the same way. If we test the inter-rater reliability of this clustering algorithm on this smaller group of pooled ratings, we get an agreement of 63 cases out of 72 (87.5%) or a Kappa statistic of 0.75.

7.2.4.2 Clustering

Looking at the clustering itself, there may be ways that results could be improved. There are various ways clustering can be performed. In particular, there are different techniques that can be used to form the cluster. Different distance measures can also be used. The initial number of input features could be increased. Compared to other studies, the number in this experiment was quite small. Jarvis et al. (2003) used 21 features in a cluster analysis. Some other automatic systems also used a lot of input features. E-rater01 used fifty initial features from which eight to ten were selected.

One drawback of the automatic assessment systems reviewed earlier is that they require a training set of data external to the set of essays under consideration. This clustering algorithm avoided the need for a training set by setting two points in space as initial points for the high cluster and low cluster. These points were selected as a point one standard deviation above or below the mean on each feature. The standard deviation is a measure of dispersion of points around the mean. If a feature is normally distributed, two-thirds of the values lie within one standard deviation of the mean leaving the other third of the points lying more than one standard deviation away from the mean. The mean plus and minus a standard deviation seemed good points to set as initial points for clusters to coalesce around. These initial settings could be varied to improve performance.

We might also consider whether initial cluster points are necessary at all. Clustering is often used for exploratory data analysis with no guidance. Jarvis et al. (2003) used unsupervised clustering on two sets of learner essays. They found that high quality essays formed a number of clusters which suggested that high quality essays had multiple profiles. The nature of these multiple profiles varied over the two different essay sets. By setting clustering points in this study, richer data that may emerge from a freer analysis may be lost.

7.2.5 Conclusion

In this experiment, a clustering algorithm has been shown to be a possible alternative to human assessment for the rating of L2 learner essays. Using a large set of lexical features which have shown relationships with essay quality in previous studies, a principal component analysis found a smaller set of independent variables that accounted for most of the variance in the data. These variables were then matched back to the original features and these matched features were used to cluster the essays into a high quality and a low quality group. In this experiment, the ratings produced by the clustering algorithm corresponded with assessments of human raters better than the human raters corresponded with each other. However, there appear to be a small number of essays dealt with differently by the clustering algorithm and both raters which may indicate a systematic weakness in the algorithm. In the next experiment, an alternative automatic algorithm using the same essays is considered.

7.3 A Bayesian algorithm

7.3.1 Introduction

In this experiment, an algorithm implements a Bayesian classifier to categorize the same one hundred essays as in the previous experiment. Fifty essays were identified as being high quality leaving the remaining fifty essays to be classified as relatively low quality. The Bayesian classifier was guided by training essays selected automatically from the set of essays. The selection was done by recognizing features that are highly likely to indicate high quality essays.

The basic premise of a Bayesian classifier is that classification of an item (in this case, an essay) can be guided by comparing the occurrence of features of an item (in this case, words in the essay) to the occurrence of features in two representative groups of items (for example, samples of high quality and samples of low quality essays). The item can be classified as a member of the group to which there is most similarity in occurrence of features.

Various essay features could be used in the analysis. Lexical statistical features such as those used in the previous experiment could be used but in this experiment the semantic content of the essays was analyzed. Occurrence of particular words in each essay was compared to occurrence in a set of predicted high quality essays.

To do a Bayesian analysis, a number of essays are needed as training samples. In this experiment, these training essays were selected automatically from within the set of essays. The method of selecting these training essays was based on observations in other studies.

Previous studies in Chapter Five and Chapter Six suggested that some simple lexical features such as essay length or lexical diversity could be effective at identifying small numbers of very good essays or very poor essays. Essay length may be a good indicator of high quality or low quality essays particularly in a timed essay format. Very long essays are likely to be relatively high quality and very short ones are likely to be relatively low quality. Lexical diversity measures were also shown in Chapter Five and Six to help isolate small numbers of good and poor quality essays. However, both essay length and lexical diversity are often less effective at distinguishing between more borderline essays.

A similar process often happens when humans rate essays. Some essays are easier to rate than others. Typically ones that are particularly high quality or low quality may often be easy to identify while the great majority that are closer to average may take more consideration.

In this study, a training sample of essays was selected from within the set of essays by using combinations of features which were highly likely to indicate good quality essays. The features chosen were essay length, lexical diversity, the number of hapax legomena and the proportion of error.

Once the training sample was identified, the occurrence of words in individual essays was tallied. These occurrences were then compared with occurrences in two sets of essays. One set was the training sample and the other set included all the essays outside the training sample excluding the particular essay under consideration. If the occurrence of words in an essay resembled the essays in the training samples, then it was deemed similar and allocated to the high quality group. If not, it is allocated to the low quality group.

At this point, it is worth commenting why this experiment used only a high quality sample whereas the previous experiment used a core high cluster and a core low cluster. Because this experiment involved analysis of each word in an essay, shorter essays included less information than longer essays. These shorter essays themselves are likely to be poor quality but their lack of information caused by their short length may affect the reliability of predicting other similar essays. In the previous experiment, any shortage of words only affected one feature, essay length, because the other features were independent of essay length.

Using Bayesian classifiers for essay grading is not a new idea. Larkey (1998) used a similar method to classify five large L1 data sets from large testing organisations. However, the training sets used in the study were very large, ranging from 233 training essays to rate 50 essays to 586 training essays to rate 80 essays. Larger groups of essays were also rated. One training set of 383 essays was used to rate a set of 225 essays. What this experiment sets out to investigate is whether relatively tiny training sets can give reliable results.

7.3.1.1 Aims of the experiment

The main aims of this experiment are to:

1) compare the performance of a Bayesian algorithm with the results of human rater assessments.

2) check the reliability of a training set of essays identified automatically from the set of essays.

7.3.2 Methodology

7.3.2.1 Participants and tasks

The same one hundred essays were used as in the previous experiment with the same holistic assessments (see Appendix 3.1 for the task).

7.3.2.2 Analysis

The analysis had two major parts:

1) the selection of a high quality training sample

2) Bayesian classification of remaining essays into high quality and low quality groups

Both of these are now described in more detail. The programs L-analyzer and L-Bayes (see Appendix 1.1) were used in the analysis.

7.3.2.2.1 Selecting the training set

To select the training set, four features were used, essay length, lexical diversity, the number of hapax legomena and an estimate of error. Since length of essay can affect the number of hapax legomena and lexical diversity, a fixed sample was used for each essay and hapax legomena and lexical diversity were calculated for that sample. Lexical diversity was calculated as the number of word types in a fixed sample of 100 words, TTR(100). The number of hapax legomena was also tallied for a 100 word sample, Hapax(100). An error estimate was calculated in the same way as in the previous experiment and an error proportion in relation to essay length calculated.

The target was to find a set of about 15-20 essays to form the training set. A set of conditions for predictors of high quality essays was constructed. These conditions were considered in descending order of strictness until a sufficient number of essays were identified. The conditions were as follows:

1) In the top quartile of each of three features, essay length, TTR(100) and Hapax(100) while also not including a high proportion of error (A high proportion of error is an error proportion in the top quartile)
2) In the top quartile of essay length and TTR(100) while being above average for Hapax(100) and not including a high proportion of error
3) In the top quartile of essay length and Hapax(100) while being above average for TTR(100) and not including a high proportion of error
4) In the top quartile of essay length while being above average for both TTR(100) and Hapax(100) and not including a high proportion of error
5) In the top quartile for both TTR(100) and Hapax(100) while also being above average for essay length and not including a high proportion of error
6) In the top quartile of TTR(100) while being above average for both essay length and Hapax(100) and not including a high proportion of error
7) In the top quartile of Hapax(100) while being above average for both TTR(100) and essay length and not including a high proportion of error
8) Above average for all three features of essay length, TTR(100) and Hapax(100) while exhibiting zero error

These eight conditions identified sixteen possible high quality essays in the proportions shown in Table 7.5.

Table 7.5: Essays identified by the training set conditions

Condition	1	2	3	4	5	6	7	8
No. of essays	9	2	2	1	0	0	1	1
Total	9	11	13	14	14	14	15	16

Table 7.5 shows that there were nine essays identified by the first condition and two further essays by the second to give eleven different essays identified by the first two conditions. The first three conditions identified 13 total essays and a total of sixteen essays were finally identified by the eight conditions.

7.3.2.2.2 Bayesian classification

These sixteen probable high quality essays identified as the training sample were also initial members of the high quality group. The next stage of the analysis was to use a Bayesian classifier to assign a further 34 essays to the high quality group leaving the remaining fifty essays to form the low quality group.

Occurrences of all words in an essay that appeared in more than one essay in the set of essays were considered. For each word in each essay, the occurrences of that word in the essays in the high sample group and in the essays of the group that includes all other essays were tallied. For example, say the first word of an essay under consideration is *once*. The occurrence of *once* is then checked in the sixteen training essays. Lets say *once* occurred in eight out of these sixteen essays. If *once* only occurs twelve times in the remaining 83 essays, *once* has a higher relative occurrence in the high quality sample than in the group of other essays. Therefore, on the basis of the occurrence of *once*, the essay seems more likely to belong to the high quality sample than to the remaining group. Of course, on its own, this one word is not a reliable predictor, but when we take into account every other word in the essay in a similar way, the Bayesian classifier will leave us a more reliable probability that the essay belongs to the high quality group or the remaining group.

This procedure was done for all the remaining 84 essays and the 34 essays with the highest probability of belonging to the high quality group were allocated to that group. The remaining fifty essays were allocated to the low quality group.

Bayesian classifiers have certain technical problems some of which are particularly related to dealing with essay type data. A set of data transformations were carried out to counter these problems. These problems and the data transformations to counter them are reviewed in Appendix 7.4.

7.3.3 Results

7.3.3.1 Bayesian algorithm

A comparison of decision agreements and Kappa statistics from the Bayesian algorithm with human ratings is shown in Table 7.6.

Table 7.6: Decision agreement and Kappa reliability for Bayesian algorithm

	Rater 1		Rater 2	
	DA	Kappa	DA	Kappa
Bayesian	.70	.40	.70	.40
Rater 1	-	-	.72	.44

This algorithm agreed with Rater 1 in 70 cases out of a 100 and with Rater 2 in 70 cases out of a 100. The two raters agreed with each other in 72 cases out of a 100. After adjusting for chance agreement, we get a Kappa statistic of r = 0.40 for the reliability of the algorithm with each rater compared with r = 0.44 for raters with each other. This shows that the performance of this algorithm is not as good as but approaches that of two human raters.

7.3.3.2 Reliability of training essays

To check the reliability of the training set essays, the ratings of the sixteen selected essays were compared with ratings for these essays by the two human raters. The number of training sample essays rated high by each rater by selection condition are shown in Table 7.7.

Table 7.7: Training set essays rated high by human raters

	condition					
	1	2	3	4	7	8
No. of essays	9	2	2	1	1	1
Rater 1	9	2	2	1	0	1
Rater 2	9	2	1	1	1	1

The table shows that both raters rated all but one training set essay as being high quality. No essay was rated poor by both raters. The essay rated poor by Rater 1 was an essay selected by condition 7 and the essay rated poor by Rater 2 was an essay selected by condition 3. All the essays selected by the first two conditions were also rated high by both raters.

7.3.4 Discussion

7.3.4.1 Comparison of algorithm ratings with pooled human ratings

Table 7.8. shows the essays the human raters agreed on and how they were treated by the Bayesian algorithm.

Table 7.8: Automatic treatment of essays agreed on by raters

		Raters	
		High	Low
Bayesian	High	30	10
	Low	6	26

The number of agreements between the Bayesian algorithm and the pooled ratings was 56 cases out of 72 or 78% and the Kappa statistic was 0.56. Out of 36 essays rated as high quality by both human judges, 30 were also rated high quality by the Bayesian algorithm. This means that six essays rated high by both raters were rated low by the Bayesian algorithm. Similarly, out of 36 essays rated low by both human raters, 26 were also rated low by the Bayesian algorithm. This means that ten essays rated low by both raters were rated high by the Bayesian algorithm. Again, we can not be sure whether these essays are being misplaced by the algorithm or whether the agreement by the raters is simply chance agreement in the rating of borderline cases. It is interesting though that the relative weakness of the Bayesian algorithm seems to be focused in its prediction of low quality essays. This may be a result of concentrating the analysis on high quality essays. This may have been an argument for also running a Bayesian algorithm on low quality essays at least as a check.

7.3.4.2 Comparison of Bayesian classifier with pooled human ratings

Using the Bayesian algorithm, essays are allocated to the high quality group by two processes. The first process is by the training set conditions which identified sixteen essays. The second process is by the Bayesian classifier which allocated a further 34 essays to the high quality group and the remaining fifty to the low quality group. Therefore, the agreement in high quality candidates cannot be credited entirely to the Bayesian classifier.

Fourteen of the sixteen training set essays were agreed on by both raters. Therefore, the essays agreed on by human raters allocated by the Bayesian classifier are shown in Table 7.9.

Table 7.9: Bayesian classifier treatment of essays agreed on by raters

		Raters	
		High	Low
Bayesian classifier	High	16	10
	Low	6	26

There were 58 essays that the raters agreed on that were allocated by the Bayesian classifier. The agreement between the results of the classifier and the raters was 42 out of 58 cases which is agreement of 72% or a Kappa statistic of 0.47.

Of the sixteen essays identified by the training set, there were fourteen whose rating was agreed on by both raters. These fourteen were all rated as high quality by the training conditions. This suggests that the training set identification portion of the Bayesian algorithm may be more reliable than the Bayesian classifier portion of the algorithm.

7.3.4.3 Adjustments to the algorithm

There may be considerable scope for adjusting the conditions for selecting the training sample. The selection of the training set essays is crucial. If an essay is mistakenly assigned to the training set at this stage, the repercussions are likely to be felt throughout the analysis. The nature and the number of the conditions is of particular interest. In this experiment, the strictest two conditions proved reliable but there were problems with some essays chosen with subsequent conditions. Also, the optimal number of essays to put in the training set needs further investigation. However, this choice may depend on the conditions themselves. More research into the reliability of these conditions in identifying high or low quality essays is vital.

In this experiment, the training set conditions all related to identifying the high quality group. The algorithm as a whole concentrated on high quality essays since it seemed that they could be identified more reliably given the tendency for them to include more words. However, training samples could be used to reliably identify some low quality essays to at least remove them from the analysis. Perhaps similar features could provide conditions for low quality essays. Very short essays with low lexical diversity and hapax legomena values and high error would probably identify low quality essays. In Chapter Five, essays with a TTR score higher than a lexical variation score were found to be low quality.

In addition, the reliability of classifiers with such small training samples needs to be considered. The study by Larkey (1998) suggests that a minimum number of training set essays may be necessary for a Bayesian classifier to work properly. Since this experiment uses a smaller number, there may be validity concerns. This weakness may be compensated for by using it alongside other algorithms as a check. For example, if a clustering algorithm or another Bayesian algorithm based on features rather than word content produced similar results, then validity may be bolstered.

7.3.5 Conclusion

In this experiment, a Bayesian algorithm has been shown to be a possible alternative to human assessment for the rating of L2 learner essays. A training set of essays to guide a Bayesian classifier was identified automatically from within the set of essays. These essays were identified by combinations of features which typically indicate high quality essays. On the basis of this training sample and using a Bayesian classifier, other essays were classified as being similar or different according to their semantic content. Similar essays were classed as high quality and non-similar essays as low quality by default. The initial training sample analysis identified sixteen high quality essays. The subsequent Bayesian classification identified a further 34 similar essays. The remaining fifty essays were classed as poor quality.

In this experiment, the ratings produced by the algorithm corresponded with assessments of human raters almost as well as the human raters corresponded with each other. The training set essays were all rated high quality by at least one human rater and fourteen out of sixteen essays were rated high quality by both raters. A weakness in the approach in this experiment seems to be in essays rated as poor quality. This suggests that more attention needs to be focused on identifying poor quality essays. In this experiment, the training essay selection portion of the algorithm seemed to perform more reliably than the Bayesian classifier portion.

7.4 Discussion

7.4.1 Comparison between clustering and Bayesian algorithms

Results show that both algorithms are almost as reliable in classifying the essays as the two native judges. Table 7.10 shows decision agreements (out of 100) for each method and rater. The results of both algorithms compare favourably with ratings by humans. The clustering algorithm seems to produce the best results but the Bayesian algorithm also seems promising.

Table 7.10: Decision agreement between raters and automatic methods

	Rater 2	Clustering	Bayesian
Rater 1	72	78	70
Rater 2	-	76	70

Table 7.11: Classification of essays agreed on by both raters

	Clustering		Bayesian	
	high	low	high	low
Pooled rater high	32	4	30	6
Pooled rater low	5	31	10	26

Table 7.11 shows the essays where there were agreements between Rater 1 and Rater 2 and how these essays were dealt with by each automatic method. There were 36 essays that were rated high quality by both Rater 1 and Rater 2 and 36 essays that were rated low quality by both raters. Of the 36 essays that were rated high by both raters, the clustering algorithm identified 32 of these as high quality whereas the Bayesian algorithm only grouped 30 of them as high quality. This means the clustering algorithm classified four of the agreed high quality essays as low quality and the Bayesian algorithm did the same for six essays. Of the 36 essays that both human raters assessed as low, the clustering algorithm also identified 31 of these as low quality whereas the Bayesian algorithm only grouped 26 of them as low quality. This means the clustering algorithm classified five of the agreed low quality essays as high quality and the Bayesian algorithm did the same for ten essays.

Table 7:12: Agreed essays by raters and algorithms

		Pooled algorithms	
		high	low
Pooled raters	high	28	2
	low	4	25

Table 7.12 shows the 59 essays that were rated the same by the two raters and rated the same by the two automatic algorithms. There were 28 essays rated high by both human raters, and both algorithms. Similarly, there were 25 essays that were rated low by both raters and by both algorithms. There are two essays that were rated high by both raters but were rated low by both automatic methods. Similarly there were four essays that were rated low by each rater but rated high by both automatic methods. These special cases suggest there may be some essays that are not being correctly assessed by automatic methods. It may be instructive to examine such essays in more detail to see if there are any characteristics that may identify them and can be incorporated into any future algorithm to facilitate more accurate rating.

With hindsight, it would also have been better to have the essays rated by one or two more raters to help evaluate whether there were any problems with the automatic assessment algorithms. These extra raters would afford greater reliability in human rating results and allow us to better spot anomalies in the results of the automatic systems. A strength of the study by Page (1994) was the comparison with high numbers of human raters.

7.4.2 Inter-algorithm reliability

Although one of the advantages of automatic assessment is that it eliminates some of the reliability concerns which plague human rating, it does come with some concerns of its own. Two important reliability concerns for humans are inter-rater reliability and intra-rater reliability. Inter-rater reliability refers to the consistency of score awarded to the same essay by different raters. If two raters score the same essay differently, it is an obvious problem. Intra-rater reliability refers to the inconsistency of rating by a particular rater. On different occasions the rater might award different grades to the same essay. While automatic assessment eliminates these two concerns, it does raise another reliability issue of its own: inter-algorithm reliability. There are potentially many different algorithms for automatic assessment. This chapter includes just two out of a great number of possibilities. Inter-algorithm reliability is concerned with the consistency of rating between algorithms. It checks to see if different algorithms rate the same essay in a similar way. The agreement between the two algorithms in this chapter is shown in Table 7.13.

Table 7.13: Agreement of two automatic algorithms

		Clustering	
		High	Low
Bayesian	High	39	11
	Low	11	39

Comparing the performance of the clustering algorithm and the Bayesian algorithm, we find that there was agreement in 78 cases out of a 100. There was agreement in 39 high cases and 39 low cases. This translates to a Kappa inter-reliability measure of 0.56. This is higher than human inter-rater reliability but there is still considerable scope for improvement. In particular, since one of the arguments for automatic assessment is to improve reliability, we might hope to achieve a higher value.

7.4.3 Essay length

It is worth considering the relationship of essay length to the various ratings. There are two key aspects to this. Firstly, we can check whether the algorithms perform better than a simple classification according to essay length. Secondly, we can check whether the algorithms are overly dependent on essay length. In order to consider these questions, the essays were classified according to length only with the fifty longest essays classed as good quality and the fifty shortest essays classed as poor quality. The results of this classification were then compared with the classifications by raters and algorithms as shown in Table 7.14.

Table 7.14: Agreement of essay length with raters and algorithms

	Rater 1	Rater 2	Clustering	Bayesian
Essay length	74	74	88	88
Clustering	78	76	-	78
Bayesian	70	70	-	-
Rater 2	72	-	-	-

The results show that essay length is a reliable predictor of rating for this set of essays. In fact, essay length correlates better with both raters than the raters do with each other. There were 74 agreements out of a 100 between an essay length model and either rater but only 72 agreements between the two raters. However, the clustering algorithm correlated more closely with both raters than essay length alone. The clustering algorithm matched Rater 1 in 78 cases compared with 74 for essay length alone and matched Rater 2 in 76 cases compared with 74 for essay length alone. This suggests that this algorithm may be an improvement over using only essay length as a predictor. However, essay length alone performs better than the Bayesian algorithm in terms of agreements with human raters with 74 matches compared with only 70 for the Bayesian algorithm.

The results also show that essay length is very strongly correlated with the results of the two automated algorithms with 88 out of a 100 matches. It is worth noting that both algorithms, although scoring 88 matches with essay length, showed different patterns of matches. This evidence suggests that these algorithms may be overly dependent on essay length.

This evidence of a relationship with essay length while inconclusive is intriguing. It might be possible to test these algorithms on sets of essays where quality is not so correlated with essay length. This would be useful for a number of reasons. Firstly, it would present an opportunity to identify other features that may be useful in calibrating quality but are being dominated by essay length in timed essays. Secondly, it presents an opportunity for examining automatic assessment in situations where length is not such an important factor. Most of the experiments in this study have involved timed essays where length is often the overwhelming factor. There are a number of types of essays where quality is likely to be less related to length. For example, essays of higher proficiency learners tend to be longer but the differences in length are not likely to effect quality to the same degree as differences in length in essays of lower proficiency learners. The length of non-timed essays may also vary to a lesser degree than the length of timed essays. In addition, essays could be controlled for length. In experiments, sometimes samples are taken of fixed length to solve the problem of length effects. However, this practice introduces the concern that longer essays may be penalized because a shorter sample may not be representative of the longer essay. One way around this is to control for length in the task design and require learners to produce an essay of a fixed length.

7.4.4 Selecting an algorithm

This chapter has proposed and compared two algorithms. Both the clustering and the Bayesian algorithm seem to be effective at correctly assigning essays as either high quality or low quality. Both algorithms presented here are quite crude implementations of clustering and Bayesian classification. Improvements in these techniques are likely to be reflected in enhanced results. For example, with clustering, various clustering techniques could be used, i.e. the conditions by which elements are added to clusters could be varied. Similarly, there are various distance measures that could be used.

Both the clustering and the Bayesian algorithms are quite complicated and so it takes time to understand exactly how they work and how to improve them. Both algorithms could also be improved by feedback. As the applications are used, results could be fed back to improve performance. For example, typical occurrences of commonly occurring words or ranges of lexical features that are high predictors of quality could be 'learned' by the application and given prominence in the analysis.

The two experiments highlight two different approaches. The clustering algorithm depends on lexical features. The Bayesian algorithm also uses lexical features to identify a training set but then classifies the remaining essays according to the semantic similarity of essays. The approaches could be tried in reverse. The clustering algorithm could be based on semantic content. In this form, it would resemble the process of latent semantic analysis. Alternatively, a Bayesian algorithm could be based on lexical features. Bringing these two approaches together in the same algorithm could enhance the algorithm by increasing the amount of information involved and thereby boosting the reliability of the results. This could be achieved either by producing a new algorithm which included both types of analysis or by doing two separate algorithms and somehow pooling the results.

Both methods have solved the training set problem in different ways. The clustering algorithm avoided the need for an external training set by setting an imagined initial cluster point based on the orientation of the different features. The initial cluster points could be chosen in a different way to possibly boost performance. One way to investigate this might be by mapping the initial cluster points in relation to the essays that are added to each cluster. If a lot of essays cluster closely around these points, then they may be good choices. If the essays seem to cluster in other places then the selection of initial cluster points could be rethought. In the Bayesian algorithm, a training set was detected automatically from within the set of essays by identifying essay features that indicate high quality essays with a high degree of probability. It is likely that other high quality or low quality conditions could be found or some of the conditions may prove more reliable than others.

7.5 Conclusion

In this chapter, two assessment algorithms have been considered. In the first, an algorithm based on clustering was used to classify 100 essays as high quality or poor quality. A number of essay features were used to cluster the essays. These features were refined from a larger group of features by a principal component analysis which finds features that are independent of each other and account for maximum variance. Once six optimum features had been identified, the essays were clustered around two cluster points, one high quality and the other low quality. The results of this clustering algorithm agreed with human raters in 78% and 76% of cases which was more agreement than between the two raters.

The second algorithm was based on a Bayesian classifier and was tested on the same rated essays as the clustering algorithm. Bayesian classifiers require training samples. A training sample of potentially high quality essays was identified by utilizing a finding of the previous chapter that essays scoring very highly on both essay length and lexical diversity tend to be good essays. Using such conditions, sixteen likely high quality essays were identified and used to train a Bayesian classifier to recognize high quality essays by semantic content. The results of this Bayesian algorithm agreed with human raters in 70% of cases. The algorithm seemed to be more accurate in classifying high quality essays than poor quality essays. The training sample selection portion of the algorithm was more reliable than the Bayesian classifier portion. There may be problems with using a Bayesian classifier with so few training essays.

The two algorithms in this chapter represent only two out of a huge number of possible algorithms that could be tried. Other classification methods could be tried including more complex systems. Also other ways of implementing clustering and Bayesian classifiers could be tried.

In the next and final experimental chapter, a more unusual approach to essay assessment that may be suitable when multiple essays are available is considered.

Chapter Eight: Capture-recapture analysis

8.1 Introduction

Chapter Six and Chapter Seven investigated various models for assessing essays using a number of complementary essay features. This final experimental chapter investigates an alternative approach that could be applied in some special circumstances such as when multiple essays are available or when essays are particularly long.

Capture-recapture is a method developed in biology to estimate animal or fish populations in the wild. It has recently been adapted for a number of vocabulary based studies in L1 (Marcus et al., 1992) and in L2 (Meara & Olmos Alcoy, 2008) as a possible way to estimate vocabulary knowledge by the 'capture' of language samples. Marcus et al. used L1 spoken samples of children while Meara & Olmos Alcoy used essays from L2 learners of Spanish. If capture-recapture analysis of an essay can provide estimates of vocabulary knowledge of the writer, then it might also be used to assess quality of the essay itself. The method could then be used to aid assessment in situations where multiple essays can be collected, for example, in coursework or for online self-assessment applications. This experiment investigates the suitability of the capture recapture method for assessing the quality of L2 learner essays.

8.1.1 Background to capture-recapture analysis

Capture-recapture methods were developed for estimating wild animal and fish populations (Seber 1982). The method is quite basic. In the simplest case, two samples are taken from the population. The individuals captured in the first sample are tagged and then returned to the population. After allowing the population to mix freely, a second sample is then captured and the numbers of tagged and untagged individuals is recorded. Three groups can now be identified.

1) individuals that were captured in sample 1 (n_1)
2) individuals that were captured in sample 2 (n_2)
3) individuals that were captured in both samples (n_{12})

In order to estimate the population size, N, the following Petersen estimate of population size, Ñ, can be used.

$$\tilde{N} = \frac{n_1 \times n_2}{n_{12}}$$

This estimate involves the following assumptions:

a) there is no change to the population over the time period of the two samples.

b) individuals do not lose their tags.

c) for each sample, each individual has the same probability of being in that sample.

This is a simple model and is not expected to perfectly represent any actual situation. As such, the assumptions may not reflect the real world. For example, in the real world, there are likely to be changes to the population in terms of births and deaths. Some animals may lose their tags. Some animals may have greater chances of being in the sample, perhaps depending on species, age, size, temperament or other factors. In addition, perhaps animals once caught are less/more wary and so more/less likely to be caught in a subsequent sample.

8.1.2 Application of capture-recapture analysis to L2 lexical studies

This model can be applied to the estimation of the number of words in the lexicon of second language learners. Samples of language use can be taken and individual words captured, tagged and recaptured. The method was used by Marcus et al. in an L1 investigation into overregularization of verbs in children using spoken samples. Meara & Olmos Alcoy used it to investigate vocabulary knowledge of L2 learners of Spanish. When using this method in a language context, there are some obvious problems with the aforementioned assumptions. The first assumption is clearly violated. There is likely to be a change in the population of words known over the time period of the experiment. For example, no matter how close together in time the samples are taken, there are likely to be new words added to the lexicon and maybe other words lost through attrition. However, these changes are likely to be minimal in relation to the total words under consideration.

A more serious problem is in relation to the third assumption. Words are not equally likely to be captured in a sample. Some words are used far more frequently than other words. There are two noticeable cases. Firstly, content words that relate to that context are more likely to occur in the sample than words unrelated to the context. In fact, many non-contextual words have little or no chance of occurring in the sample. For this reason, Meara & Olmos Alcoy argue that it might be better to think of the population under investigation, N, as the total number of words available for a particular task rather than the total words in the lexicon of a learner. They suggest controlling for context by getting learners to write two essays on an identical task.

Secondly, function words tend to occur far more frequently than content words and so are more likely to occur in the sample. This might be expected to lead to an underestimate of total words available for the task. If we think of it in terms of catching fish, function words may act as huge fish, grabbing the bait and denying the chance for some smaller fish to get into the samples. In reality, all words have differing chances of being included in the sample. Marcus et al. cope with this by using a Jackknife estimate (Burnham & Overton, 1978) over multiple captures. By using multiple recaptures rather than just one, a more detailed model can be constructed. For example, Marcus et al. used five captures. This meant that some words were captured in only one sample, some in two, three and four samples, and some in all five. This enabled them to broaden their model to allow for words to have different probabilities of being in the sample. This is not possible with a two sample analysis. In a two stage analysis, one way to cope with high frequency function or contextual words might be to take them out of the analysis. If we analyze only the words that occur once in each essay, then this is likely to eliminate function and highly contextual words.

If, despite possible violations of assumptions of the model, this method can produce an estimate for productive vocabulary size of the learner, it seems reasonable that this same estimate should inform about the qualities of the essays themselves. This depends on the assumption that essays suggesting a writer has a relatively large vocabulary are also relatively good quality essays.

For the method to work, we need to check that the essays involve enough words to produce a meaningful estimate. In order to do this, in this experiment, a capture-recapture analysis is conducted on two identical tasks by the same learners. Two analyses are carried out. One uses all words in an essay and the other uses all words except high frequency words. For this experiment, high frequency words are defined as words that occur more than once in either language sample. The sizes of the estimates of the two analyses might be expected to differ for several reasons. One reason is that they will be measuring different things. Following Meara & Olmos Alcoy's lead, the estimate of the full analysis can be thought of as an estimate of the productive vocabulary of the learner available for the task. The estimate of the second analysis will be an estimate of the productive vocabulary available for the task minus function and highly contextual words.

The two estimates may also vary because we expect the first analysis to give a underestimate of actual vocabulary size because of the effect of high frequency words. The second estimate should, in theory, be more accurate.

For assessment purposes however, we are not interested in absolute values but rather relative values. If the values from the two analyses correlate strongly, then we can assume that the full analysis gives a reasonable estimate despite the violations of the assumptions.

8.1.3 Aims of the experiment

The aim of the experiment is to test:

1) if a capture recapture analysis without high frequency words gives a different estimate to the full analysis.

2) how many words are necessary to produce a reliable estimate.

8.2 Methodology

8.2.1 Participants and tasks

Two written tasks were attempted by a group of 36 university-level Japanese learners of English. The tasks were identical but were conducted at an interval of several weeks. Students were asked to produce a sample of writing based on a six-caption cartoon prompt. The cartoon used is Cartoon Task 1 in Appendix 3.1. The students were given 25 minutes to write as much as possible.

8.2.2 Analysis

A capture-recapture analysis was performed using a specially designed computer program, L-capture (see Appendix 1.1). Two analyses were carried out. In the first analysis, all the different word types produced by each learner were used. For each essay, the total number of word types on the first task (n_{t1}), the total number of word types on the second task (n_{t2}) and the total number of word types used in both tasks were calculated (n_{t1t2}). The Petersen estimator ($\tilde{N}$) was calculated for each student as follows:

$$\tilde{N} = \frac{n_{t1} \text{ x } n_{t2}}{n_{t1t2}}$$

A second analysis was conducted with words that occur no more than once in each task. This should eliminate frequently occurring function words and words strongly related to the picture context. Again, the same computer program, L-capture, was used for the analysis. Using these words only, the new population being estimated is slightly different since it does not include multiple-occurring words. If this new population under investigation is N_α, then the Petersen estimate of this can be expressed as $\tilde{N}_\alpha$. This estimate was calculated for all students as follows:

$$\tilde{N}_\alpha = \frac{n_{at1} \times n_{at2}}{n_{at1t2}}$$

where n_{at1} is the number of word types on the first task excluding multiple-occurring words, n_{at2} is the number of word types on the second task excluding multiple-occurring words and n_{at1t2} is the number of word types occurring in both tasks excluding multiple-occurring words.

8.3 Results

Table 8.1: Results of full capture-recapture analysis

	n_{t1}	n_{t2}	n_{t1t2}	$\tilde{N}$
mean	52.69	59.14	28.67	110.3
standard deviation	10.15	14.28	7.63	26
coefficient of variation	19.26	24.15	26.60	23.56
highest value	99	121	64	187.2
lowest value	40	35	19	53.85

Table 8.1 shows the results of a full capture-recapture analysis. The average number of words available for the task, Ñ, was 110.3. There was quite a large range of Ñ from 53.85 to 187.2 perhaps suggesting a range of quality in the essays.

Table 8.2: Capture-recapture analysis using words only occurring once in a task

	$n_{\alpha t1}$	$n_{\alpha t2}$	$n_{\alpha t1t2}$	$\tilde{N}_\alpha$
average	27.44	33.08	8.92	109.58
standard deviation	5.91	9.12	3.35	39.96
coefficient of variation	21.53	27.53	37.6	36.46
highest value	51	71	23	200.66
lowest value	18	16	5	32

Table 8.2 shows the results of a capture-recapture analysis using words that only occur once in a task. The average estimate, $\tilde{N}_\alpha$, is 109.58 which was very similar to the average estimate in the full analysis. This is deceptive however since this was an estimate of a different population, total words available for the task excluding function words and core content words. The similarity in population averages underlines the fact that a full analysis underestimated the number of words available for the task. The range of values was wider in this reduced analysis than in the full analysis. Values ranged from 32 to 200.7 and also reflected in the higher coefficient of variation, 35.6 in this analysis compared to 29.2 in the full analysis.

The second analysis should give a better theoretical estimate since the data was adjusted to match the assumptions of the model. To see if this translates into a more accurate and reliable result, the statistics and results of each analysis can be examined..

Although, the Petersen estimate was about the same in both analyses, there were fewer words used in the second analysis. Particularly noticeable was the small number of words that appeared in both samples. There was an average of 8.92 and a range from 5 to 23. The low number of words at the bottom of this range could be a concern as it may lead to unreliability. The larger range of scores in the reduced analysis may show that the score is more sensitive but may also reveal unreliability because of the smaller numbers involved in the analysis. Looking at the statistics, there must be some concern about whether the reduced analysis contains enough words to produce a reliable estimate.

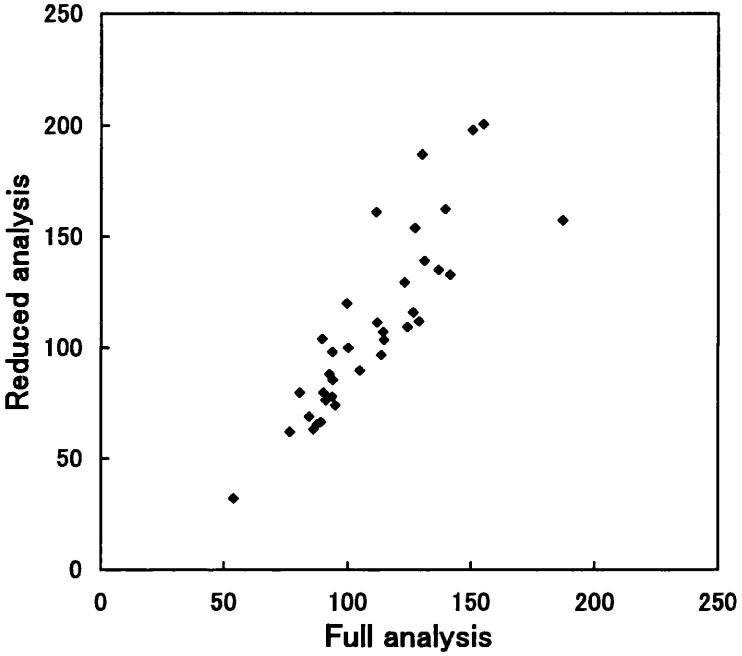


Figure 8.1: Estimates for each essay on both analyses

By comparing the results, the relative merits of the two analyses can be appraised. If the results are very similar, then it means that the reduced number of words does not give an improved performance. However, if the results are quite different, then the conclusion is less clear. It may mean the reduced analysis is better and that including the high frequency words in the analysis has distorted the results. However, it may also mean that the reduced analysis is not as good, perhaps because the low number of words leads to unreliability. It may also mean that neither analysis is very accurate. To test the similarity of results, a correlation analysis was carried out on the results of the full analysis and the reduced analysis. The results correlated with r = .85. This was very significant but not as high as we might wish for measurement reliability. It seems to indicate that both estimates were broadly measuring the same thing but there was a small difference. The nature of the difference is less clear.

Figure 8.1 shows the estimates for each essay on the two analyses. The graph shows a broad agreement between the two analyses. However, the variation seems to increase with larger scores and the two analyses seem to produce similar scores at the lower end of the scale. More variation in absolute terms is to be expected because of larger values. For example, 10% of 200 is larger than 10% of 50. However, the variation at the upper end seems to exceed this. The difference at the upper end is highlighted by an outlier essay that got the highest estimate of close to 200 on the full analysis but only the sixth highest estimate on the reduced analysis. The relative agreement at the bottom of the scale suggests that the reduced analysis may be quite robust. The worry about the reduced analysis is that the small number of words may lead to unreliable estimates. If this was the case, then differences between the estimates are likely to be noticeable at the bottom end of the scale. The differences at the top end of the scale seem to point to fluctuations caused by the number of multiple occurring words in the full analysis. This suggests that the increased accuracy possible by a reduced analysis may outweigh any potential loss of accuracy by using less data. However, this interpretation is speculative. The results are far from conclusive and further investigation of estimates against quality assessments would be helpful.

8.4 Discussion

8.4.1 Sampling reliability

The capture-recapture method depends on sampling from a population N. Sampling involves error and this error could lead to considerable range of scores over repeated samplings. A wide range of sampling scores is problematic for a prospective measurement tool. Given the very small sample sizes involved in these analyses, the effect of this error could be serious. To investigate the extent of this sampling error, a population of size N was sampled numerous times with sample size S1 and S2. A capture-recapture analysis was applied to these samples and the range of $\tilde{N}$ estimates compared. For comparison with the experimental data, similar population and sample sizes were taken. In the full analysis, the mean estimate of population size, $\tilde{N}$, was 110.3 and average sample sizes were 53 and 59. In the population using only words occurring once, the mean population estimate, $\tilde{N}_\alpha$ was 109.6 and the sample sizes were 27.4 and 33. Therefore two computerized simulations were undertaken to reflect these different parameters. The L-capsim computer program (see Appendix 1.1) was used for these simulations. The first simulation was to mirror the full analysis and used a population of N=110 and samples sizes of 53 and 59. In the second simulation, the population size was also 110 with sample sizes of 27 and 33. Using these simulations, the precision with which writing samples of these sizes can accurately predict the true population can be investigated. The results of the simulations are shown in Table 8.3.

Table 8.3: Values of Ñ for two simulations

	simulation 1	simulation 2
N	110	110
first sample size	53	27
second sample size	59	33
number of samplings	1000	1000
average population estimate Ñ	111.165	110.76
standard deviation of Ñ	10.27	29.20
Coefficient of variation	9.24	26.36
High	147.27	317.33
Low	85.26	63.47

Not surprisingly, simulation 1 involving the larger samples showed much less variability in the range of population estimates with a coefficient of variation of 9.24. Simulation 2 had a large range of values and a coefficient of variation of 26.36. If this value is compared to the coefficient of variation of the scores of the L2 learners, the variation is quite similar. This suggests that for low sample sizes, the variation inherent in any individual value of the estimator $\tilde{N}_\alpha$ may be as great as the variation between different learners raising doubts as to its suitability for measurement purposes. Furthermore, the reliability of sample size using N=110 with varying sample sizes (but both samples S1 and S2 of equal size) shows that at low sample sizes there is great variation.

Figure 8.2 shows the range of values for various sample sizes with the upper line showing the mean plus one standard deviation and the lower line the mean minus one standard deviation. This graph suggests with a sample size of less than 40, there is likely to be a wide range of values. With a sample size of just under 60, the estimator should be within plus/minus 10 of the true population value.

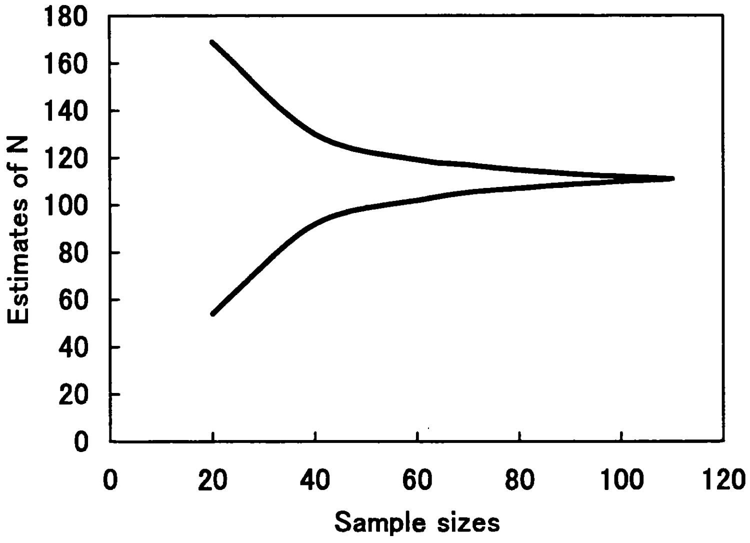


Figure 8.2: Range of N estimates for various sample sizes for population size 110

8.4.2 Minimum essay length

In the sampling simulation, a sample size of 60 words or so is necessary to give a Petersen estimate within about 10% of the true population value. However, this sample size is not essay length. In the full analysis, sample size is the number of word types in the essay.

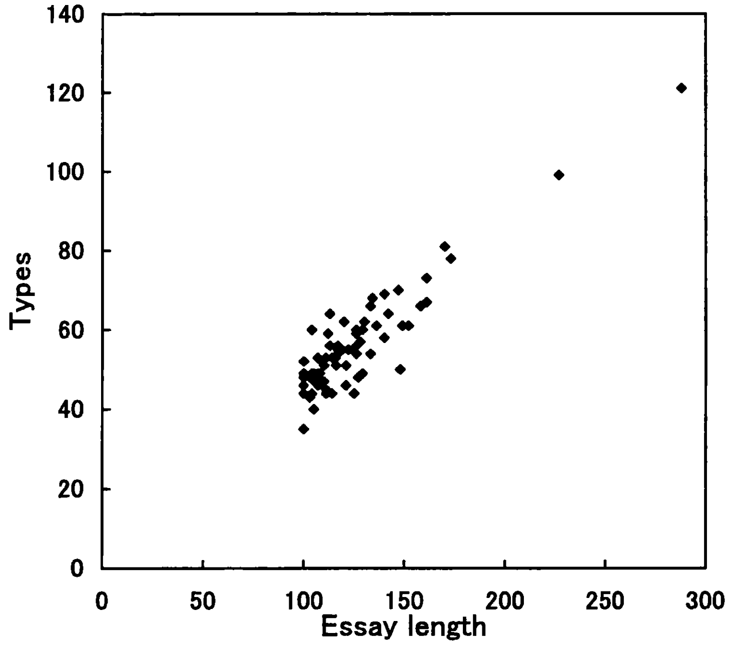


Figure 8.3: Number of word types and essay length

Figure 8.3 shows the number of word types plotted against essay length for the essays in this study. This suggests that essays of about 120 words usually produce about 60 word types. The number of word types occurring no more than once in either essay is shown in Figure 8.4.

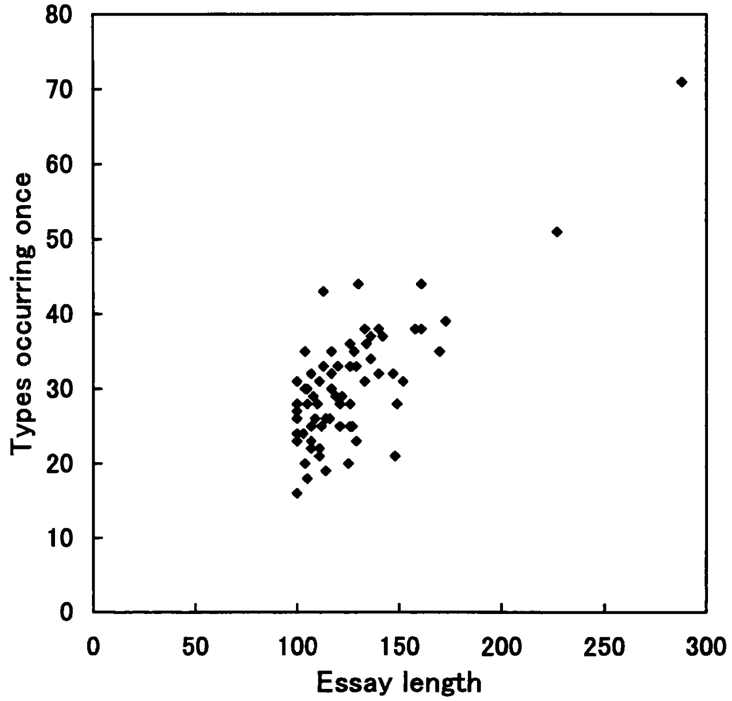


Figure 8.4: Number of types occurring only once and essay length

Although most essays in this experiment produced far fewer of these words, we can predict from the linear trend that about 250 words would be necessary to produce an estimate within 10% of the true population value.

The conclusion is that to get the best estimate and in order to allow for the assumption of all words having an equal probability of occurring in the sample, essays of around 250 words are necessary. If this assumption is to be ignored then upwards of 120 word essays may be sufficient. Of course, the latter estimate may be susceptible to error due to violation of the assumption. These estimates are only a guideline. They may help explain the slightly low measurement reliability between the results of the two analyses. The full analysis may not be accurate because of the violations of the assumption regarding equal likelihood of word capture. The reduced analysis may also not be accurate because of a shortage of words.

8.4.3 Using capture-recapture analysis for assessment.

The obvious problem with using this method for assessment purposes is that it requires at least two samples of written work. Many assessment situations cannot accommodate the collection of multiple writing samples. However, it could be an excellent choice for continual assessment situations where students submit written assignments over a period of time, say, a one year university course. Here not only could it give a guide to essay quality but could also measure development in proficiency. With multiple samples, the system could probably be used with some accuracy. In Marcus et al., capture-recapture was used with a Jackknife estimate over not only two time points but a total of five captures or recaptures. Repeated samples enable use of a more realistic model which takes into account the fact that words have differing probabilities of occurring in a sample. This means all words in the sample can be incorporated into the analysis and so helps minimize sampling variability. This considerably improves the precision of any estimates. With multiple samples of writing, less words will be necessary for each capture. The method should be more accurate at estimating the size of the productive vocabulary used even for essays on different topics.

For the typical one time testing situation, the capture-recapture method seems a little more difficult to incorporate into a system. One possibility could be if learners produced two pieces of writing in a testing situation and these were assessed. Some tests, such as the IELTS written portion, use two written tasks. Research could be useful as to whether two samples of differing content could be incorporated to produce an estimate. Another possibility could be with more advanced learners who are asked to produce an essay that is long enough to be split into two halves and assessed in this way. Given the results of this study, to produce enough types for a reasonably reliable estimate for two samples would need an essay of at least 240 to 500 words. Of course, a split half methodology may also introduce other problems. For example, there may be stylistic differences between two halves of an essay which could affect the analysis.

8.5 Conclusion

The experiment in this chapter involved using capture-recapture analysis to find an estimate of a learner's vocabulary knowledge and using this estimate to determine the quality of the learner's essay. This experiment has highlighted a number of problems with employing the capture-recapture method for the measurement of essay quality. Firstly, unlike capture of wild animals where an individual animal can only be captured once, it is possible for words to appear multiple times in a sample. Secondly, words have differing probabilities of appearing in a sample. Function words are more likely to appear than content words and contextually relevant words are more likely to appear than contextually incompatible words. This complicates the process of getting a reliable estimate. Thirdly, sampling reliability dictates samples be as large as possible but L2 learners written output tends to be sparse.

In this experiment, two analyses were carried out to find an estimate of essay quality. One analysis included all the words in the sample. The other included only words that appeared once. This was to control for the influence of words that appear often in an essay such as function words. Including these words may give a biased estimate of essay quality.

The results of the two analyses correlated highly which suggested that the full analysis may give a reliable estimate despite its theoretical limitations. The weakness with the second analysis is that it depends on a small number of words and may be affected by sampling variability. The results of a computer simulation suggested that about 60 words per sample would be necessary for a reliable estimate. A sample of 60 words translates to an essay length of about 120 words for a full analysis or about 250 words for a reduced analysis. These results are intriguing but more research is necessary into how these estimates relate to quality assessments. Using them to estimate essay quality also depends on the further assumption that an essay that suggests a writer has a large vocabulary is also a good essay.

Another major problem with this method is the need to repeat a task for a second sample. This makes it difficult to use in regular testing situations. An alternative approach might be to split a text into two and then do an analysis on each half. Of course, the minimal essay length would double using this approach. The capture-recapture approach seems ideal for a continual assessment situation. If more samples of writing can be collected, then a more complex model can be used which incorporates words with differing probabilities of occurrence. Because this kind of model could incorporate all the words in an essay without violating the basic assumptions, it could provide a more valid and reliable estimate with relative short written samples.

This chapter has concluded the experimental section of the thesis. The next chapter is a discussion of some of the issues raised by this experimental work.

Chapter Nine: Discussion

9.1 Introduction

This chapter discusses some issues raised by the results of the experimentation in this thesis. The first part looks at some of the lexical features which have appeared in this study and gauges their strengths and weaknesses for automatic assessment. The second part reflects on how this research fits in with research in L2 vocabulary studies. The third part looks at some of the issues involved with automating the assessment process. The fourth part considers some of the limitations of this study. In the final part, some further avenues of research into automatic assessment are explored.

9.2 Lexical features

In this study, various lexical features were examined for their suitability in gauging quality of L2 learner essays. In this section, some essential properties of these features are investigated. Length of text is considered separately because of its unique role in assessment.

9.2.1 Properties of lexical features for assessment purposes

The following are three essential properties of a feature for use in assessment of low level learners:

1) It is able to identify differences in essays of varying quality particularly in low level learners.
2) It shows reliability over different pieces of work by the same learner.
3) It can be calculated and interpreted easily and automatically.

The following property is also attractive:

4) It exhibits some validity as an indicator of quality.

9.2.1.1 Identifying quality differences in essays of low level learners

The first point, an ability to identify differences in quality, is an obvious requirement for assessment purposes. Many of the features have already been shown to be related to essay quality in other studies. For these measures, the question is how they perform with low level learners. For example, this study suggests that care needs to be taken with measures of extrinsic lexical diversity. The Lexical Frequency Profile (LFP) (Laufer & Nation, 1995) has been shown to correlate with assessments of written quality both by the creators with Israeli learners of English and also by Goodfellow, Jones & Lamy (2002) with L2 learners of French. Both of these groups of learners were lower intermediate level. However, the results of the experiment in Chapter 5.5 with low level learners suggest the LFP might not be suitable for lower level learners on some tasks. This may also be the case with P_Lex (Meara & Bell, 2001). Meara & Bell found that P_Lex was effective for lower intermediate to advanced learners but there were also problems with P_Lex with the learners in the same Chapter 5.5 study. Advanced Guiraud proposed by Daller, van Hout & Treffers-Daller (2003) was found to discriminate between high beginner learners in Chapter Six but it did not perform so well as part of the clustering algorithm in the Chapter Seven experiment which involved many lower level learners. One possibility for this effect with extrinsic measures could be a problem with the task rather than the learners. It may be that the tasks did not facilitate the use of lower frequency vocabulary. In that case, tasks may need to be checked carefully to ensure they elicit vocabulary from a range of frequency levels.

A number of other features showed a positive correlation with essay quality, for example, TTR(100), the TTR of a 100-word sample. Some features promised an ability to recognize differences between learner essays and native essays but their ability to identify differences between learner essays of differing quality is yet unclear and needs more research. This is the case with the following features: relative distance between *a* and *the*, the number of hapax legomena and mean word length. As with extrinsic measures, it is quite likely that the performance of individual features may vary according to the learners, tasks and compatibility with other features.

9.2.1.2 Reliability across tasks

Reliability of a feature over different tasks is important if we hope to use that feature with various tasks rather than on a single specific task. For research purposes, task specific measurements are a viable alternative but for testing in education, a more versatile measure that can be used with a variety of tasks is preferable. A major criticism from teachers about automatic assessment packages in L1 situations is related to the inflexibility of task selection (Herrington & Moran, 2001). With those packages, task selection is, in effect, often taken away from the teacher and put into the hands of the testing organization.

When checking for reliability over different tasks, we assume that learners perform consistently well on different tasks. Therefore, a learner should produce essays of similar quality on the two tasks. This may not always be the case. There are many factors that can affect individual performance. We might expect some differences in individual performance across different tasks but most learners should perform consistently. Also, differences in absolute scores may occur across different tasks reflecting differences in difficulty or context but this should not compromise reliability which depends not on absolute scores but on scores being consistent relatively. In other words, scores may vary across tasks but ordering according to quality should be consistent.

Reliability of a feature is related to its occurrence. Features that occur few times tend to be unreliable. This helps explain the poor performance of the extrinsic measures of lexical diversity in some experiments. On certain tasks, some learners only produced a small number of advanced words. This was also a potential problem identified with some of the features used by Ferris (1994). Because many features in that experiment only occurred a small number of times, they were liable to great fluctuations over different tasks.

By contrast, features that rely on larger amounts of data tend to be more reliable. This was seen throughout this study with features such as essay length, word types and TTR which tended to be reliable across tasks and learners. However, the measure of distinctiveness despite incorporating large amounts of data showed problems with reliability.

This study also hinted that in some cases, reliability testing may be overly optimistic. Some features may be quite task specific and so testing reliability on different tasks may suggest the feature is unreliable. These features may be more reliable on very similar or identical tasks. In Chapter 3.3, distinctiveness was found to be quite task specific. It showed relatively high reliability when learners completed two identical tasks but much lower reliability levels with learners on different tasks. In many assessment situations, learners are responding to exactly the same prompt, so task specificity itself is not a problem. However, lack of reliability over tasks is likely to indicate the lack of a strong relationship with quality. It is worth noting that the measures such as essay length, word types and TTR tend to show reliability across all kind of tasks.

9.2.1.3 Automatic calculation

A major aim of this thesis is to investigate the issues involved in attempting to produce a simple computer application to assess essays. Therefore we need some simple measures that can easily be calculated by a computer program. Some features can be calculated directly and simply such as essay length or the number of word types. Others may not easily be calculated exactly but a pretty good estimate could easily be produced. An example is mean sentence length. It is not easy for a simple program to determine exactly where each sentence begins and ends but mean sentence length could be estimated by dividing the total number of words by the number of sentence ending punctuation marks. It is worth remembering, though, that there are occasions when this kind of approach can be seriously affected by stylistic choices. For example, there are uses of full stops as decimal points which could affect an estimate of mean sentence length.

Other features may prove more difficult to calculate. For example, the number of error-free T-units has been shown to be a very good indicator of proficiency (Larsen-Freeman & Strom, 1977). A T-unit let alone an error free T-unit cannot be easily calculated nor easily predicted automatically. It is obviously related to sentence length and clause length and yet its strength lies in its difference to these measures. If we look back at the features identified by Ferris (1994), we can divide them broadly into features that are easy to calculate directly, are possible to estimate and are more difficult to estimate. Some examples of these are shown in Table 9.1.

Table 9.1: Variables from Ferris (1994) in terms of their ease of calculation

easy to calculate	possible to estimate	difficult to estimate
number of words	words per sentence	deictic reference
mean word length	number of relative clauses	coherence features
1st/2nd person pronouns	negations	synonymy
repetition	comparatives	complementation

However, there seems plenty of scope for developing measures from features that are easy to calculate or estimate.

9.2.1.4 Validity

The validity of automatic assessment systems is largely based on reliable prediction of human ratings. Validity is not the most important concern for individual features but it is certainly a useful property. For automatic assessment, face validity can be particularly important given the reservations some people hold about using computers in assessment (Herrington & Moran, 2001). Even the e-rater system which produces fairly accurate results comes under criticism because its features are seen as overly relying on essay length.

Some features may exhibit more validity than others. For example, lexical error as a feature would seem to exhibit strong validity. It is widely recognized that more proficient learners are likely to make fewer errors while less proficient learners make more errors. The reality may be slightly more complex than this impression. Larsen-Freeman & Strom (1977) found that a linear progression of decreasing error does not always occur. They report that for errors involving articles, while high quality essays had the fewest errors as would be expected, poor quality essays actually contained fewer errors than medium quality essays. Nevertheless, the relationship between error and essay quality may be quite easy to appreciate and lends validity to using an error estimate in assessing learner essays. In other cases, the basis for validity may be less clear. In those cases, validity may depend on making certain assumptions or conditions. For example, uniqueness and distinctiveness could have validity if one argues that as learners become more proficient their vocabulary knowledge should increase. As they know more words, it follows that they will be more likely to use unique and more distinctive vocabulary than less proficient learners who know fewer words. Of course, the more realistic the assumption, the stronger the case for validity. As we saw in Chapters Three and Four, this argument for uniqueness and distinctiveness may not stand up to close scrutiny.

Validity may be applicable under certain conditions. For example, we might argue that the LFP is valid for lower intermediate learners and above because we expect them to have an adequate range of vocabulary that can be reliably identified at the different frequency levels. However, it may not be valid for lower level learners or in relation to tasks that do not elicit a suitable range of vocabulary. Typical usage of *a* and *the* might be argued as being a valid indicator for discriminating between L1 speakers and L2 learners of any level if one argues that these words may be two of the final words to be acquired completely and used in a native-like way for learners from some L1 backgrounds.

Of course, face validity does not ensure reliability. This can be seen in the case of mean word length. The case for validity is strong. Zipf (1932) showed that more frequent words tend to be shorter. This is reflected in mean word lengths in letters of 5.25, 6.57, and 7.29 for the headwords of the 1000, 2000 and University Word List frequency levels respectively. Therefore if more proficient learners are more likely to use more infrequent words, then we might expect mean word length to be higher for more proficient learners. However, only weak correlations with quality suggest that although the assumption may be a sound one, short L2 essays may not provide more proficient learners with sufficient opportunity to deploy enough lower frequency words to reliably show up in mean word length.

Validity considerations may become more complex when using features in combination. On one hand, validity should be enhanced if a variety of measures tap into different aspects of knowledge. On the other hand, there is a need for balance. E-rater has been criticized for relying on measures that all seemed to be affected by essay length.

9.2.2 Length of essay

It is worth considering length of essay as a special case. It has been shown to have a major correlation with essay quality and it also often affects the correlation of other variables with essay quality. Moreover, its role in both these effects is controversial. Various studies have found essay length to be an important indicator of quality of essays in many situations. In this study, essay length was shown to be both more reliable over two tasks than uniqueness and distinctiveness and also more able to identify high and low quality essays than these same two statistics. Ferris (1994) found essay length to be the most significant indicator of proficiency in a study of 160 learners from Arabic, Chinese, Japanese and Spanish L1 backgrounds. Larsen-Freeman & Strom (1977) showed that essay length broadly correlated with proficiency even in tasks with no time constraint. The PEG automatic essay scoring system uses a measure of essay length as one of a set of features. However, other studies have found no significant relationship between essay length and essay quality such as Perkins (1980) whose results showed an increase in essay length with an increase in proficiency but this increase was not significant. However, it may be relevant that the rubric in this experiment advised students that rating would be based on essay length as well as essay quality.

On balance, the evidence suggests that in many cases, essay length is a good indicator of essay quality. The degree of correspondence between essay length and essay quality may depend on various factors such as the level of the learners, the task context and the task conditions. I think a strong relationship is likely to hold for lower level learners on time constrained tasks. This was borne out by the results in Chapter Six and Seven. In the two-dimensional model in Chapter Six, essay length accounted for over 61% of the variance but the other dimension, lexical diversity only accounted for about 4%. In Chapter Seven, a classification by essay length alone seemed to be able to account for quality better than human raters. For many low level learners, a small vocabulary precludes them from writing a lot. They may also find the physical process of accessing vocabulary and getting words down on the page more of a challenge. Therefore, it seems quite likely that lower level learners will write shorter essays. Of course, this relationship will not be perfect. Some more proficient learners may opt to use more time editing and correcting what they have written while others might spend the extra time producing more text.

Many other features are also affected by essay length. Variables that involve counting the occurrence of a feature are likely to be affected directly. As the length of the essay increases, the count of the feature is also likely to increase. This often enhances the performance of the feature in predicting quality by itself in cases where length also correlates with quality. In these cases, correlation is improved but the relative contributions by the feature and essay length remain unclear. Therefore, where the features are included in a set of variables, it is more usual for the feature to be controlled for length.

Where the feature involves a ratio as in the type-token ratio, length effects may be more problematic. With many ratios, for example, the TTR, the feature has an inverse relationship with essay length. This inverse relationship will work against the ability of the feature to predict quality. As the length of essay increases, the value of the ratio is likely to decrease. In these cases, it is more important to control for essay length. However, essay length effects can be pervasive. A lot of research has gone into producing a lexical diversity statistic that is free of length effects. The most successful to date, is probably the D estimate developed by Malvern et al. (2004). However, there is some evidence that in certain cases, this measure is still not completely free of length effects (McCarthy & Jarvis, 2007).

In assessment situations, since essay length may be positively related to essay quality, it is important to control for length in any features where essay length corresponds negatively with that feature. In the case of features that correspond positively with essay length, it is probably preferable to control for essay length where an independent measure of essay length is included as well.

The effect of length on essay evaluation is a complex and problematic one. In many written tests, the length of the essay may give a clue to quality. This is especially true for lower level learners for whom writing a lot in a time-constrained task may be an indicator of fluency. However, this holds less true for higher proficiency learners and for testing situations with no time constraint. The experiments in this study have mostly involved low level learners and time constrained tasks. As a result, the influence of essay length has been an overwhelming factor in the assessment models. To find out more about other essay features that can help assess essays, it may be worth investigating testing situations where essay length is less influential. This could be done by examining responses to length controlled written tasks where essay length will not be an issue. The influence of essay length may also be lessened when using non-timed tasks or essays of higher proficiency learners.

It is also worth noting there tends to be a sensitivity to the correlation of essay length with essay quality in the wider assessment community. Testing organizations are often at pains to point out that raters reward quality rather than quantity. Length of essay is not usually explicitly referred to in scoring guidelines. However, this seems to contradict evidence such as Cumming, Kantor & Powers (2002) that essay raters actively consider essay length when rating essays. On one hand, this could be an argument for the frailty of human raters. Even though essay length is not included on the scoring rubric, it seems to influence their rating anyway. However, it could equally be seen as a weakness in many scoring rubrics that a basic element such as length is not included. Managing the essay length problem is certainly one of the most challenging issues ahead in automatic assessment.

9.3 L2 vocabulary research

Recently, Nation argues for using multi measures in vocabulary research (Nation, 2007). One simple reason for using multiple measures is that should one of them fail due to task design, student characteristics or other unforeseen problems, another could act as a backup measure. Another reason seems to be that using a variety of similar measures can help safeguard validity in research by identifying slightly different aspects of knowledge. With an array of complementary measures measuring slightly different aspects, a more accurate picture of learner vocabulary can be constructed. It is in the complementation of various measures where this study may assist vocabulary research.

In the area of lexical statistics of written productions, there has been something of a tradition of considering a range of measures. A major reason for this has been the lack of a standard measure of vocabulary and the obvious shortcomings in some measures that have often been used. A great deal of research has been involved in trying out new measures and looking for superior alternatives for established measures. This intensive search for better measures has lead to some significant developments. For example, the long recognized effect of essay length on lexical diversity has led to the development of more complex measures such as the D estimate (Malvern & Richards, 1997).

Perhaps as a result of this goal of a perfect measure of lexical diversity, most studies concentrated on simply finding the measures that were best in terms of reliability and correlations with written quality or learner proficiency. Consequently, there has been less emphasis in developing the complementary aspect of Nation's call. Little stress has been placed on exploring why lexical features give different results or on the similarities or differences in the performance of these features. In this respect, an important contribution is that of McCarthy & Jarvis (2007) who explored the performance of various measures of lexical diversity over differing essay lengths. They found that many of the features work well if used with texts of certain lengths and they specify the ranges of use for various measures of lexical diversity. This certainly diminishes the chances of researchers using inappropriate measures for a particular context and stresses the fact that most of these measures are useful given an appropriate context. Some of the experiments in this study have given clues about properties of measures of lexical diversity. In Chapter Six, strong correlations were found between the D estimate and Yule's K suggesting they may be measuring very similar constructs. Both of these measures correlated with other measures of lexical diversity less strongly.

There seems to have been little research on how lexical measures could complement each other practically. For complementation, there needs to be firstly, a more precise identification of how measures vary. In this respect, Meara and Bell's (2001) classification of intrinsic and extrinsic measures of lexical diversity has been a useful one. By making this distinction, it has enabled not only the similarities of these two types of measure to be fully appreciated but also their relative benefits and weaknesses explored. This, in turn, opens up the opportunity for their complementary use in the field. Once the complementary nature of measures becomes apparent, it is possible to use them together in a more meaningful way. Chapters Five, Six and Seven included findings that essay length and lexical diversity together can often account for essay quality better than either feature by itself. However, the relative contributions may vary according to task and level of learner. In the Chapter Six experiment, essay length was a dominant feature in a two-dimensional model incorporating essay length and lexical diversity. Similarly, in Chapter Seven, essay length was the dominant feature in the clustering algorithm. However, in an experiment in Chapter Five, lexical diversity seemed to account for quality better than essay length.

For vocabulary assessment, complementation seems an area of vital importance so the lack of focus is somewhat surprising. Vocabulary use has been shown to be a key indicator in assessment whether through lexical error, profiles of vocabulary frequency or demonstrations of sophisticated vocabulary usage. Many written tests include vocabulary profiles such as the ESL composition profile (Jacobs et al., 1981). These usually involve a holistic evaluation of vocabulary use even though more detailed descriptors may exist in the rating guidelines. However, a way to produce holistic measures of vocabulary use in essays has yet to be developed.

In vocabulary research in general, there is also a lack of a holistic way to assess vocabulary knowledge. Vocabulary research has identified various facets of vocabulary knowledge. The most common two-dimensional model is the vocabulary breadth and depth model. Vocabulary breadth or size has been extensively studied with measures like the Vocabulary Levels Test (Nation, 1990) or the Eurocentres Vocabulary Size Test (Meara & Jones, 1990) Depth of vocabulary has been measured by the Vocabulary Knowledge Scale (Wesche & Paribakht, 1996). Meara (1996) has advocated a slightly different way of thinking about vocabulary knowledge focusing on the nature of the lexicon as a whole rather than individual words and identified two possible dimensions, vocabulary size and organization. The organization dimension can be investigated through word association tests. There has been an increasing appreciation on the fluency/accessibility aspect of vocabulary use. Some include this as a third dimension as in Daller, Milton and Daller-Treffers (2007) who visualize a three-dimensional model of breadth, depth and fluency. A fluency/accessibility element also seems to be implicit in Meara's organization dimension. Although, there has been a lot of research on measuring each facet of vocabulary, there has been little effort in putting these measures together to give an aggregate vocabulary score to a candidate's performance or to an essay.

In vocabulary production, it is not always easy to gauge these different facets. In a spoken interview, a skilled interviewer may have the flexibility to explore these different facets in some detail. However, this is much more difficult in written tests. The lack of control in elicitation has left researchers trying to find increasingly ingenious ways to squeeze more information out of sparse productions of words. To make matters worse, a lot of the commonsense measures such as lexical diversity do not lend themselves clearly to any multi-faceted model. In addition, their relatively poor performance in predicting quality suggests they are inadequate holistic measures by themselves. In some cases, researchers have viewed gauging vocabulary knowledge from written productions as something of a dead-end and Meara & Olmos Alcoy (2008) argue for using non-traditional written production tests such as Lex30 (Meara & Fitzpatrick, 2000). One result of these problems with written work is that we are less confident in making assessments of the candidate as we might with spoken performance and more likely to make assessments restricted to the written sample in hand.

Practically bridging this gap between different facets of vocabulary knowledge is an area that is badly neglected. There is a plethora of tests and measures available but little research on how to use them together. One of the issues in complementation is how to incorporate various measures of different facets to produce a more holistic measure of vocabulary knowledge. In this sense, the model in Chapter Six was a crude two dimensional model which I call quantity and content. Although this does not fit in cleanly with a vocabulary breadth, depth and fluency model, quantity could be seen as a proxy measure for fluency in low level learner tasks especially on timed tasks. Accessibility of the lexicon is clearly a factor in how much learners write in a limited time task. Lexical diversity, whether intrinsic or extrinsic, or a vocabulary distribution measure like hapax legomena could be viewed as a proxy measure for a more core vocabulary breadth. Two dimensions are likely to give a lot more predictive power than a single measure. If we wanted to complete the trio, we could perhaps propose a proxy measure for vocabulary depth. Lexical error might be a possibility since lexical errors often occur in words the learner knows but uses incorrectly. Chapter Eight offered some alternatives for using more features together. In particular, supervised methods such as Bayesian analysis allow for control on the input variables and could be used with various kinds of model.

In this respect, I believe that there are many avenues for using measures in a more complementary way in both vocabulary research and vocabulary assessment. I believe this study contributes some simplistic ideas on how we can integrate different facets of vocabulary knowledge in order to come up with a more holistic evaluation.

9.4 Automating the process

A major aim of this study was to develop an algorithm that could automatically analyze essays input in electronic form. To that end, as far as possible, all analyses were done using specially designed computer programs (see Appendix 1.1). However, many of the essays were not initially input into the computers by the students themselves. Most experimental data was generated by timed essays performed during class time. It was not feasible for the students to input these directly into the computer in the classroom and so they were handwritten and input by myself at a later date. This introduced various opportunities for error to enter the data. I had to read the essays, interpret the handwriting and then input it into the computer. There was scope for errors at several stages. Sometimes words or punctuation were difficult to interpret, for example, distinguishing between commas and full stops caused many problems. On other occasions, typing errors may have been introduced by me where no such error existed in the original. Another task involved essays produced as homework by learners in their own time (Chapter 5.3) and these were input into computer form by the students themselves and delivered by email. When students input their own essays, input errors become just one part of overall student performance error. A simple test for the reliability of transcription might have improved the methodology. By introducing an additional human checker, unclear words or punctuation marks could have been independently assessed. This may have helped minimize any impact of transcription errors.

Using computers in a timed essay at this moment in time raises another concern. When writing an essay by hand in a timed format, we do not usually worry about the physical writing process affecting the overall performance. In fact, we may even view the physical process of producing text as one facet of language skill, and it may only become an issue if a learner is suffering from some sort of physical handicap or injury. Producing language on a computer keyboard seems to be a different case at this moment in time. Typing skill may vary between learners independently of language skill. In Japan, first year undergraduates have different experiences of using computers, especially when using English as input. This may change at some point in the not too distant future, and it is probable that typing essays in class time will become more of a required skill. Where learners use their own time to input essays, computer literacy seems less of a concern. However, it may still have some effect. For example, if learners who are poor at typing have to spend relatively more time inputting text, it may make them more likely to write less.

All the programs used in this analysis were specially-designed except where acknowledged (see Appendix 1.1). Programs to handle essay files and do simple analyses are relatively easy to construct. The analyses in experiments from Chapter Three to Six were relatively straightforward to do on the computer. The two automatic algorithms in Chapter Seven were somewhat more complex. They both involved bringing together several analyses and there were a number of stages that needed human input in these processes. One stage was in the calculation of an estimate of lexical error. This estimate was derived by eliminating legitimate words by checking against a list of the most common words in English, a list of proper nouns, and a user list. Words which were not eliminated by these lists were checked by humans to further eliminate any more infrequent words and acceptable variants. Words eliminated at this stage were added to the user list of additional acceptable words or proper nouns so that over time the number of words needing to be checked would decrease. The words remaining after this check were deemed to be errors and a count of these words was used as a lexical error estimate. Whether this estimate of error is an important enough element to warrant this human intervention is something that needs more investigation.

In the clustering algorithm in Chapter Seven, there were another two stages that required human input. The first was part of the PCA analysis. PCA depends on a complex matrix transformation called a singular value decomposition (SVD). This transformation is simple for very small matrices but soon becomes very complex for larger matrices. In this experiment, it was necessary to find the SVD of an 11 by 11 matrix. This stage could be automated but was beyond my computer programming ability. Therefore, I had to bridge this stage by using some external specialized software for calculating an SVD for a large matrix (Bluebit, 2003).

Another occasion where human input was needed was in the comparison of PCA output variables and the input features. The input features were compared with the PCs produced by the PCA to find input features that approximate these PCs. This stage of the analysis was also done by human hand by studying graphs of the original features plotted against the PCs (Appendix 7.1) and seeing which feature had the most linear relationship with each PC. This could possibly be automated by using a regression model but would make the procedure considerably more complex and I am not convinced that a program could make the decision as well as a human. This may lead to a loss of performance in this algorithm if fully automated.

Comparing the two algorithms in Chapter Seven, the Bayesian algorithm is more automated than the clustering one. In fact, if the lexical estimate was taken out, the Bayesian analysis would be fully automated. This perhaps slightly balances out its inferior performance in classifying the essays in comparison with the clustering algorithm.

9.5 Limitations of the present study

There are quite a few limitations of the study. Some have been hinted at during the course of the experiments. I will address a few of the main limitations. Firstly, I look at the danger of reading too much into the results of small scale experiments. Secondly, I look at limitations regarding the tasks and subjects in the experiments. Finally, I describe problems with the design of the rating structure in the classification experiments in Chapter Seven and consider how spelling errors could be handled more effectively.

The experiments in this study were mostly of an exploratory nature. An automated scoring system needs some way of assigning scores to essays. The experiments in Chapter Five were designed to explore whether certain variables held promise for this. They involved only a small number of learners. Given the limited nature of these studies, it is unwise to read too much into the results. Some of the results may not be true generally but only for the group in that particular study.

This study has been aimed at low level Japanese undergraduates all sharing the same L1 background. Therefore, it would be unwise to extrapolate results to other proficiencies of learners or learners from other L1 backgrounds. It is quite likely that features that help identify proficiency in Japanese learners may not work for other L1 speakers. One example might be the patterns of usage of the items *a* and *the* analyzed in Chapter 5.3 which seemed to vary clearly between Japanese learners and native speakers. One reason for this may be the lack of such articles in the Japanese L1. Learners with other L1s may well vary considerably in patterns of usage of these articles. Further research on all aspects of automation may be necessary before using with learners from other L1 backgrounds.

Another potential weakness is in the tasks. There were several reasons for the initial choice of the cartoon tasks (Appendix 3.1 and 3.2). One was to provide an appealing task that was simple to understand for learners of varying proficiency. Most learners find the cartoons amusing and are keen to write a story about them. Also, the tasks are pure language tasks as they do not involve any content input from the learners. Another reason for choosing this type of task was for task standardization. If more researchers use standardized tasks, results from different studies can be more easily compared. This type of cartoon task has been used in several other studies such as Meara & Olmos Alcoy (2008) and Daller, van Hout & Treffers-Daller (2003).

There was one possible weakness of the cartoon tasks that emerged during the course of the experiment. There is some evidence that they are restrictive in the range of vocabulary they elicit. The tasks tend to produce very similar answers which include a lot of the same vocabulary. Much of this vocabulary is fixed by the context. This consideration may also impact on some measures like uniqueness. Using uniqueness to measure proficiency rests on an assumption that more proficient learners will use vocabulary that less proficient learners do not. However, a task requiring precise vocabulary does not necessarily engender uniqueness. Correct choice of vocabulary is important but inappropriate use of vocabulary may inadvertently boost uniqueness scores. The vocabulary may also be restricted by frequency band. In particular, there is little evidence of vocabulary being used from lower frequency bands. This may have accounted for the extremely low values of LFP and P_Lex in Chapter 5.5. Learners completed these tasks using mostly words from within the 1000 most frequent words of English. It is not clear whether this usage reflected the learners' limited vocabulary or whether the task could be completed satisfactorily using low level vocabulary. However, this same kind of task was used with considerable success to elicit spoken output in German and Turkish in Daller, van Hout & Treffers-Daller (2003).

It is also worthwhile considering the subjects in the experiments. They were on the whole low level learners studying at Japanese university. Some were English majors and some were non-language majors. Some of the English majors were first-year students and some were third-year students. First-year students usually exhibit a range of abilities reflecting their different backgrounds before university. English majors usually show particular improvement particularly on productive tasks by the third year of their study. Standardized language tests scores are usually not available for these students but those that do take the TOEIC test usually score between 250 to 500 for first year students and 350 to 700 for third year students. When investigating assessment involving finding quality differences, it is useful to have essays reflecting differing proficiencies. I sometimes felt that some of the essay sets composed of only first year essays, while displaying some noticeable quality differences, had ranges that were often not very wide, with many essays of similar quality. This made rating difficult at times. When rating, it was relatively easy to isolate good essays and poor essays but more difficult to discriminate between a large number of medium quality essays. All raters involved with rating the essay sets containing only first year learners reported similar problems.

Essay sets involving first year essays and third year essays offered a chance to collect essays from learners with a wider range of proficiencies. In addition, one study compared essays from native speakers and learners. In this case, although there was less control on the essay task, there was a clear difference in proficiency that could easily be investigated without recourse to human rating. An obvious concern in using comparisons between learners and natives is whether results can be interpreted as being relevant to inter-learner comparisons.

Another weakness is in the experimental design used in Chapter Seven. A test format was chosen that reflected real life placement situations which are often about fitting bodies into classes rather than finding more meaningful strata of proficiencies. It was also a simpler rating procedure. From a rating perspective, it takes less rater training to discriminate between the best fifty candidates and the rest than to split them into proficiency bands which necessitates training to find consistency of rating between judges. A downside of this is that rater reliability may be affected considerably by borderline cases. Judges are likely to agree on very strong and very weak candidates but more likely to disagree about average ones. The selection of only two groups for placement was, on reflection, a poor choice. If we assume that essay quality follows a normal distribution, then splitting into two groups splits the data at the point where there are a maximum number of learners at the borderline level. This is likely to undermine inter-rater reliability. As an illustration consider a set of 100 learners who belong to three proficiency levels, thirty of high proficiency, thirty of low proficiency with the remaining forty of middling proficiency. Imagine our expert raters can perfectly recognize the thirty high level and thirty low level learners but cannot distinguish the forty middling learners so they assign twenty to each group. Even if the reliability of the sixty high and low level learners is 100%, the random assignment of the others would lead to an expected agreement of just 80% for the whole 100 learners. On average, the raters would agree on half of the twenty learners to be allocated to each group leaving forty out of the fifty in each group agreed on. If we adjust this agreement measure for agreement by chance, we get a Kappa reliability estimate of 0.6. If learners were split into three groups, this would probably more naturally fit the spread of quality if distributed normally.

The handling of spelling errors warrants further consideration. Spelling errors can be a useful resource for automatic assessment but may also be a problem to deal with. They are useful because they are one relatively easily identifiable form of error. Error is likely to influence human raters as found by Engber (1995). Occurrence of spelling errors may be an indicator of error more generally. However, spelling errors can also be problematic. Automatic packages are often ambiguous about how spelling errors should be handled. In fact, whether errors are corrected or not, they may cause statistics to be unreliable. Correction may reward the learner to the same degree as if the word had been properly spelled. On the other hand, non-correction could lead to the learner not getting credit for partial word knowledge, for example, when comparing words to word lists. In other cases, such as the calculation of distinctiveness, non-correction may mean that the learner gets credit for using a rare word when the learner may actually have misspelt a common word. The P__Lex package handles spelling errors in the user interface but this could become cumbersome if used with many essays containing multiple errors. One way to ease this burden is to keep a record of errors in the program and so enable less interface time with subsequent use of the program. One way to approach the correction debate might be to correct spelling errors to reward even partial knowledge of a word but also record them so that an error statistic becomes part of the assessment system. Spelling errors could be corrected by using a spellchecker module as part of the assessment package.

9.6 Future research

I believe this thesis has only scratched the surface of the automatic assessment question. There is plenty more research to be done in all aspects of the study. I will focus on the choice of system and the choice of variables.

9.6.1 Choice of system

In Chapter Seven, two approaches were evaluated, a clustering and a Bayesian algorithm. However, there seem to be countless methods in the fields of classification and pattern recognition that could be adapted to this problem. One way of looking at possible methods is to split them into supervised methods and unsupervised methods. Supervised methods involve presenting a system with guidelines on how to do a classification. Supervised methods usually involve some kind of training set. Bayesian classification is a supervised method since it depends on sets of training data that the classifier uses to predict the quality of each of the other essays. Cluster analysis is an unsupervised method since it does not use a training set. However, it is not truly unsupervised in this study since initial clustering points were input. A true unsupervised method would find optimal clusters itself as in the study by Jarvis et al. (2003). It might be interesting to compare the supervised version of clustering used in this study to a truly unsupervised method to see if initial clustering points are actually necessary. There are various other classification and pattern recognition algorithms of both supervised and unsupervised types that could be used with this kind of work. One possible system is a neural network. Meara, Rodgers & Jacobs (2000) proposed using a neural network system in conjunction with lexical signatures to classify essays according to quality. They found that a neural network system could be trained to distinguish essays by quality using lexical signatures. Unfortunately, they did not continue by taking the system to the crucial stage of using the neural network to predict quality of essays outside the training set. One drawback with neural networks is that they usually require quite extensive training sets if they are to work accurately. Nevertheless, they are one possible area of research.

Not only are there other methods that could be used but there are also other ways of implementing these methods. Even continuing with cluster analysis or Bayesian analysis, there are various ways these can be developed. In the case of clustering there are various clustering algorithms and ways of measuring distance. There is also a more basic way in which the two methods varied. The clustering algorithm was based on lexical features whereas the Bayesian classifier depended on semantic content. It would be possible to design a clustering algorithm using a semantic approach and a Bayesian algorithm with the lexical features. An alternative would be to use a combination of both lexical statistical information and semantic information. In fact, this is the approach that would seem to make most use of the data.

These kinds of methods often work best after a period of trial and error helps establish the ideal specifications. As one uses them more and more, one learns how to get the best performance out of them. This was very evident with the work using Bayesian classifiers. I encountered various problems trying to get output in a form that I could use. One particular problem was that the Bayesian probabilities of belonging to a particular group often came out as extreme values of 0 or 1. This was problematic because I wanted to progressively allocate the essays to a quality group. I eventually solved this problem and several others with the help of some data transformations in Rennie et al (2003) (see Appendix 7.4 for more details of these transformations).

A further area of research is to take automatic rating away from regarding the human judge as a gold standard of rating. Most automatic scoring methods depend on comparison with expert judges. However, this approach can be problematic. As has been noted, human rater scores are often very unreliable and so systems that mimic human raters depend on a flawed standard. By using multiple judges, the reliability of humans can be improved but other problems remain. One problem with judges is reflected in the high correlation of human ratings with essay length even though length is rarely mentioned in any scoring rubric. Up to now, the main role of automatic assessment has been in terms of convenience. Automatic assessment can be quicker, cheaper and more accurate than small numbers of human raters. However, the next challenge for automatic assessment could be to improve the rating process itself. One appeal of automatic scoring techniques is that they offer the opportunity to move beyond simply predicting the often-flawed impressions of human judges to recognizing elements of writing that can account for quality of writing and can be applied in a more consistent and reliable manner.

9.6.2 Selection of variables

Another area that needs to be explored in more detail is the features that can be used in these models. In this thesis, I was largely influenced in my choice of feature by the fields of authorship attribution and vocabulary acquisition. Many statistics developed in the field of author attribution are useful but were not actually developed with assessment in mind so may not be the most appropriate measures. The appeal of this field is that a lot more research has involved the use of complex statistical methods than in the field of vocabulary acquisition and applied linguistics in general.

I hope that these experiments have suggested that the use of a selection of features is often better than a single approach. Single features like the length of an essay or lexical diversity may be valuable and they may show a correlation with quality but if they do this correlation is likely to be limited reflecting the complex judgments behind quality rating. It is unlikely that one measure can encapsulate all the nuances of quality.

Experiment 5.4 hinted that strong correlations may not always be vital between features and essay quality. A standard correlation coefficient measures the strength of a linear relationship between features and quality. However, by being open to other forms of relationship, we may be able to identify features that can also help without exhibiting the strong linear relationship. One such case was highlighted in this study. Lexical variation and TTR were shown to have a weak overall relationship with essay quality but strong in some domains. A high score or a low score for lexical variation or TTR was very likely to be associated with high or low quality respectively. However values not so extreme were less likely to be predicted in terms of quality. This is a simple observation but one that makes these extreme values potentially very useful. Using these extreme values, it may be possible to reliably identify essays of high and low quality using a sampled TTR measure such as TTR(100) as was attempted in the Bayesian algorithm in Chapter Seven. This is comparable to a skilled human rater who can often flick through a sheaf of essays and pull out ones that can be confidently rated as high or low quality.

When an array of features is used, then a decision has to be made as to which features to include. One consideration may be a trade-off between accuracy and validity. To find a method that produces the most accurate results, it may be best to let a model decide itself what features to choose. The model will do this by finding the features that produce the best statistical evidence. The problem with this is that.then the features may not appear to represent a balanced model of quality. This will affect validity. This was the case with Sheehan's (2001) criticism of the model used in e-rater. By allowing a computer program to select the variables according to a best fit approach, the result was that all features seemed like proxy measures for essay length. Page (1994) describes an alternative approach for Project Essay Grade where variables are selected beforehand which reflect a set of underlying qualities of writing. However, if features are selected with validity in mind, they may well not perform so well statistically.

What seems clear from this research is that there are many systems and features that could be useful in identifying quality of learner essays. One major challenge is how to incorporate these systems features into an automatic assessment system. Given the almost unlimited permutations of factors influencing test design such as testing purpose, number of learners, proficiency level, type of decision to be made etc, it seems quite unlikely that one single program will be sufficient. Perhaps a suite of programs that could be selected according to the parameters of each testing situation might be more appropriate. This suite could include some of the algorithms and features explored in this study but might also be augmented by some already established packages such as VOCD (Malvern et al., 2004), P_Lex (Meara & Bell, 2001), RANGE (Heatley, Nation & Coxhead, 2002).

9.7 Conclusion

This discussion chapter has confirmed that through this study, not only has progress been made toward a flexible automatic assessment package but a great deal has been learned along the way about all parts of the assessment process.

It is worth remembering that although the features and assessment systems utilized in this study aim to predict essay quality, they are all quantitative in nature. These features and systems do not involve direct assessment of essay quality but rather predict quality ratings of humans by quantitative means. The quantitative elements considered involve counts, comparisons of occurrences or more complex numerical analyses. It is also worth remembering that high correlations of features with quality do not mean that the quality is engendered by that particular feature. Essay length has been shown to correlate highly with quality determined by human raters in a number of experiments in this study. However, this does not mean that essays were rated as high quality because they were long. Essays that were rated high quality often happened to be long. Especially in timed essay formats, essays written by high proficiency learners are likely to be high quality and long while those written by low proficiency learners are likely to be low quality and short.

A lot has been learned about the types of features that are best suited to include in an assessment model. In particular, a number of potential problems have been identified. When working with low level learners, it is important to use features that are sensitive enough to register differences. With external criteria, we need to ensure that the criteria are calibrated to reflect differences in the learners. If not, poor measurement is likely to add to reliability concerns. Reliability is also affected by the occurrence of features. Low frequency features are intrinsically less reliable. Where features involve counts, it is important to make the range of the counted class as broad as possible.

Although many features show a relationship with essay quality, no one feature is likely to account for quality by itself. Complementary use of features is important. In this respect, this study can add to general research in L2 vocabulary use. For example, in both Chapter Six and Chapter Seven, the complementary relationship of essay length and lexical diversity was highlighted. In Chapter Six, essay length and lexical diversity together seemed to account for quality better than either feature by itself. In the clustering algorithm in Chapter Seven, essay length and lexical diversity formed two of the major dimensions on which the clustering was conducted.

Despite obvious limitations with the scale of some of the studies, evidence about a range of features has been collected in this study. Their relative merits and limitations have been compared and their performance tracked on a variety of tasks. Some of the models used for essay assessment may also have applications in modeling vocabulary knowledge and use.

Both feature selection and system design depend on the automation process. Features need to be calculated automatically which in some cases means making do with estimates or finding proxy measures. For example, simple estimates of error and advanced types have been used. Lessons have also been learned about automating the process into a simple program or suite of programs.

Despite some key limitations in this study, there is still plenty of scope for further research. Some problems have been identified with some of the tasks and with experimental design. However, there is almost limitless choice for further research in terms of a huge range of features and statistical techniques that could be applied.

Chapter Ten: Conclusion

10.1 Introduction

In this study, a case has been made for the automatic assessment of written productions of L2 learners. Firstly, some lexical features that can assist in identifying quality in L2 essays have been identified. Secondly, some simple systems to generate automatic assessments using some of these features have been deployed. The first three experimental chapters concentrated on features and the final three chapters on systems. In this conclusion, the major findings of the study are briefly summarised.

10.2 Uniqueness index

The first experimental chapter looked at the case for using lexical signatures to assess essays. This was based on Meara Jacobs & Rodgers (2002) finding that L2 learners showed a high degree of unique lexical choice in their essays. They wondered whether a uniqueness index based on lexical signatures could be harnessed to help aid assessment. In the first chapter, a simple uniqueness index was constructed and its reliability checked over similar tasks written by the same learners. Although, their finding of considerable individual lexical choice in learners was confirmed, this uniqueness index was found to have certain problems which may hamper its application for assessment purposes.

One problem is an unstable property of uniqueness which causes sudden changes in score with only slight changes in essays. Another problem is the complexity of the calculation which includes sampling. This complexity was found to detract from reliable measurement of the statistic. A third problem is that uniqueness is not linear in its relationship to quality. Both high quality and low quality essays may score highly on an index of uniqueness. A final problem is that lexical signatures utilize only a very small number of words in the essay. In these experiments, this number was sometimes as low as eight words but usually about 15-20. This may affect validity as well as performance. All these problems together probably contributed to the result that a uniqueness index score could not prove reliable over tasks by the same learners.

There may still be some hope for this kind of measure and there are various avenues for future research. One would be to control reliability testing more strictly by testing over two productions of the same task. With distinctiveness, this was found to improve reliability considerably. Uniqueness is likely to be as task specific as distinctiveness. More control on the words used in the analysis may also help with reliability. A further avenue of research is to find a strategy to handle the non-linear aspect of uniqueness. Both high quality and low quality essays may score highly on uniqueness. Perhaps other features could be used alongside uniqueness as indicators as to whether scores are at the high end or low end of the spectrum. Finally, the most interesting avenue might be to explore the co-occurrence properties of lexical signatures in more detail. For example, the kind of words that work well together could warrant attention. This might also help in identifying effective words for the analysis.

10.3 Distinctiveness

In the second experimental chapter, a more robust concept of distinctiveness was investigated as an alternative approach to the uniqueness based approach of lexical signatures. Distinctiveness offers a solution to some of the problems found with lexical signatures and the concept of uniqueness. Firstly, distinctiveness has much better mathematical properties which make it more stable. Secondly, it is a much easier statistic to calculate and does not involve any sampling concerns. Although, like uniqueness, it does not have a linear relationship with quality, there may be a way to solve the problem. It seems that high quality essays may be differentiated from low quality essays by appealing to a distinctiveness profile. Finally, distinctiveness depends on all the words in an essay which seems to make it a more stable and valid indicator.

Despite these technical improvements, distinctiveness did not show much more reliability over different tasks than the index of uniqueness. However, distinctiveness proved to be relatively reliable over identical tasks which suggests it is very task specific. Despite introducing more stability by using all the words in an essay, distinctiveness was prey to another weakness. Because unique words and relatively rare words earn maximum reward in a distinctiveness score, a high proportion of the score depends on rarely occurring words. Rare occurrences are by nature less reliable. For example, spelling errors score high credit intended for unique sophisticated vocabulary. Even controlling for spelling errors by eliminating unique words did not seem to improve reliability. In fact, elimination of errors led to lower reliability. In addition, it was not easy to see how the discrimination of high quality and poor quality distinctiveness profiles could be achieved in practice. This is an aspect that could be developed.

Although both the lexical signature approach in Chapter Three and the distinctiveness statistic in Chapter Four had considerable technical problems, my feeling is the underlying problem is that a unique/distinct dimension does not sufficiently synchronize with quality. An array of factors impacting on a unique/distinct dimension do not fit cleanly into a distinct is better argument. For example, distinct lexical choices may sometimes be a result of use of sophisticated vocabulary not available to learners with less knowledge. However, on other occasions, distinct choices may be the result of learners using error variants of less distinct words. Other considerations in lexical choice, such as the appropriateness or precision in meaning, may not align well on this dimension. Lexical choice seems rather too complicated for a unique or distinct based dimension to give an adequate estimate of quality. My feeling is that if some of the technical problems can be overcome, then both of these measures may find a minor role in assessment.

10.4 Other features

The next series of experiments involved considering lexical features that could be used to help assess L2 essays as part of a battery of measures. In these studies, some established measures such as LFP, P_Lex and various measures of lexical diversity were investigated. In particular, their reliability and ability to predict quality in learner essays were considered. In addition to these established measures, some features that have been used in authorship attribution but have not been picked up widely in L2 vocabulary studies were also considered. These included word distributional features such as the proportion of hapax legomena or mean word length. Statistics such as Yule's K and entropy widely used in authorship attribution but which are commonly thought to be inapplicable to L2 essays were also included. The results were highly speculative but gave some ideas to take into the multi-feature assessment models in Chapter Six and Seven.

For example, mean word length and mean sentence length seemed to show some correlation with assessments of learner essays. Yule's K and entropy also correlated with differences in essay quality. The results were not as strong as Malvern & Richards' D estimate but they showed a lot more stability than might be expected for such short essays. Given that the D estimate is difficult to incorporate into assessment systems because of its complexity of calculation, these may be considered as alternatives. The number of hapax legomena was shown to discriminate between learner essays and native written samples. Measures based on Zipf word ranks also showed potential for discriminating between native and learner essays. In another experiment, lexical variation and TTR were shown to be surprisingly robust predictors of proficiency on some basic learner essays. Problems with essay length were minimized because all essays were quite short. In addition, lexical variation seemed to be a better indicator of poor quality essays than TTR. A final experiment suggested that LFP and P_Lex need to be used with care with low level learners. Because they both assume that essays contain vocabulary from a range of frequency levels, there are dangers with very low level learners and tasks which only produce vocabulary from the first 1000 words.

10.5 Essay length and lexical diversity

The final three experimental chapters looked at systems rather than features. Chapter Six looked at a simple model of essay assessment using two dimensions of quantity and content. In this model, quantity was measured by essay length and content by a measure of lexical diversity. Various measures of lexical diversity were considered and compared. In the analysis, quantity was given an equal weight to content by standardizing the values on each dimension. The lexical diversity measures included were TTR(100), Guiraud Index, Yule's K, Hapax(100), Malvern & Richards' D estimate and an estimate of Advanced Guiraud. Graphs suggested that quantity was the dominant dimension with this set of essays in terms of predicting essay quality. Of the content measures, Guiraud Index and Advanced Guiraud seemed to perform best in combination with quantity in isolating some very good essays and very poor essays at the two ends of the spectrum.

Regression analysis confirmed essay length as the dominant feature accounting for over 61% of the variance. The lexical diversity measures only accounted for a further 4% with each measure performing a similar contribution. Guiraud Index and Advanced Guiraud both also showed high correlations with essay quality but they both also showed a strong correlation with essay length. This suggested the association with essay length may be accounting for the association with essay quality. There was a very high correlation between the D estimate and Yule's K which suggests that these may measure a similar aspect of essays. After controlling for essay length, TTR(100) and Advanced Guiraud showed the strongest partial correlation with essay quality. Whereas, Guiraud Index lost most of its association with essay quality once essay length was factored out, Advanced Guiraud maintained a high correlation with quality which confirms it as a potentially useful extrinsic measure of lexical diversity.

An important finding of this experiment is that although correlation analyses show that lexical diversity is not an overall accurate predictor of quality, the essays that score at the extremes on both dimensions, i.e. long essays with high lexical diversity or short essays with low lexical diversity, are very accurate predictors of high quality and low quality essays respectively.

10.6 Clustering algorithm

Chapter Seven looked at two rather more complex systems for assessing essays automatically. These systems were more complex in that they included various multivariate techniques plus they involved different kinds of data in the analysis. In the first of two experiments, a clustering algorithm was used to classify one hundred essays as relatively good quality or relatively poor quality. These automatic assessments were compared with human holistic ratings. The clustering algorithm involved finding a small number of features that best accounted for variance in the essays. These features were found by means of a principal component analysis. This small number of features was then used to cluster the essays according to relative proximity. The features were weighted according to the amount of variance they accounted for in the data. The six features were found to be in descending order of influence: essay length, TTR(100), mean sentence length, estimate of lexical error, mean word length and Hapax(100). The results of the clustering algorithm agreed with two human raters with an agreement reliability of 78% and 76% per rater or a Kappa adjusted for chance agreement of r = 0.56 and r = 0.52 respectively. This was higher than the agreement between the two human raters of 72% (r = 0.44).

The advantage of this method is that different features can be included in the model depending on the essays. One problem with this method is that the algorithm is very complicated and parts such as matching PCs with features is difficult to do automatically. Other aspects of the clustering algorithm such as the clustering procedure, may warrant further research. In this case, clustering was controlled by setting initial cluster points. Further research might be directed at the patterns unsupervised clustering would produce.

10.7 Bayesian algorithm

In the second of the two experiments in Chapter Seven, a Bayesian algorithm was used. This classified essays according to proximity to a training set of high quality essays. This proximity measure was based on semantic content and depended on occurrence of individual words in the essays. The training set essays were found from within the set of essays and predicted to be high quality. The identification of training essays was inspired by the finding of Chapter Six that high scoring essays on both essay length and lexical diversity tend to be reliably high quality. Accordingly, sixteen essays scoring high on essay length, two forms of lexical diversity, sample word types and sample hapax legomena, but low on error were identified as high quality training samples. A further 34 essays that were deemed most similar to these training set essays by a Bayesian classifier were added to this group. The remaining essays formed a lower quality group.

This Bayesian algorithm agreed with human ratings with an agreement rating of 70% or a Kappa statistic of 0.40. This was slightly less than the agreement between two raters. The Bayesian algorithm can be split into two parts. Some essays are allocated by the training set selection and some by the Bayesian classifier. Results suggested that the training set selection portion was more reliable that the Bayesian classifier portion. There may be too few essays included in the training set for the Bayesian classifier to operate properly. This algorithm predicted high quality essays better than low quality essays. This may be as a result of both the training set selection and the Bayesian classification focusing on high quality essays.

More research might focus on each part of the Bayesian algorithm process. For the training sample selection, the exact feature conditions that can identify exemplar training sample essays is important. For the Bayesian classifier part, the number of training essays required is unclear. Also Bayesian classifiers based on semantic information and feature information could be compared.

Since the two algorithms were used on the same essays, it was a good chance to check relative performance of the algorithms or inter-algorithm reliability. The two systems were found to agree with an agreement rate of 78% or Kappa of $r = 0.56$. The influence of essay length on the algorithms was also checked. The clustering algorithm seemed to perform better than a model based on essay length alone in terms of agreements with human raters. However, essay length performed better than the Bayesian algorithm and also correlated better with raters than raters did with each other. Both algorithms had a very high correlation with essay length which suggests they may be overdependent on essay length.

The results of these experiments suggest that the clustering algorithm is superior in terms of performance. However, the Bayesian algorithm was superior in the sense that it is a completely automatic analysis that requires no human intervention whereas the clustering algorithm still contains a few steps that are not fully automated. Complete automation in this clustering algorithm may be accompanied by a slight loss of reliability. However, I feel that these basic algorithms are just a taste of what is possible if statistical techniques are combined with different types of essay information such as lexical features and semantic content. For example, running these two different analyses and combining results may improve reliability with human ratings further. The two techniques employed were just two of many others and applied in quite a crude manner. There are many more sophisticated methods that could be used.

10.8 Capture-recapture analysis

Chapter Eight investigated capture recapture as a candidate for essay assessment. This method is used in biology to estimate wild animal populations. It has been adapted to estimation of learner vocabulary size by considering samples of written work. It might also be used to assess essay quality in some situations if it is assumed that an essay that appears to be written by a learner with a large vocabulary is likely to also be a good essay. Despite various technical problems, this method was shown by means of a computer simulation to produce relatively reliable scores at written sample sizes of 120 words and up. The weakness of this method for assessment purposes is that a number of written samples are necessary. This may make it difficult to use for assessment of traditional one-off written tests but make it an excellent alternative for continual assessment situations or where multiple writing samples are available.

10.9 Conclusion

This study highlights that automatic assessment is a realistic aim but there is no one feature or method that can account for essay quality. Systems that incorporate a variety of features may be more reliable. Moreover, essays vary according to learners, tasks, and test conditions. For each particular context, a different combination of features and system may be effective. Therefore systems that incorporate a range of features and methodologies may stand the best chance of achieving reliable prediction of essay quality.

This study only represents a tiny fraction of the features and techniques that could be used and I believe there is a lot more research that can be done. Above all, I believe these experiments demonstrate that the key to reliable automatic essay scoring is the accumulation of information including lexical features and semantic content and thorough analysis using a variety of statistical techniques.

Bibliography

Arnaud, P.J.L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein-Braley, & D.K. Stevenson (eds), *Practices and problems in language testing* (pp. 14-28). Essex: University of Essex.

Borland software corporation (1983). *Borland Delphi Personal version 6* (computer program).

Bluebit (2003). Online matrix calculator. Retrieved from *http://www.bluebit.gr/matrix-calculator/.*

Burnham, K.P., & Overton, W.S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika, 65* (3), 625-33.

Chodorow, M., & Burstein, J. (2004). Beyond essay length: Evaluating e-rater's performance on TOEFL essays. *TOEFL Research Reports, 73.*

Crossley, S., Salsbury, T., McCarthy, P., & McNamara, D. (2008). Using Latent Semantic Analysis to explore second language lexical development. In D. Wilson, & G. Sutcliffe (eds), *Proceedings of the Twenty-first International Florida Artificial Intelligence Research Society Conference* (pp.136-141). Menlo Park, CA: The AAAI Press.

Cumming, A., Kantor, R., & Powers, D.E. (2002). Decision making while rating ESL/EFL writing tasks: A descriptive framework. *The Modern Language Journal, 86*, 67-96.

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge.* Cambridge: Cambridge University Press.

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics 24* (2), 197-222.

Engber, C.A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing 4* (2), 139-155.

Evola, J., Mamer, E., & Lentz, B. (1980). Discrete point versus global scoring for cohesive devices. In J.W. Oller & K. Perkins (eds), *Research in language testing* (pp. 177-181). Rowley: Newbury House.

Ferris, D.R. (1994). Lexical and syntactic features of ESL writing by students at different levels of L2 proficiency. *TESOL Quarterly, 28* (2), 414-421.

Gaies, S.J. (1980). T-unit analysis in second language research: Applications, problems and limitations. *TESOL Quarterly, 14*, 53-60.

Goodfellow, R., Lamy, N., & Jones, G. (2002). Assessing learners' writing using lexical frequency. *Recall, 14* (1), 133-145.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique.* Dordrecht: D. Reidel.

Hatch, E., & Lazaraton, A. (1991). *The research manual: Design and statistics for applied linguists.* New York: Newbury House.

Heatley, A., Nation, I.S.P., & Coxhead, A. (2002). RANGE programs. Retrieved from *http://www.vuw.ac.nz/lals/staff/Paul_Nation.*

Herrington, A., & Moran, C. (2001). What happens when machines read our students' writing? *College English, 63* (4), 480-499.

Holmes, D. (1994). Authorship attribution. *Computers and the Humanities, 28*, 87-106.

Hunt, K.W. (1965). Grammatical structures written at three grade levels. *Research Report No. 3.* Urbana, Illinois: National Council of Teachers of English.

Ishikawa, S., Uemura, T., Kaneda, M., Shimizu, S., Sugimori, N., & Tono, Y. (2003). *JACET8000: JACET list of 8000 basic words.* Tokyo: JACET.

Jacobs, H.L., Zingraf, S.A., Wormuth, D.R., Hartfiel, V.F., & Hughey, J.B. (1981). *Testing ESL composition: a practical approach.* Rowley, MA: Newbury House.

Jarvis, S., Grant, L., Bikowski, D., & Ferris, D. (2003). Exploring multiple profiles of highly rated learner compositions. *Journal of Second Language Writing 12*, 377-403.

Jolliffe, I.T. (2002). *Principal component analysis.* New York: Springer Verlag.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. *Psychological Review, 104*, 211-240.

Landauer, T., Laham, D., & Foltz, P. (2003). Automated Scoring and Annotation of Essays with the Intelligent Essay Assessor. In M. Shermis, & J. Burstein (eds) *Automated Essay Scoring: A cross disciplinary approach* (pp. 87-112). New York: Routledge.

Larkey, L.S. (1998). Automatic essay grading using text categorization techniques. *Proceedings of the 21st Annual International SIGIR Conference on Research and Development in Information Retrieval*, 90-95.

Larsen-Freeman, D., & Strom, V. (1977). The construction of a second language acquisition index of development. *Language Learning, 27* (1), 123-134.

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307-22.

Laufer, B., & Nation, P. (1999). A vocabulary-size test of controlled productive ability. *Language Testing, 16* (1), 33-51.

Lee, Y., Gentile, C., & Kantor, R. (2008). Analytic scoring of TOEFL CBT essays: scores from humans and e-rater. *ETS TOEFL Research Reports, 81.*

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English.* Harlow: Longman.

Linnarud, M. (1975). *Lexis in free production: An analysis of the lexical texture of Swedish students' written work.* University of Lund, Department of English: Swedish-English Contrastive Studies, Report No 6.

Malvern, D.D., & Richards, B.J. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (eds), *Evolving models of language* (pp 58-71). Clevedon: Multilingual Matters.

Malvern, D.D., Richards, B.J., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development.* New York: Palgrave MacMillan.

Marcus, G.F., Pinker, S., Ullman, M., Hollander, M., Rosen, T.J., & Xu, F. (1992). *Overregularization in language acquisition. Monographs of the Society for Research in Child Development, 57.*

McCarthy, P.M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing, 24* (4), 459-488.

McNamara, D.S., Louwerse, M.M., Cai, Z., & Graesser, A. (2005). Coh-Metrix version 2.0 Retrieved from *http//:cohmetrix.memphis.edu.*

McNeill, B.R. (2006). *A comparative statistical assessment of different types of writing by Japanese EFL college students.* Unpublished PhD thesis, University of Birmingham, UK.

Meara, P.M. (1996). The dimensions of lexical competence. In G. Brown, K. Malmkjaer & J. Williams (eds), *Performance and competence in second language acquisition* (pp. 35-53). Cambridge: Cambridge University Press.

Meara, P.M. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26* (1), 32-47.

Meara, P.M. (2007). *P_Lex ver2.0* (computer program) Swansea: Lognostics.

Meara, P.M., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16* (3), 5-19.

Meara, P.M., & Fitzpatrick, T. (2000). Lex30: an improved method of assessing productive vocabulary in an L2. *System, 28*, 1, 19-30.

Meara, P.M., Jacobs, G., & Rodgers, C. (2002). Lexical signatures in foreign language free form texts. *ITL Review of Applied Linguistics, 135-136*, 85-96.

Meara, P.M., & Jones, G. (1990). *The Eurocentres Vocabulary Size Test.* 10KA. Zurich: Eurocentres.

Meara, P.M., & Miralpeix, I. (2007a). *D_Tools ver2.0* (computer program). Swansea: Lognostics.

Meara, P.M., & Miralpeix, I. (2007b). *Vocabulary size estimator* (computer program). Swansea: Lognostics.

Meara, P.M., & Olmos Alcoy, J.C. (2008). Words as species: an alternative approach to estimating productive vocabulary size. Retrieved from *http://www.lognostics.co.uk/vlibrary/index.htm.*

Meara, P.M., Rodgers, C., & Jacobs, G. (2000). Computational assessment of texts written by L2 speakers. *System, 28*, 345-354.

Mosteller, F., & Wallace, D.L. (1984). *Applied Bayesian and classical inference: The case of the Federalist Papers.* New York: Springer-Verlag.

Muncie, J. (2002). Process writing and vocabulary development: Comparing Lexical Frequency Profiles across drafts. *System, 30*, 225-235.

Nation, I.S.P. (1990). *Teaching and learning vocabulary.* Rowley: Newbury House.

Nation, P. (2007). Fundamental issues in modeling and assessing vocabulary knowledge. In H. Daller, J. Milton & J. Treffers-Daller (eds), *Modelling and assessing vocabulary knowledge* (pp. 35-43). Cambridge: Cambridge University Press.

Nihalani, N. (1981). The quest for the L2 index of development. *RELC Journal, 12* (2), 50-56.

Oser, E. (1934). *Vater und Sohn.* Konstanz: Sudverlag.

Page, E.B. (1994). Computer grading of student prose, using modern concepts and software. *Journal of Experimental Education, 62* (2), 127-142.

PASW Statistics 18 (2009). *Release 18.0.0.* (computer program).

Perkins, K. (1980). Using objective methods of attained writing proficiency to discriminate among holistic evaluations. *TESOL Quarterly, 14* (1), 61-69.

Read, J. (2000). *Assessing Vocabulary,* Cambridge: Cambridge University Press.

Rennie, J.D.M., Shih, L., Teevan, J., & Karger, D.R. (2003). Tackling the poor assumptions of naïve Bayes text classifiers. *Proceedings of the Twentieth International Conference on Machine Learning, Washington D.C. 2003,* 616-623.

Seber, G.A.F. (1982). *The estimation of animal abundance and related parameters.* London: Griffin.

Shannon, C. (1948). A mathematical theory of communication. *Bell System Technical Journal, 27* (4), 379-423.

Sheehan, K. (2001). *Discrepancies in human and computer generated essay scores for TOEFL-CBT essays.* Unpublished manuscript.

Sichel, H.S. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist, 11*, 45-72.

Sinclair, J. (1991). *Corpus concordance collocation.* Oxford: Oxford University Press.

Smith, R. (2004). The Lexical Frequency Profile: Problems and uses. *JALT2004 at Nara conference proceedings*, 439-451.

Ure, J. (1971). Lexical density and register differentiation. In J.E. Perren & J.L.M. Trim (eds), *Applications of Linguistics* (pp. 443-452). Cambridge: Cambridge University Press.

Wesche, M.B., & Paribakht, T. (1996). Assessing second language vocabulary. *The Canadian Modern Language Review*, 53, 13-41.

Yu, G. (2007). Lexical diversity in MELAB writing and speaking task performances. *Spaan Fellow Working Papers in Second or Foreign Language Assessment, 5,* 79-116.

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: CUP.

Zipf, G.K. (1932). *Selected studies of the principle of relative frequency in language*. Cambridge, MA: Harvard University Press.

Appendices

1.1 A list of specially-designed computer programs

L-unique

Functions: This program calculates the uniqueness index and the distinctiveness statistic for each essay in a set of essays. This program also allows the user to check the frequency of lexical items in the essay corpus as a whole as well as in each essay. It can produce a profile of usage for a particular lexical item across each essay in the corpus. The distinctiveness statistic adjusted so as not to include unique words is also returned.

Inputs: Up to 100 essays can be input simultaneously as text files. Once an initial analysis is complete, the user can select the lexical items to use in the uniqueness index analysis according to frequency in the corpus of essays.

Outputs: Uniqueness index values for each essay and various distinctiveness statistics for each essay are saved to text files.

L-analyzer

Functions: This program calculates a range of lexical statistics for each essay in a set of essays. These statistics include the number of word tokens, the number of word types, TTR, lexical variation, the number of hapax legomena, mean word length, mean sentence length, Yule's K, entropy, the number of words in the JACET 1000 word list as well as counts of various word types such as *a*, *the*, and punctuation marks. It also produces TTR(n) and Hapax(n), the number of different words and the number of hapax legomena for a word sample of a specified size, n, for each essay (default sample size is 100 words).

Inputs: Up to 100 essays can be input simultaneously in text file form.

Outputs: Each count or statistic for all essays is returned in a separate text file.

L-distribution

Functions: This program calculates a range of word distributions for each essay in a set of essays. These include word frequency distributions, ranked word distributions, word length distributions and TTR curve distributions for each essay. The program also returns the most commonly occurring n words in an essay in rank order (n can be specified by the user with default of n=5).

Inputs: Up to 100 essays can be input simultaneously as text files.

Outputs: Word frequency distributions, ranked word distributions, word length distributions, TTR curve distributions and the most commonly occurring words for each essay are saved to separate text files.

L-cluster

Functions: This program splits a set of 100 essays into two sets of 50 predicted higher quality essays and 50 predicted lower quality essays using a clustering algorithm based on a number of essay characteristics.

Inputs: Lexical statistical information on each essay in text file form.

Outputs: Two text files are returned, one identifying essays in the predicted higher quality group and the other identifying essays in the predicted lower quality group.

L-Bayes

Functions: This program classifies a set of 100 essays into two sets of 50 predicted higher quality essays and 50 predicted lower quality essays using a Bayesian algorithm which uses the occurrence of all words in the essay corpus.

Inputs: One hundred essays in text file form.

Outputs: Two text files are returned, one identifying essays in the predicted higher quality group and the other identifying essays in the predicted lower quality group.

L-capture

Functions: This program performs a capture-recapture analysis to estimate vocabulary size based on two essays written by the same learner.

Inputs: Two sets of essays in text file form. Up to 100 essays can be input for each set.

Outputs: Capture-recapture estimates of vocabulary size based on each essay are returned in text file form

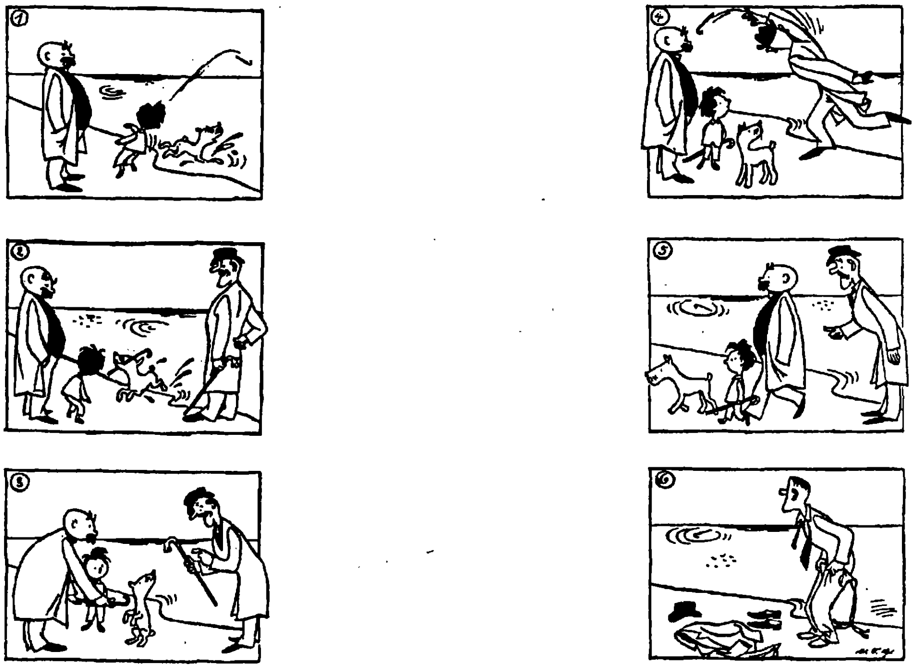
L-capsim

Functions: This program produces S multiple capture-recapture population estimates of a given population size N, using two samples of size s_1 and s_2. Up to 10000 estimates can be produced.

Inputs: The number of iterations (S), population size (N) and two sample sizes (s_1 and s_2)

Outputs: The multiple estimates of population size are saved to a text file.

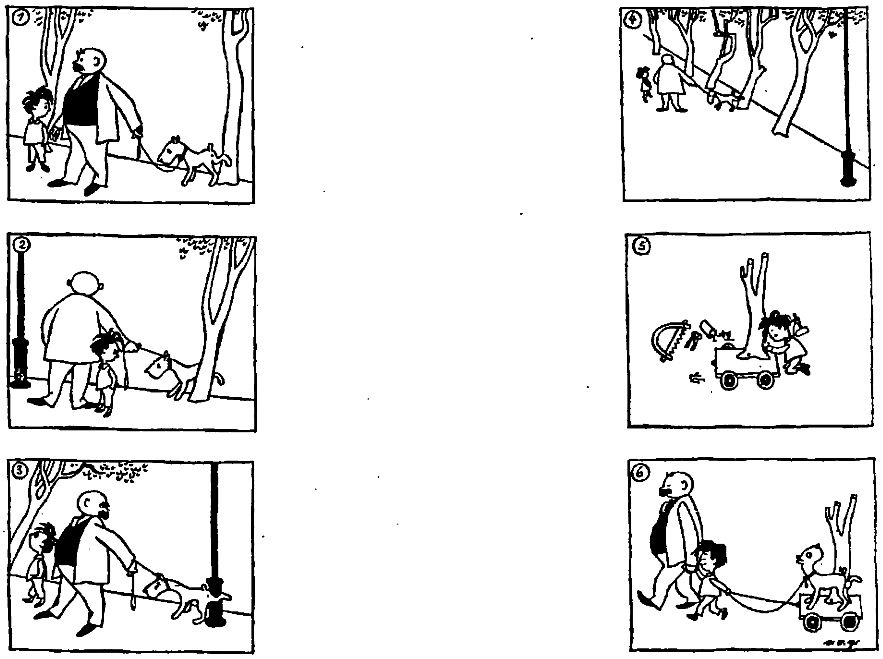
3.1 Cartoon Task 1



(Oser, 1934)

Write a story using the pictures. You have 25 minutes to write as much as you can. You should write at least 100 words.

3.2 Cartoon Task 2



(Oser, 1934)

Write a story using the pictures. You have 25 minutes to write as much as you can. You should write at least 100 words.

3.3 Sample essays for Cartoon Task 1

Three sample essays are presented. Essay A is relatively long, Essay B is relatively short and Essay C is of about average length.

Essay A (230 words)

There are two people in the picture. The man is father "Bob" and this girl playing with a dog is Bob’s daughter "Lucy". They are very close. They walk with dog to sea near their house. Today is very sunny day. Dog’s name is "Lucky". Lucky likes take bring peace of wood, or ball or something. Lucy threw peace of wood. Lucky bring this wood. Lucky is very fun this game. Lucy say to Lucky good job good boy. Lucky thought that I want to bring something again. This is very fun! Bob threw his stick to sea. Lucky ran into the sea and take it came back soon. The man was wathing all. The man thought. Is that dog bring my stick? Let’s try!! The man threw his stick into sea. Bob Lucky and Lucy just watch. Lucky didn’t take bring his stick. The man thought why that didn’t take bring to me? They went back their home. Lucky didn’t want to bring stranger’s thing. Lucky want to bring his keeper’s thing. Only Bob’s and Lucys. After all, that man put off his wear and take bring by himself. Today is sunny day but, now is not summer. Water is cold. That man had a catch cold and ferver. That man is very poor. This time Bob and Lucy and her dog spend very good time in their house.

Essay B (100 words)

A boy and a dog are playing. And a man stand near them. "Catch this!" said boy, and he through the stick. Dog runs in the water and come back to take the stick to the man. A man come hear. "This dog is clever this stick through away and take to me. OK?" he said. "Catch this!" said him and he through the stick far away. but they going to opposite. "Is that take to me?" The man asked. But they are not answered. "I going to take the stick myself." He said, and he took off him clothes.

Essay C (140 words)

A boy throw his grandfather’s stick and his pet dog ran into the sea for get the boy’s grandfather’s stick. The dog could get stick and come back from the sea. The other old man that seeing this situation throw his stick too. He thought, the dog get his stick too. But the dog didn’t get into the sea. The dog get only his keeping owner’s stick. So, the dog ignore the old man’s action. So the old man had to into the sea for get his stick by himself. So he put out his shoes clothes and hat. It is funny. Pet is get stick for he wants to loved his keeping owner. The old man should keep a pet if he doesn’t want to get his stick by hisself. I think the old man is poor a little.

3.4 Combinations

The number of different combinations of r items selected from a population n is ${}_{r}C_{n}$ as follows:

$$ {}_{r}C_{n} = \frac{n!}{r!(n-r)!} $$

where n! = n x (n-1) x (n-2) x……x 3 x 2 x 1

Example

The number of different 6-word subsets possible from a population of 30 words is

$$ {}_{6}C_{30} = \frac{30!}{6!(30-6)!} = \frac{30x29x28x27x26x25}{6x5x4x3x2} = 593775 $$

Increasing the number of words to 50 or the subset size to 8 or 10 only makes the number of combinations even larger.

For example, when selecting six words from 50, there are an unbelievable number of combinations. If you selected a different subset every 10 seconds of every day (and night), it would take more than 5 years to get them all (15,890,700 different subsets). If you made it a 10 word subset from 50, it would take 81 years!! (but only if you speeded up and found a subset every second and got your spouse and two kids doing the same).

3.5 Sample essays for Cartoon Task 2

Three sample essays are presented. Essay A is relatively long, Essay B is relatively short and Essay C is of about average length.

Essay A (250 words)

I'm a Bob. This is my father, Gobline. And my dog, Kanehiko. We always walk together. Today, too. Now, Kanehiko is coming out his water from his body. "Oh, very long time." I said. Still he was doing. "Very very long time." I had a little anger while he was doing. Still he kept doing. My father, Gobline turned back and pulled a lope. We had a little anger that the same feeling. "Very very very long time!" I said as I was going ahead. Still he wasn't stopping! He didn't separate with the tree. I said "Does he like a tree??" I'm very angry. Because, I want to go home and eat a chocolate. So, I went to my house and came into my room. There were many tools there. In fact, I want to become artist in my future. I like to make something with tools. And I started to make a tree that Kanehiko loves with a board and four small tyres. I feel that my hammer jumped on steel pins. "Ton, ton, ton." it danced with me. After I finished my great work, I turned to go to the place that Gobline and Kanehiko were. He was keeping to do still there. My father were tired and sat down around there. I gave him to new tree. Kanehiko move to the tree! Success!! Kanehiko gladed, Gobline gladed, too. Of course and me! I was satisfy very much. We started to go home. "Wait my chocolate, pleace"

Essay B (101 words)

One day a old man and a young child were walking the street with a dog. The dog stopped and raise the dog's back leg by trees of the street one by one. A few minutes later, the young child and old man were getting angry about the dog's doing. After walking, the young children thought about it, and started making tree which was cut and put on the truck. Next walking, the dog get on the truck and from then walking became smoothly, but the dog would not like to get off it. And keep to raise the dog's leg.

Essay C (147 words)

There was a dog he likes walk to street he always walk to street with his boss and boss's childlen. One day he walked street with his boss and boss's children. First, he tried do toilet on a tree's bottom. But he didn't. Then, he didn't toilet on a light's bottom too. His boss was angry. Boss's children was boring. And he didn't toilet next tree's bottom too. He tried do tilet on all tree's bottoms. But he didn't on all tree's bottoms. He could not choice best tree's bottom. His boss and boss's childlen really boring to him. The boy was think. And he had very good idea. He made a move tree for the dog. Now, the dog go to street everyday with his boss and boss's children. He must not choice best tree's bottom for toilet. Because he has a tree for himself.

3.6 Index of uniqueness using words occurring frequently in two different tasks

These words occurred in close to 32 texts in both assignments. The words used are shown in Table A1 and the results of the analysis in Table A2. Uniqueness index values are similar to the main analysis in terms of average number of signatures per subset and uniqueness index scores. The correlation coefficient (r) between task scores is again very low at 0.11. Figure A1 shows the uniqueness index scores for both tasks using these words.

Table A1: Thirty words appearing frequently in both tasks

day for go good one that then there this thought too were father at by can do don't move of put want didn't had my take are have their with

Table A2: Unique lexical signatures and uniqueness index scores for both tasks

	Task 1	Task 2
Average unique signature per subset	17.98	17.47
Average uniqueness index score	0.28	0.27
highest	0.65	0.54
lowest	0.06	0.06
Correlation (r) (Task 1 & Task 2)	0.11	

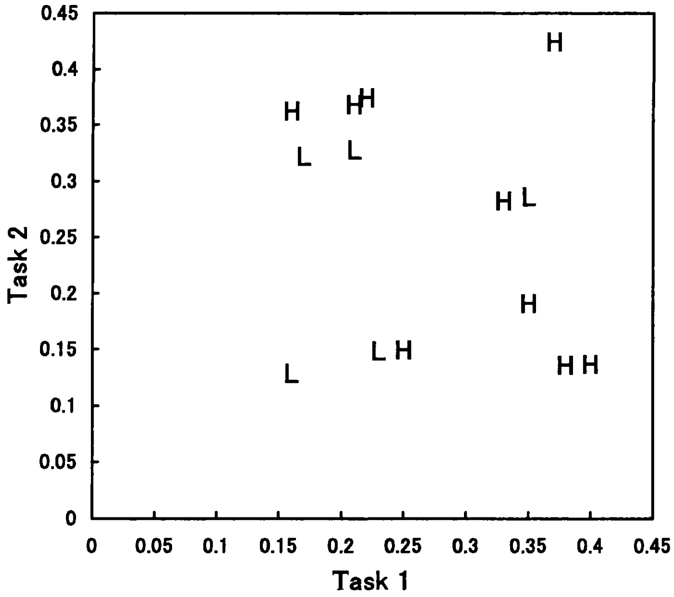


Figure A1: Uniqueness index scores using words from both tasks

5.1 Sample learner essays and native text for review task

5.1.1 Sample learner essays

Three sample essays are presented for each task. Essay A is the longest, Essay B is the shortest and Essay C is of about average length.

Essay A (418 words)

The Cars is feature animation film which was released in July in 2006 in Japan. This movie got a Golden Glove Winning in the best animation section. The director is John A. Lasseter. He is a member of Pixar executive producer and he directed many pixer's movie. This movie was a story of the car. The main character is rookie racing car who name is Lightning McQueen. He has dreamed all his life of winning the Piston Cup Championship. Opening started the exciting race scene with a big sound. It attracted our feelings because it sound was real and the animation was so clear and beautiful. McQueen had to take part in next race in Los Angeles in California. McQueen and his transport truck Mack begin a journey across the country to California on the highway. However Mack slept as driving and McQueen was left on the road and got to Radiator Springs. While McQueen was running away from police car, Sheriff, he destroyed the road in Radiator Springs. He took to a traffic court and required to repair the road. At first he didn't like this town but he met Sally, Mater and Doc Hudson and knew the history of Radiator Springs. He loved Sally and made best friend, Mater. This movie told real friendship and kindness. The characters were very attractive like human and interesting story draw us into the cars world. At last he repaired the road completely and maintained his body by Luigi and Guido. He gave happy all citizens and they appreciated him. Its may difficult to kind to others but the kindness is natural feeling. McQueen changed his heart by good friends easily so good friends are so important to get a win. Suddenly Mack and many TV reporters and camera came to the town and he was taken by Mack to California. In the race, he had no partner so it was hard to win the race but Radiator Spring's friend came to the race and supported McQueen. Thanks to them he got big power but in the last rap, old racing car clashed. He helped him in the verge of win. Everyone admired McQueen. Finally the Radiator Springs became flourishing because McQueen decided Radiator Springs as the stronghold. He noticed the importance of kindness. This movie told that we can get something important by the failure. The last was happy end and this movie was easy to understand so people really loved it from children to adults and enjoy watching.

Essay B (269 words)

Mucc is one of Japanese rock bands. Now, they are becoming popular all over the world. They were really dangerous. Their songs express human's bad emotion and social contradiction clearly. Their bass sounds were queer like a snake that crawls under ground. The guitar sounds made heavy and aggressive unison. The lyrics were painful. They often used these words, death, hate, despair, bleed, loneliness and so on. If you listen to their old songs when you droop, you might commit suicide. Their songs were lethal and they had reality. Their world was only dark. However, they are changing. They start to sing about "over the pain". This album "Gokusai" means "colored vividly". Yes, they got colors. The songs in "Gokusai" express hopes, future, lives. Not only those good things but also they have pain yet. Therefore, the songs have reality and reach our hearts. The sounds are also aggressive, but they have some kindness. Especially, "Utagoe" and "Yasasi Uta" have new style, they had never performed like them. Miya, the leader of the band said that he can not recognize this social disease and he can not forget these pains yet, however, it can not be helped, we have to live, and he can get good things from this world. This idea can be found in Utagoe's lyric, "I love this world that may not have enough merit to live" (ikirukachi mo naiyouna sekaiwo aisiteru). This year, they are going to great with 10 year's anniversary. They will release 2 best album named best of mucc and worst of mucc. It will be interesting to see what they do.

Essay C (344 words)

This movie is one of the "GHIBLI" movies. A producer is Hayao Miyazaki. He is very famous producer in Japan. He also set up the original and scenario of this story. In addition, he wrote lyrics for a song, which is a main song "Kimi wo Nosete". People don't know about this story is what country and time. Mr. Miyazaki had visited Wales, the UK before he made this movie. Therefore, you can see beautiful scene *looks like the UK. Also, he said this movie set is in that country. On the other hand, you can guess the time of this story. Pazus' father had taken a picture of Laputa. And it was printed "1968 7" so this story is between 1970's and 1980's. However, the flight technology is more developed than now. The material of this story is from "Gulliver's travels". The story is about "Laputa", which is the land floating in the sky. There are many characters and they are quite important of Raputa. Two main characters are Pazu and Sheeta. They are very brave and gentle to everyone. Pazu is a boy and works at a coal mine. He is bereaved his parents. Sheeta is a descendant of Laputa family. Their enemy is Musuka. He is also a descendant of Laputa and he wants to be a king of Laputa and revive that land. All of them are so individual. Laputa is rare land and Pazu's father discovered it but anyone didn't believe him. Pazu's dream is to go Laputa and he wants to know about it. On the other hand, the army and the pirates chase Sheeta because she is the throne of Laputa and she has a piece of flying stone which is the most important part of Laputa. Pazu and Sheeta save Laputa from Musuka who is the army and family of Laputa throne. We can feel love for Laputa and partners of the main 2 characters. Pazu and Sheeta experience many dangers but their bond becomes strong. We can enjoy the story and beautiful scenery of Laputa and the sky.*

5.1.2 Sample native text

Pelican Pete's until recently was a sidewalk bar located in the vicinity of Nagoya Dome. Though it was a cosy little place that sat four but stood dozens if the sidewalk was used with reckless abandon, owner Tip decided late last year that it was time to move into bigger and better quarters. At its new location in Ikeshita behind the Chikusa Kuyakusho, there are now enough seats for everyone on most nights. Pelican Pete's promotes an extremely friendly beach bar atmosphere kinda like Cheers meets Margaritaville that should please many locally displaced parrotheads. For some the beach atmosphere and Tip's loud aloha shirts alone are enough to make them forget what part of the world they are in. But for the more discerning, it is the menu that will send them on a cruise. The extensive menu at Pelican Pete's is what you would expect to be on a menu back home if you are from North America. Predominantly American, the items boasted on the menu are salads from 500 yen sandwiches from 500 yen and dip nachos 700 yen. But what caught my eye were the burgers. I have a hard time finding a good hamburger in Nagoya so Pelican Pete's is a godsend. The plain burger starts at 400 yen but Pete's menu offers up to 18 different toppings that can be added to make your own. If this still isn't enough, there is always beer. Since Tip's favorite drink is beer it is no wonder that extra effort is put into the selection at Pelican Pete's. Forty five types of beer from 400 yen from 20 countries around the world.

5.2 List of function words from the British National Corpus

(Leech, Rayson & Wilson, 2001)

Pronouns

it, I, you, he, they, she, we, who, them, him, me, her, one, us, something, nothing, himself, anything, itself, themselves, someone, everything, herself, anyone, everyone, whom, myself, yourself, none, no-one, somebody, anybody, his, plenty, mine, lots, ourselves, yours, hers, ours, whoever, theirs

Determiners

the, a, an, their, its, my, your, no, our, every

Determiner/pronouns

this, that, which, what, all, some, these, any, many, those, such, more, own, same, another, much, each, few, most, both, several, half, whose, little, former, whatever, less, enough, latter, either, fewer, neither

Prepositions

of, in, to, for, with, on, by, at, from, as, into, about, like, after, between, through, over, against, under, without, within, during, before, towards, around, upon, including, among, across, off, behind, since, because, rather, until, according, up, despite, near, above, per, along, away, throughout, outside, round, beyond, worth, down, inside, instead, plus, past, front, apart, onto, beside, below, beneath, amongst, via, unlike, addition, prior, concerning, next, except, alongside, till, ahead, depending, regarding, toward, opposite, following, amid, underneath

Conjunctions

and, but, or, if, when, than, while, where, whether, so, though, cos, nor, &, unless, once, even, whereas, whilst, albeit

Interjections and discourse markers

yeah, oh, no, yes, mm, ah, mhm, aye, ooh, hello, hallo, dear, eh, ha, aha, hey, bye, yep, goodbye

Notes

1) Words are listed by category where they have not been included in a previous category. For example, in the BNC list, *his* occurs in the most frequent list of determiners but is excluded here because it is included in the list of pronouns.

2) Multi-word lexical items are abbreviated if part of the word has appeared earlier. If all the components of a multi-word item have appeared earlier, then that multi-word item is not included. This means the form of these lists is quite different to those found in Leech, Rayson & Wilson (2001). For example, *according to* in the prepositions list is abbreviated to *according* because *to* appears earlier in the list. *Along with* does not appear because both *along* and *with* appear earlier. This highlights a weakness of this primitive form of automatic calculation. It cannot make distinctions between single word items and multi-word items. Such distinctions could be a feature of future more sophisticated versions.

6.1 Sample essays for TV task

Three sample essays are presented. Essay A is the longest, Essay B is the shortest and Essay C is of about average length.

Essay A (328 words)

I don't think watching television is bad for children. It is true that there are some harmful TV program for children such as programs including sexual or violent expressions. There is an evidence that it causes bad affection for children. However there are a lot of good reasons for watching televisions. Firstly, there are educational programs which treat science, history and mathmatics. They are usually broadcasted during day time. Why is it during day time? I guess it's for the case when children is absent from their school, they can watch and learn something without going to school. Actually the programs are more interesting than a class of school. I used to watch such a program when I was absent from school. Secondly, there are some kinds of news program. For children it's painful to read a news paper, because they can't consentrate and there are so many words they don't know due to a lack of experience of reading books. Therefore watching news on TV can be the important connection between children and world. Phewhaps some people says that news program is also difficult for children, but there is a TV program which teaches children very understandably such as "kodomo-news (news for children)". Thirdly, there are variety show and comedy show as a good reason. Some people believe that they are harmful, because in the shows there are only funny things. It's true that watching too much or only the comedy shows makes children stupid, but it is not because of the program, but a lack of studying. Or rather, the comedy programs can be an important role that children can find a way to express themselves widely, and make their characters. As a conclusion, I can understand people who say that TV program is harmful, because there are actually bad programs. However, if there no TV program except for bad one, TV industries must haven't lasted. This is why TV program should be watched by children.

Essay B (50 words)

I disagree with that whacting televivion is bad for children. Although there are many bad program for children, children can learn many things from television. For example, children can learn name of things and lunguage. Also, we can look other country's scenery and other beautiful things. It grows children heart.

Essay C (147 words)

I agree with the statement. My opinion is children should read good books, such as business books than watching television. This is because, I think, most TV problem is made for all people, from kids to elder people. It should be easier to make people understand. While, books is good materials to get new information. Usually, books are consisted by a lot of ideas of author. We can get those ideas author has gotten through their life by just paying only a few thousand yen. There may be information, which author doesn't say on TV, in books. Author takes much more time to make their books. Even some people say that watching television is waste of time. I think so too, and children should keep time to watch less than 30 minutes a day. Because of these reasons, I support the statement "Watching television is bad for children.

7.1 Principal components graphed against standardized variables

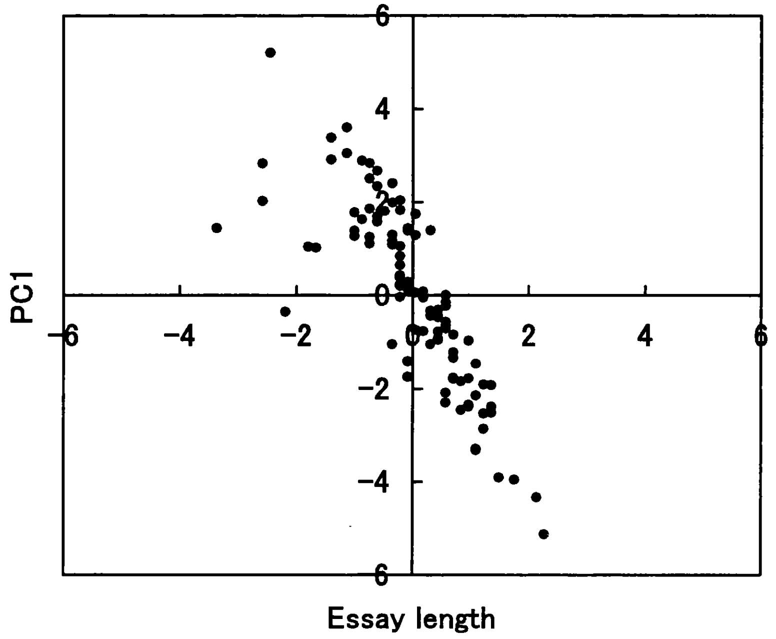


Figure A2: PC1 v essay length

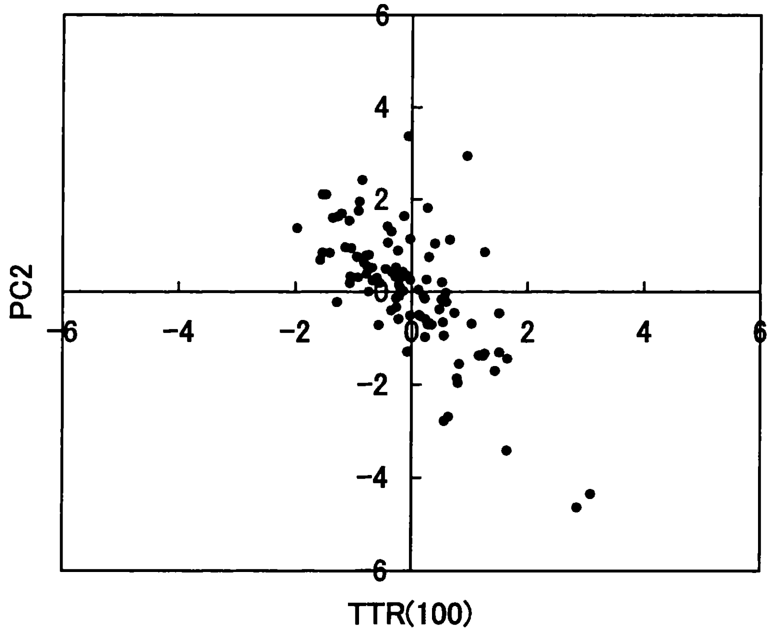


Figure A3: PC2 v TTR(100)

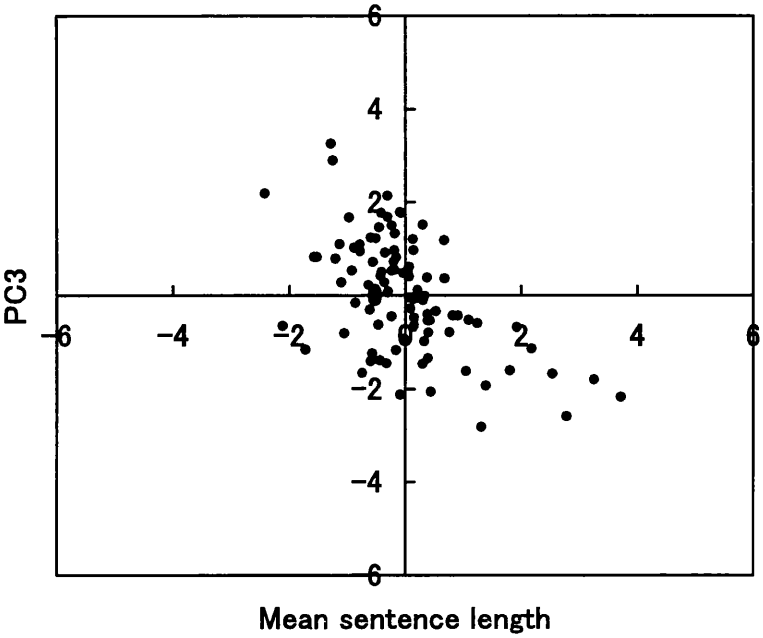


Figure A4: PC3 v mean sentence length

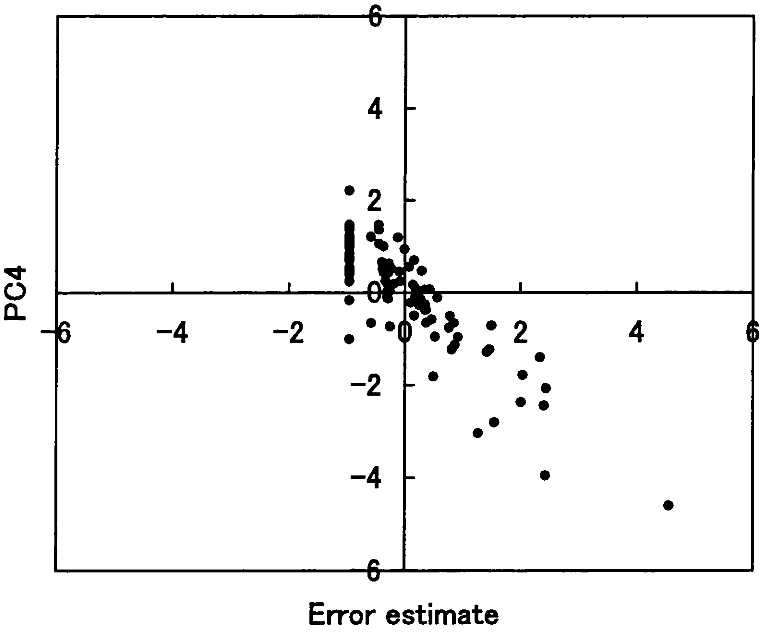


Figure A5: PC4 v error estimate

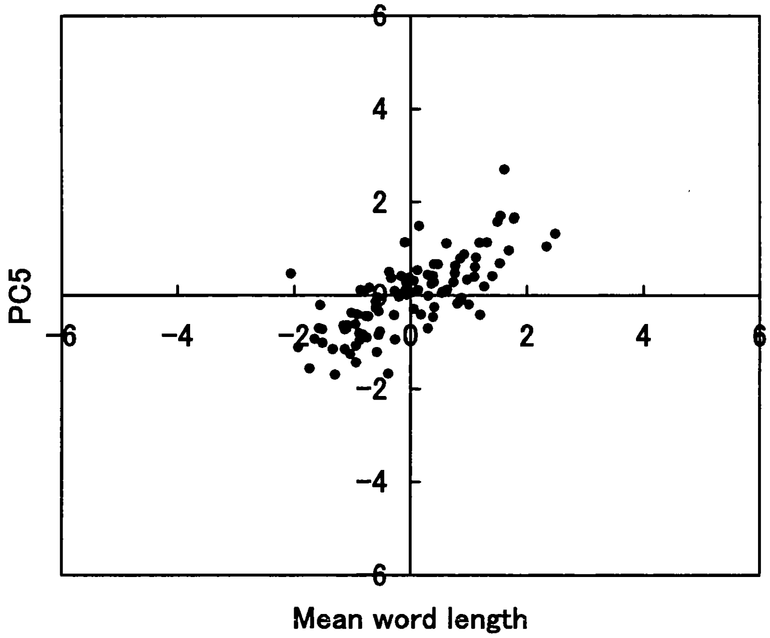


Figure A6: PC5 v mean word length

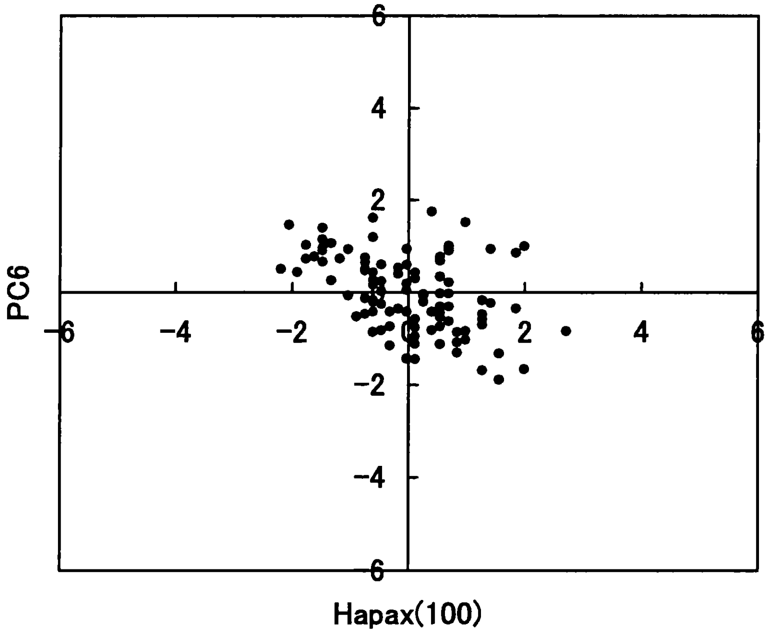


Figure A7: PC6 v Hapax(100)

7.2 Euclidean distance

The Euclidean distance between two points x and y in n-dimensional space is as follows:

$$\mathbf{d(x,y)} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

Weighted euclidean distance can be defined as follows:

$$\mathbf{d(x,y)} = \sqrt{w_1(x_1 - y_1)^2 + w_2(x_2 - y_2)^2 + ... + w_n(x_n - y_n)^2}$$

Example:

In the clustering algorithm, the distance of each essay represented in six-dimensional space was calculated to each initial cluster point.

The initial high cluster point (C_h) was:

$c_h = (1,1,1,-1,1,1)$

The first essay point, $_1e$, was:

$_1e = (0.83, 0.70,-0.45,0.29,0.63,0.45)$

The weightings for each dimension were:

(1, 20/34, 13/34, 12/34, 6/34, 6/34)

The distance between the first essay and the initial high cluster point was:

$$d({}_1e, c_h) = \sqrt{(1-0.83)^2 + \frac{20}{34}(1-0.7)^2 + \frac{13}{34}(1+0.45)^2 + \frac{12}{34}(-1-0.29)^2 + \frac{6}{34}(1-0.63)^2 + \frac{6}{34}(1-0.45)^2}$$

$= \underline{1.25}$

7.3 Kappa statistic

The Kappa statistic (K) can tell us the inter-rater reliability between raters adjusted for chance agreement.

$$\text{K} = \frac{(P_0 - P_e)}{(1 - P_e)}$$

where P_0 is proportion of agreement observed and P_e the proportion of agreement expected by chance. For two response categories, P_0 and P_e are calculated as follows in Table A3:

Table A3: Two rater representation

		Rater 2		
		1	2	Sum
Rater 1	1	A	B	A+B
	2	C	D	C+D
Sum		A+C	B+D	N

$$\text{P}_0 = \frac{(A+D)}{N}$$

$$\text{P}_e = \frac{(A+C)}{N}\frac{(B+D)}{N} + \frac{(A+B)}{N}\frac{(C+D)}{N}$$

Example: The agreements of two raters are as follows in Table A4:

Table A4: Example two rater representation

		Rater 2		
		Good	Poor	Sum
Rater 1	Good	36	14	50
	Poor	14	36	50
Sum		50	50	100

$$\text{P}_0 = \frac{(36+36)}{100} = \frac{72}{100} = 0.72$$

$$\text{P}_e = \frac{(50)}{(100)}\frac{(50)}{(100)} + \frac{(50)}{(100)}\frac{(50)}{(100)} = \frac{1}{4} + \frac{1}{4} = 0.5$$

$$\text{K} = \frac{(0.72 - 0.5)}{1 - 0.5} = \frac{0.22}{0.5} = \underline{\mathbf{0.44}}$$

7.4 Technical problems associated with Bayesian classifiers

The following problems have been associated with using Bayesian classifiers with essay data:

- A need to avoid absolute probabilities 0 and 1 since the Bayesian calculation includes multiplication of probabilities and complement probabilities
- Problems with modeling multiple occurrences of a word in an essay
- When dealing with long essays, Bayesian classifiers often polarize at values of 0 and 1 which makes incremental assignment to categories difficult
- Bayesian classifiers may be overly influenced by long essays
- They may also be overly influenced by words that occur in many essays
- Training samples of differing sizes may distort results

As a way around these problems, an algorithm suggested by Rennie et al. (2003) was used. This addresses the above concerns using a set of transformations as follows:

- A transformation of adding one to each word occurrence
- Using a log value of word occurrence
- Discounting weights of words according to the number of essays they occur in
- Normalizing essays for length
- Using complement probabilities, that is, in this case, the probability that an essay does not belong to the high quality group
- Using log weightings
- Normalizing log weightings