An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)

A Dissertation

Presented for the

Doctor of Philosophy

Degree

The University of Memphis

Philip M. McCarthy

August, 2005

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.

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

ProQuest Information and Learning Company
300 North Zeeb Road
P.O. Box 1346
Ann Arbor, MI 48106-1346

To the Graduate Council:

I am submitting herewith a dissertation written by Philip McCarthy entitled "An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD)." I have examined the final copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for a degree of Doctor of Philosophy with a major in English.

Charles Hall, Ph.D.
Major Professor

We have read this dissertation and recommend its acceptance:

Teresa S. Dalle, Ph.D.

Emily Thush, Ph.D.

Arthur C. Graesser, Ph.D.

Max Louwerse, Ph.D.

Accepted for the Council:

Karen D. Weddle-West, Ph.D.
Assistant Vice Provost for Graduate Studies

For Yuko

Acknowledgements

As a linguist, it should come as little surprise that the form of my acknowledgement section should be the result of a corpus investigation. Many days of compiling and looking over a number of such sections helped to established that there were two over-arching forms of acknowledgement. First, there were the *robotic* and *soulless* acknowledgements which amounted to "Thank you, really, now give me the degree and let's get on with the real work." And second, there were the somewhat *gushy*, and *melodramatic* versions that amounted to "Thank you, really. I love you, really."

Seeing the merits of both acknowledgement styles, I tried valiantly to synthesize the forms into a blend of heartfelt gratitude and cold, hard facts.

I failed.

In the end, *the gratitude* far outweighed *the facts*, meaning that what follows is nothing more than melodramatic gush. Your indulgence for the next two pages, therefore, would be greatly appreciated.

First, unparalleled and hitherto unheard of thanks to my mentor, Charles Hall, for his patience, brilliance, humor, kindness, for his endless faith in me, and for his endless support of me. I am forever in your debt.

Thanks to Teresa Dalle for four years of unremittingly wonderful advice. Thank you above all for getting me started at the University of Memphis by making my funding possible. I am incredibly grateful.

Thanks to Emily Thrush, who also supplied four long years of support, advice, and great patience. Thank you also for making possible my first conference presentation and my first publication.

Thanks to Max Louwerse, for his invaluable roll in developing the MTLD lexical diversity tool used in this dissertation. Thanks also for the endless readings of endless chapters and for keeping me in line and on target.

Thanks to Art Graesser, for all his help, support, and time, especially when my ideas seemed to be coming exclusively from left field. Deep in left field. Way deep.

Thanks to Danielle McNamara, for funding me at the Institute for Intelligent Systems, for supporting me, and for not only listening to all my mad ideas, but for actually encouraging me to go out and do them. You are an inspiration, and I am so very grateful.

Thanks to Scott Jarvis for providing the corpus I used in Chapter 5 of this dissertation. And thanks also for providing the encouragement and interest that motivated much of this work.

Thanks to Scott Crossley, David Dufty, Christian Hempelmann, Carl Cai, and Xiangan Hu, for listening again and again and again and again to all my tales of lexical diversity. None of this would have been possible without you.

Thanks to Dr. Tigyi and his family for their support of my wife and me here in the U.S. Thanks for giving us the opportunity to come to this country and attend our universities. We owe you more than can ever be repaid.

Thanks to all the people at the Institute for Intelligent Systems, all my colleagues, all my students, and everyone on the magnificent Memphis Strangers soccer team for years and years of great times, great debate, and great fun.

Thanks to Steven Pinker, for writing *the Language Instinct*, and for changing my life by doing so. Without your book, and your kind and thoughtful responses to my posts, I would never have taken up linguistics.

Thanks to my family, who have always been there for me, especially when I deserved it the least. But especially thank you to my father, who can have no idea how proud I am that he is my dad.

And finally, thanks to Yuko: my wife, my best friend, and my greatest supporter. Thank you for making all this possible. *Kotoba dewa iitsukusenai hodo kansha shiteruyo.*

ABSTRACT

McCarthy, Philip, M. Ph.D. The University of Memphis. August, 2005. An Assessment of the Range and Usefulness of Lexical Diversity Measures and the Potential of the Measure of Textual, Lexical Diversity (MTLD). Major Professor: Charles E. Hall, Ph.D.

Lexical diversity encompasses a wide variety of measures, all of which seek to quantify the range of vocabulary deployed in a given text. Researchers use these measures of lexical diversity in many fields, including language acquisition, stylistics, neuropathology, and linguistic forensics. Unfortunately, text length confounds all the measures, leading to questions of the conclusions of some studies. Many alternative measures have been introduced but all have failed to overcome the problem of text length correlation. We introduce and test a new measure of lexical diversity: the measure of textual, lexical diversity (MTLD). We test MTLD and 13 of the best known traditional measures of lexical diversity against the largest corpus yet established for such a test: 23 genres of spoken and written texts, comprising 414,000 words. The results of these tests supply evidence that none of the traditional measures avoid correlation with text length. MTLD, however, does not correlate with text length over the ranges tested suggesting that MTLD is the first reliable measure of lexical diversity. The significance of such a measure is that researchers and educators will be able to assess the lexical diversity of both spoken and written texts without concern for the differing text lengths. We also test all the traditional measures against a further corpus of NS and NNS. In these tests, both MTLD and some of the traditional measures predicted differences in the results. We conclude that MTLD is the only sophisticated measure that avoids correlation with text

length but that using other sophisticate measures, in conjunction with MTLD, may be the best approach to analysing texts.

TABLE OF CONTENTS

	PAGE
List of Tables	xiv
List of Figures	xvi
CHAPTER	
1. LEXICAL DIVERSITY	1
What is Lexical Diversity?	1
Why the Need for a Quantitative Measure of Text?	2
What is the Problem With Lexical Diversity?	3
What is the Purpose of This Dissertation?	4
What Special Terms Will Be Used in This Dissertation?	5
How is This Dissertation Structured?	6
What is the Scope of This Investigation?	7
What is the Importance of This Study?	8
Summary	21

2. ESTABLISHED MEASURES OF LD 22

What is TTR? 22

Is TTR Ever Useful? 25

What are Receptive Measures? 30

Can the Problem of TTR Be Corrected? 32

Where Next for LD? 59

3. ASSESSING COMPLEX MEASURES OF LD 62

How did we Select Material for the Test Corpus? 63

What was the Method Used for Analysing the Texts? 66

How Did We Select Texts? 71

Which LD Measures Were Included in the Study? 75

Results 77

Discussion 85

4. INTRODUCING THE MEASURE OF TEXTUAL, LEXICAL DIVERSITY (MTLD) 87

How Is Textual, Lexical Diversity Distinct From Lexical Diversity? 88
What Is the Measure of Textual, Lexical Diversity? 94
How Is MTLD Calculated? 95
How Are Incomplete Factors Scored? 99
Why Is a Factor Size of TTR 0.71 Used? 101
Why Is There a Minimal Factor Size? 106
Why Is There Reverse Order Analysis of the Text? 107
What Is the Reliability of MTLD? 109
Is MTLD Sensitive? 111
Is MTLD Effective Over a Wider Range Than Other LD Measures? 113
What Percentage of the LD Scores Can Be Accounted for by Text Length? 119
Summary 122

5. APPLYING MTLD AND OTHER COMPLEX MEASURES TO A CORPUS OF NATIVE AND NON-NATIVE SPEAKERS 124

Which Materials Were Used? 125
Which Measures Were Included in This Analysis? 127

Which Questions Are to Be Addressed in This Analysis? 127
Question 1: Can Convergent Validity Be Established
Between MTLD and Other Complex LD Measures? 128
Conclusion 132
Question 2: Do the Differences in Results That
MTLD and Other LD Measures Produce Help Us
to Better Understand Textual Differences Produced
by NS and NNS? 133
Conclusion 135
Question 2a: Do the Differences in Results That
MTLD and Other LD Measures Produce Help Us
to Better Understand Textual Differences Produced
by 9th Grade NS and NNS? 136
Question 2b: Do the Differences in Results That
MTLD and Other LD Measures Produce Help Us
to Better Understand Textual Differences Produced
by 7th Grade NS and NNS? 140
Conclusions 143
Question 3: Do the Differences in Results That LD
Measures Produce Help to Indicate the Grade of the
Student Writer? 144
Conclusion 148
Post Hoc Question 151

6. CONCLUSION 154

What Was the Purpose of This Dissertation? 154

Why Was This Research Necessary 154

What Did the Dissertation show? 156

What Is the Significance and Implications of These Findings? 159

What Future Research Might Be Conducted Using MTLD? 160

How Else Might the MTLD Technique Be Used? 161

What Alternative Measures Might Be Used? 163

What Are the Limitations and Problems for MTLD as a Measure of LD? 165

Conclusion 167

7. REFERENCES 168

LIST OF TABLES

Table	PAGE
1. TTR Comparison of Four Texts Showing Differing LD Developments.	64
2. Text Lengths Used for the LD Analysis.	69
3. Comparison of this Study's Range of Material Against Previous Similar Studies.	72
4. The Number of Files for Each Genre and their Source.	74
5. Significance Correlations for the 7 Measures (a).	77
6. Significance Correlations for the 7 Measures (b).	79
7. The Standard Deviation Scores for Genres with High LD Scores.	84
8. Calculating the Remainder Score for a Hypothetical Text.	101
9. ANOVA Results for Contrasts by Measure Across 2277 Section Sizes of Texts.	114
10. The Four Best Performing Optimal LD Measure Ranges.	116
11. The Best Performing Ranges for the Best Performing LD Measures from Chapter 3.	118
12. Variance Explained by Text Length (100-2000 Tokens) for all LD Measures.	120
13. Variance Explained by Text Length (100-500 Tokens) for all LD Measures.	121
14. Details of Subjects Used in Chapter 5 Study.	126
15. Correlations for U and D^a Against Other Measures.	130

16. ANOVA Results for Measures Predicting Differences for NS and NNS. 131

17. ANOVA Results for Measures Against 122 9th Graders. 134

18. Mean and Standard Deviation Differences Against 121 9th Graders. 137

19. MTLD and D^b Differences. 138

20. ANOVA Results of Measures for 71 7th Graders in Predicting Years of English. 139

21. Mean and Standard Deviation Differences Against 71 9th Graders. 141

22. Distinguish Sub-Groups of 7th Graders on Years of English. 142

23. Mean and Standard Deviation Differences Against 84 Subjects. 143

24. ANOVA Results for Measures Against 71 7th Graders to Predict Years of English. 146

25. Analysis Across Groups of 5th, 7th and 9th Graders with 2 Years of English. 147

26. The Parts of Speech as a Percentage of Overall Tokens per Group. 148

LIST OF FIGURES

Figure	PAGE
1. TTR Comparison of Four Texts Showing Differing LD Development.	27
2. Peaks and Valleys Formed by TTR Analysis (a).	29
3. Peaks and Valleys Formed by TTR Analysis (b).	29
4. Peaks and Valleys Formed by TTR Analysis (c).	30
5. The TTR Curve of *Red Badge of Courage* with Logarithmic Curve Estimation.	40
6. Flowchart for MTLD.	98
7. TTR Factor Sizes Against Degree of Difference Scores for Three Different Genres.	105

Chapter 1: Lexical Diversity

When you measure what you are speaking about and express it in numbers, you know something about it, but when you cannot express it in numbers your knowledge about it is of a meagre and unsatisfactory kind.

William Thomson (Lord Kelvin) (1824 - 1907)

What Is Lexical Diversity?

Lexical diversity (LD) measures the range of vocabulary deployed by a speaker or writer. Greater LD is widely held to be indicative of greater linguistic skills, speaker competence, or even a speaker's socioeconomic status (Avent & Austermann, 2003, Carrel & Monroe, 2004; Grela, 2002; Ransdell & Wengelin, 2003). Somewhat synonymous to LD is the term 'lexical richness' (LR) (Tweedie & Baayen, 1998), however, this is often used to mean lexical diversity with a weighting system built in for rarer words (Daller, Van Hout, & Treffers-Daller, 2003).

Very much related to LD is textual *difficulty*, as in novels such as *War and Peace* and *Ulysses* which score highly by any measure of LD while generally credited in

literature circles as among the most challenging of novels. Similarly, casual conversations, not generally considered to be textually difficult, offer very low LD scores relative to other registers. LD is also related to, but distinct from *lexical knowledge*. Differing linguistic purposes lead to differing degrees of LD, so we cannot necessarily extrapolate from someone's output in one register what their overall vocabulary knowledge may be.

In sum then, LD is concerned with *productive* vocabulary rather than *receptive*; LD measures the *range* of lexical deployment. LD is *mode* and *purpose* dependant, meaning that LD scores alter depending on the purpose of the text, and the medium through which the text is produced. And LD is a specific, *verifiable measure* of lexical deployment where higher is often, but by no means always, indicative of greater range or skill. These features mean that LD is a *quantitative* rather than a *qualitative* measure of text.

Why the Need for a Quantitative Measure of Text?

Intuition tells us that the vocabulary used by Shakespeare is far more diverse than the vocabulary used by, say, a freshman composition student. Intuition also tells us that spoken language is less diverse in terms of vocabulary usage than is written language. It also tells us that 5th graders vocabulary diversity is generally less diverse than 6th graders, which in turn is less diverse than 7th graders, and so on. The problem, however,

is that while opinions on the quality of a text are clearly useful, such opinions could be greatly enhanced and better understood if they were tested empirically, using quantifiable measures (Lee et al., 1995).

Quantifiable knowledge provides an objective and verifiable approach to assessing texts. As a consequence, predictions can be formed and texts can be judged scientifically, leading to the possibility of improvements in education methods, testing, and student placement. LD alone is by no means the answer to everything, but as one quantifiable measure, often used in conjunction with many other measures, LD offers an extremely useful tool whose long history and broad application make it extremely valuable to researchers and teachers.

What Is the Problem With Lexical Diversity?

Essentially, the problem of LD is best explained by reference to Heap's Law (Heap, 1978). Heap's Law applies to many fields; however, in linguistics, Heaps' law refers to the fact that the more text that is gathered, the less is the likelihood of new items appearing. As such, the first few words of any given text are likely to be unique, unused items, whereas the 1000^{th}, 1001^{st}, 1002^{nd} etc. are likely to have been used before. The upshot of this is that LD measurement is confounded by length of text: the longer the text, the more likely there are to be repeated items. And as the greater the repetition of items, the less diverse the text appears, it is extremely difficult to know to what extent an

LD measure is measuring vocabulary diversity, and to what extent it is measuring the length of text. This problem for LD measurements effectively means, therefore, that we can only be certain of diversity differences when we measure texts relatively: that is, when we compare texts of equal lengths.

What Is the Purpose of This Dissertation?

Lexical diversity itself is not a linguistically disputed concept, but LD is only a convenient, over-arching concept which manifests itself as a measurement in a large variety of forms. These forms are the various methods through which LD is calculated, each of which has its advantages and disadvantages, but each of which has been found to be less reliable than was hoped. When speaking of LD, then, we must talk in terms of measures of LD, with each measure having its supporters and its reasons for being supported, along with its distracters, and reasons for being questioned.

The problems associated with the methods through which LD is measured mean the quest of researchers in lexical diversity, and the goal of this dissertation, is to establish the range and usefulness of existing LD measures, and, if necessary to establish a new LD measure that avoids the problems of existing measures.

What Special Terms Will Be Used in This Dissertation?

In this dissertation, LD will be used as a covering term referring to an idealized measure of the diversity of vocabulary deployed in a text. The problem of LD, or the "inherent flaw" of LD as it is often called (Malvern, Richards, Chipere, & Duran, 2004) will refer to Heap's Law described above: that LD measures struggle to separate the effects of text length from the diversity of vocabulary in the text. Two further terms, *reliability* and *sensitivity*, will feature often in this dissertation and require some explanation as their use here is not consistent across the literature.

For reliability, we will mean that the method of calculation used by any LD measure provides a score of LD that is not a function of the length of the text. Reliability, in this sense, is established by calculating the LD of parts of a text and comparing the sum of those parts to the whole. Such a method for calculating LD reliability has long been accepted in this field (Hess, Sefton, & Landry, 1986).

Sensitivity, in this dissertation, will refer to the need for an LD score to be sufficiently diverse to be useful for researchers and teachers. For example, a binary measure LD score would be less useful than a continuous measure. Similarly, an LD score that can distinguish between lower-moderate diversity and higher-moderate diversity is more informative than is one which groups both scores as merely moderate. The greater the meaningful range of the LD score, therefore, the more sensitive we shall call the measure.

How Is This Dissertation Structured?

Chapter 2 gives a history of the many varieties of LD measures that have been published. Beginning with the most traditional and widely used measure of LD (Type-Token Ratio), the chapter shows the advantages of each measure, the applications which have used such measures, and some of the fundamental problems with each of the measures.

Chapter 3 tests 13 different published measures of LD on 23 genres of 18,000 words each. Although some measures show some success, and some measures suggest usefulness over particular ranges or under particular conditions, we shall see that each measure is influenced by text length.

Chapter 4 introduces a new measure of LD: the measure of textual and lexical diversity (MTLD). This measure is subjected to the same battery of tests as were the 13 LD measures in Chapter 3. MTLD passes all tests of reliability and is then tested for sensitivity. In this test too, MTLD is successful.

Chapter 5 is a test of the application ability of LD measures in the field of applied linguistics. The tests feature MTLD and the best performing measures from Chapter 3 against a corpus of essays from native and non-native speakers of English. The first object of this chapter is to establish convergent validity between MTLD and other more established LD measures. Having established this, we investigate the ability of the best performing LD measures to differentiate texts in terms of native/non-native speakers, years of English studied, and grade level of student. The results suggest that the greatest

increase of LD occur within the first two years of language learning, following which, LD steadily stabilizes before plateauing at around six years of English. Results also suggest that some LD measures, often working in combination, can distinguish the relevant grade level of students of English up to the limits imposed by the parameters of the corpus.

Chapter 6 concludes this dissertation. We begin by summarizing the findings and then suggest future research that has arisen as a result of the studies conducted in this dissertation. We also speculate about other LD methods that may be used in the future. Given that the best way to understand LD may be by conducting an analysis through a basket of measures approach, we outline some other possible means of calculating or viewing LD.

What Is the Scope of This Investigation?

The problem of LD is most apparent over shorter texts, particularly those of around one-hundred words to a few hundred words in length. For longer texts, the problem of LD persists; however, the effect of the inherent flaw is not nearly so pronounced (Malvern et al., 2004). In this dissertation, therefore, our investigation will focus on the shorter end of text lengths: from 100 to 2000 words in the first study (Chapters 3 and 4), and then 100 to 400 words in the second study (Chapter 5). Practical restrictions on existing measures mean that we shall hold the minimum length of text in

this investigation to 100 words. The formula of certain measures, for example the LD measure known as D (Malvern et al., 2004) simply cannot operate under 50 words, and its creators do not exhibit confidence in its output until around the 100 word level (see Chapter 2).

We would also like to make clear that the investigations in this study in no way try to cover every avenue or possible use of LD. The tests conducted here will merely point out the general weaknesses of existing LD measures over limited texts or limited size. Further tests will suggest that the measure MTLD appears to offer greater reliability and sensitivity than existing measures, especially if used in conjunction with existing measures; however, the wide and varied applications of LD mean that a vast array of further testing will be needed in order to establish the limitations of MTLD.

What Is the Importance of This Study?

While this dissertation cannot definitely assure researchers and teachers that MTLD is the best available measure of LD, the evidence its puts forward will be a strong indicator that such a method will become a valuable tool. Such a tool is undoubtedly needed, as LD is one of the most common metrics used in textual analysis.

Lexical diversity is used in fields which vary from stylistics to neuropathology and from second language acquisition to forensics. In *stylistics*, for example, authorship of historic texts can be identified using differences in LD across authors (Smith & Kelly,

2002): In *neuropathology*, LD can offer help in detecting the onset and development of Alzheimer's (Bucks, Singh, Cuerden, & Wilcock, 2000): In *language acquisition*, the lexical development of children can be assessed (Singh, 2001): In *forensics*, we may even be able to tell whether witnesses are fabricating statements (Colwell, Hiscock, & Memon, 2002).

LD has long been a subject of investigation. Skinner (1937) made a ranking of words based on frequency of use. Skinner used this table to establish patterns based on word associations. One year later, Carrol (1938) introduced what appears to be the first mathematical method of measuring LD, known as *k*. Carrol's k is a constant derived from a "diversity curve," itself formed from a relationship between total number of words used, and number of different words used, to measure the diversity differences in samples of student writings.

Since that time, LD has appeared as a measure in a great range of studies, from studies of dementia (Bucks et al., 2000) to studies of personality type and their effect on composition processes (Carrel & Monroe, 1993); from studies on writing styles (Youmans, 1991) to studies on military morale (Carpenter & Hersh, 1985); from studies of mother and child interaction (Phillips, 1973) to studies of the effects of socioeconomics on language output (Randsell & Wengelin, 2003); and from cross-cultural, cross-linguistic studies of bilingual communities (Daller et al., 2003) to comparisons of various psychotherapy sessions (Reynes, Martindale, & Dahl, 1984). Such diversity of use suggests that a reliable and sensitive measure of LD is much needed, will be much applied, and may lead to a wealth of information.

For a greater appreciation of LD's diverse application it is well worth considering the range of LD use across the literature. As such, we will now look at the use of LD in such studies as second language acquisition and bilingualism; gender, dialect, and IQ; first language speaking, writing and reading; speech and hearing for those with specific language impairment (SLI) and psychotherapy.

LD Used in Second Language Acquisition and Bilingualism

In studies of bilingualism, Daller et al. (2003) investigated the German used by German-born Turks living in Turkey, against the Turkish use of German-born Turks living in Germany. Daller et al. used teachers and frequency lists to rate the difficulty or rarity of lexical usage across a variety of LD measures. The addition of these lists helped these researchers to make comparisons across two different languages, in this case establishing that LD was able to establish characteristics of bilingual speakers in non-bilingual environments. Such a study is generally confounded by the fact that different languages (especially languages of different families) have naturally occurring differing LD rates. For instance, the genitive in German is marked by a separate word, whereas in Turkish the genitive appears as a morpheme. Through such studies using LD measures, the rate and areas of language decay in bilinguals can be formulated and assessed.

Ransdell and Wengelin (2003) used LD as one of their measures in a study of whether socioeconomic or sociolinguistic factors were the best predictors of children's

L1 reading and writing ability. Ransdell and Wengelin tested one hundred 4th grade Florida students, a third of whom were bilingual, and found that socioeconomic status appeared to play a more important role in determining children's writing and reading levels than other sociolinguistic factors. While Ransdell and Wengelin did find that the bilingual children lacked certain English skills, these were attributed to less L1 exposure rather than as a result of a bilingual environment. More importantly for LD, Ransdell and Wengelin concluded the scores obtained from lexical testing showed that low socioeconomic status children appeared to be every bit as good as high socioeconomic children in terms of their ability to vary vocabulary usage. Ransdell and Wengelin's research is helping to establish a "Sociolinguistic Status" (SLS) index which stands to offer strong evidence that multiple language exposure is far from detrimental to the cognitive development of children.

Carrel and Monroe (1993) included LD as one of their measures for assessing the learning styles of three kinds of composition students. Carrel and Monroe split the group of student writers into 'traditional freshman composition students,' '"basic" (sic) writers', and 'non-native ESL freshman composition students.' The subjects were all given the *Myers-Briggs Type Indicator* (MBTI) and, following the completion of their composition courses, were tested on their final products of composition. Of interest to LD measures, Carrel and Monroe's results suggested that non-native writers whose personality type rated *high* on 'feelings' produced work of greater lexical diversity; and non-native speakers who rated *low* on the 'thinking' scale also produced high rates of lexical diversity. Similarly, non-native students also produced good LD results if their personality type indicated *high* 'intuition' or *low* 'sensing.' From this, Carrel and Monroe

concluded that non-native students of English whose personality types reflected *intuition*, *feeling*, and *perceiving* were better able to express a wider range of vocabulary in their writing. Such evidence lends support to various claims (for example, Kirby, 1988; Lawrence, 1984) that individual learning styles play a significant role in the effectiveness of pedagogical approaches.

Malvern et al. (2004) conducted a study of a French oral examination by native English speakers using the transcripts collected from both teachers and students. In that study, LD was compared to several other measures of language proficiency including the teachers' assessment of *range of vocabulary*, *complexity of structures*, *content*, *accuracy*, and *pronunciation*. The study was of particular concern to language learning and testing as much discussion concerning oral interview techniques and assessment is currently being conducted. For example, Malvern et al. cite He and Young (1998) who question whether oral assessments can resemble natural conversation. Van Lier (1989) also points out, the "power differential" between interviewer and interviewee and the elicitation demands of the situation confound the naturalness of the interchange. According to Young and Milanovic (1992), the testers contribution is structured towards the goals of the exercise, whereas the contribution of those tested is more reactive, again limiting the naturalness of the conversation. Malvern et al. acknowledge that the interviewer is bound to have greater control of topic and turn-taking, once more reducing the effectiveness of the interchange to reflect true conversational ability (Johnson & Tyler, 1998; Lazaraton, 1992; Moder & Halleck, 1998).

The particular examination transcripts used for Malvern et al.'s (2004) study was the British GCSE for French (all GCSE examinations are typically taken at around the

age of 16). One important feature of this examination is that the examiners simultaneously conducted interviews as they scored the candidates. Furthermore, the examiners of the students were typically the students' own teacher. Malvern et al. acknowledge that such procedures lead to results that may not scale up, but as such an examination exists on a national level, the investigation is important.

The results of Malvern et al.'s (2004) study showed no significant correlation between LD and the teachers' rating of *Range of Vocabulary*, the latter being only the *impressions* gained from the interview. Malvern et al. conclude that the teachers' ratings are almost certainly affected by "halo effects" and the "sheer difficulty" of simultaneously testing and rating students on elements such as apparent richness of vocabulary and use of less-common terms. LD, of course, as an independent, verifiable measure, would not have been similarly affected.

A further major finding of this study was that LD scores were actually higher for students than for teachers. Malvern et al. (2004) conclude that this is probably due to teachers oversimplifying their lexis for the benefit of the students. This finding, however, was not supported by individual analysis of teacher to student interaction. Instead, Malvern et al.'s results indicated that teachers were gearing their lexical diversity at a class level. As such, there was no indication of teachers accommodating individual students by, for example, simplifying their language for weaker students, but rather that teachers had a predefined notion of the ability of their classes and geared their speech to the general level of the group.

Malvern et al.'s (2004) conclusions from this study may offer important lessons for assessment of foreign language students. Firstly, it appears that rating students while

interviewing them, whether analytically or holistically, may invite halo effects. Secondly, while some accommodation at a class rather than individual level may be beneficial for certain students, an interviewer who has no prior knowledge of the students' abilities may provide a more even playing field through which to assess a student's proficiency.

LD Used to Show Differences Between Groups

Dizney and Roskens (1966) used a variety of LD measures to differentiate high IQ children within a chronologically aged control group. Their results suggested that "bright" children tended to use more words, repeat those words more often, qualify their nouns, and speak faster. The girls in this study are reported to produce a slightly higher diversity of lexical usage, although Dizney and Roskens report the difference was not significant.

Singh (2001) studied the conversational speech of thirteen males and seventeen females to see if lexical diversity measures were able to differentiate the genders. Feldstein, Dohm and Crown (1993) had earlier shown differences in speech rate, but statistically significant differences in language use had not previously been gained. Singh's study concluded that LD measures were able to "easily" separate the groups with up to 90% accuracy. The major differences were reported as females producing more verbs, shorter sentences, and greater repetition, whereas males were reported as being overall lexically richer in their output. While such findings are acknowledged as being

tendencies, and though Singh acknowledges the study's limited number of subjects, such findings are also reported as being important for teaching methodologies to different genders and to differing treatments for aphasiac patients.

Biber (1987) used LD as one of 40 measures for differentiating American English from British English. Biber used corpora of British and American texts from 15 different registers, concluding that British English is more edited, less interactive, and less abstract than American English. The reason for such differences Biber ascribes to the greater tendency for British speakers and writers to subscribe to prescriptivist styles.

LD use in forensic studies has developed over many years and has its origins in assessments of "language style." For example, Osgood (1960) used LD as a predictor of emotional or highly motivated output. Osgood examined suicide notes and compared them to *pseudocide* notes predicting that greater motivational levels would lead to more frequent use of high frequency words. The consequence of this would be lower diversity. Such a prediction indeed turned out to be the result.

Carpenter and Hersh (1985) used LD to indicate how military intelligence may gain information about enemy troop morale from intercepted texts. In this study, Carpenter and Hersh acknowledged that the military use classified codes for most of their communication, but that personnel communications, though censored, are often intercepted and able to be stylistically analyzed. Carpenter and Hersh were not so much interested with explicit statements within communications, such as locations and troop numbers, as they are with the form of the communication and how that may reflect the author's morale. If morale can be determined from the form of correspondences then the enemy, Carpenter and Hersh argue, may be able to better judge the prudence of their own

actions. Carpenter and Hersh's study analyzed the letters of two British officers written during the American War of Independence and attempted to show that LD increases as the morale of the authors decreases. Carpenter and Hersh concluded that LD could be treated as a "stylistic index of deteriorating morale" that might be a gleamed from a text without even the author knowing of its presence.

Smith and Kelly (2002) investigated authorial *styles* using a variety of LD measures. Their stylometric investigation was concerned with dating ancient works by comparing their LD scores with the LD scores of works whose dates are known. Smith and Kelly's study focused on Greek and Roman playwrights: Euripides, Aristophanes, and Terence. Through the results of LD scores, Smith and Kelly were able to conclude that Euripides and Terence produced lexically richer works as they matured, whereas Aristophanes' work lessened in its variety over time. Smith and Kelly concluded that many other undated works, or works of uncertain date, could be judged using similar methods.

Such investigations of textual styles through quantitative measures such as LD have also strongly influenced a development in linguistic forensics. Colwell, Hiscock and Memon (2002), for example, use LD as one of their determiners for statement fabrication. In their test, of which LD was a primary measure, as many as 93% of fabricated statements were detected. As a consequence, Colwell et al. were able to conclude that eyewitness testimony can be far more beneficial when analyzed through quantitative as well as qualitative measures.

LD Used in First Language Speaking, Writing and Reading

LD has long been used in the field of first language development. Bondy Bougere (1969), for example, used LD measures to help assess to the development of first grade students' reading abilities. As lexical diversity measures were shown to significantly predict word recognition achievement, Bondy Bougere was able to recommend the addition of LD measures to teacher assessments of student reading skills.

In Phillips' (1973) study of mothers' speech to young children, Phillips offered strong support to Snow's (1972) claims that the speech of mothers to children was "simpler and more redundant than their normal speech." In her study, Phillips used transcripts of mother child interaction, the children forming three groups (8-month, 18-month, and 28-month olds). Phillips analyzed transcripts for 10 linguistic features of which LD was one. The results offered support to both her hypotheses: adult speech differs depending upon whether the addressee is a child or an adult, and that adult speech increases in complexity and diversity as the child's speech develops. In fact, the measure of LD was one of the more significant indicators of variance in speech with marked levels of increased diversity across each of the three groups of children.

Much more recently, Lee et al. (1995) used LD as one of their *quantitative* measures to help assess the *quality* of children's writing. In their study, 9 year-olds were asked to either free-write on a theme or to *re*-write a story that had been told to them (i.e., to produce work from a scaffolded prompt). Correlations with independent holistic assessors showed the re-write condition - measured in terms of cohesive indices, mean

length of utterances, composition length and LD - produced higher overall writing quality. Such empirical and quantifiable evidence helps to provide support for such arguments as scaffolded assignments.

LD Used to Help Evaluate Speech and Hearing for Those With Specific Language Impairment

Ertmer, Strong, and Sadagopan (2002) studied the speech progress of a young, deaf girl who has been fitted with a cochlear implant. Their work included lexical diversity measures and compared the child's spoken output with that of normally developing children. Ertmer et al. pointed out that the optimum age for cochlear implants is a hotly disputed issue (see, for example, Bollard, Chute, Popp & Parisier, 1999; Conner, Heiber, Arts, & Zwolen, 2000) so case studies of the development of children are much needed.

Stokes and Fletcher (2000) used various approaches to LD to study the productivity of Chinese-speaking children with SLI (specific language impairment). Interestingly, Stokes and Fletcher's study found no difference between the impaired group and the normally developing group in terms of verb use, but a significant difference between the groups in noun use. Such findings suggest that language impairment may affect some parts of speech far more than they affect others.

In an analysis of the conversational speech of patients suffering from dementia of Alzheimer type Dementia (DAT), Bucks et al. (2000) used eight LD methods to help shed light on the conventional wisdom associated with the fluency of DAT patients. Bucks et al. describe dementia as "the breakdown of intellectual and communicative functioning accompanied by personality change," and that it is often characterized by word-finding deficits, impaired performance on verbal fluency tasks, circumlocutionary responses, and discourse impairment issues. Such linguistic problems lead Bucks et al. to theorize that LD measures may provide earlier diagnosis of DAT. Bucks et al. also theorized that LD measures could evaluate patients more discreetly than formal psychological tests, that the measures may increase diagnostic specificity, and that prognosis, interventions, and treatment may also be assisted by better patient assessments. Bucks et al.'s experiment compared the speech of eight DAT patients to sixteen healthy patients acting as control. Using LD measures to assess transcripts of conversations, Bucks et al. results showed that LD measures correctly predicted all 24 subjects' conditions as either DAT or non-DAT. Based on this success, Bucks et al. speculated that LD measures might enable better understanding of how, why, when, and to what degree lexical changes correlate with semantic memory breakdown.

Thordardottir and Weismer (2001) used LD measures to study the use of high frequency 'general all purpose' (GAP) verbs used by children with SLI. Thordardottir and Weismer reported many findings that suggest SLI for children leads to problems with verb use (for example, Kelly, 1997; King & Fletcher, 1993; Watkins & Rice, 1993). Such difficulties are often related to what is seen as an over reliance of a small group of semantically ambiguous verbs known as GAP verbs. Comparing a SLI group to a

normally developing group of children, and using LD as one of their determiners of usage, Thordardottir and Weismer were able to find that GAP use is indicative of all children and, consequently, would seem to be a normal developmental stage. Moreover, Thordardottir and Weismer argue that GAP verbs may stand as prototypes for more complex later emerging verbs.

LD Used in Psychotherapy

In psychotherapy studies, LD was used by Reynes, Martindale, and Dahl (1984) to compare patient language use under various types of therapy sessions. Reynes et al. used transcripts from 25 psychoanalytical sessions and divided the sessions into three kinds of language output: *working*, *neutral*, and *resistance*. The resistance sessions focused on the patient exhibiting speech disturbances whereas the working sessions focused on areas of the transcripts in which the patient was seen as trying to deal with issues or work through them with the analyst. The neutral sessions fell on a continuum between the two extreme cases. Reynes et al.'s results offered evidence that LD was higher during the working sessions. These working sessions produced a far higher count of "primary process" speech, involving words with connotations more geared towards the ego: connotations such as drive, sensations, and defensive symbolization. Thus, Reynes et al. were able to claim that psychoanalytic working sessions enabled patients to express greater egocentric issues which, they argue, can be supported by the higher scores of LD.

Summary

In this chapter we have outlined what LD is, who uses it, and what its uses are. We have also introduced the inherent problem of LD: that text length plays a significant role in measurements of LD rather than text content. We have outlined what a sound LD measure needs to provide researchers and teachers: reliability and sensitivity. That is, a LD measure must avoid being confounded by text length, and must provide a score of LD that broadly differentiates texts. We have outlined the structure of this dissertation, indicating that the great majority of published LD methods will be tested, that a new measure of LD will be introduced, and that the best performing measures of LD will be tested for their application in the field of applied linguistics. We now move to Chapter 2 and introduce the rich and varied history of LD calculations.

Chapter 2: Established Measures of LD

In this chapter, we will see why an accurate measure of LD is no simple calculation. We shall be introduced to each of the major measures that have attempted to produce a reliable score for lexical diversity, and we shall see the pros and cons of each of these measures. We shall also see that attempts to measure LD can be placed into a four broad categories, each of which has produced different approaches to solving LD. We shall ultimately see that even the latest technique for solving LD brings with it a wide variety of possible problems. We begin by considering the most widely used and possibly most problematic measure of LD: Type-Token Ratio (TTR).

What Is TTR?

Type-token ratio (TTR) has long been the traditional method of measuring LD (Templin, 1957). Types are the unique lexical elements used in the text; for example, in the sentence "The big boy hit the small boy," the types used are *the*, *big*, *boy*, *hit* and *small*. Tokens are the individual instances of lexical items used in the text. Thus, in our example, there are seven words in total, or seven tokens: *the, big, boy, hit, the, small,* and *boy*. 'Type-token ratio' is the division of types by tokens forming a measure that ranges

from 0 to 1 where a higher number indicates greater diversity. In this example, the TTR would be 0.71.[1]

While TTR may look like a practical way of measuring vocabulary use, the finite tokens for infinite purposes aspect of language becomes a fatal problem for actually measuring the degree of diversity in a language sample. While a text may be any length, the vocabulary available is always finite. Therefore, as the text proceeds, the likelihood of tokens being repeated increases, meaning that longer texts will nearly always have lower TTRs, and shorter texts will nearly always have higher TTRs. Thus, as the text length increases, TTRs will fall, forming a hyperbolic curve which, if the text were long enough, would eventually end at a point of zero. The problem with TTR, therefore, is that the more language in the sample, the lower our score of language diversity becomes. This, of course, is misleading at best.

The underlying problem with the assumption behind TTR may best be described by analogy. A typist's speed can be measured by words per minute: as the minutes increase, so does the number of words. Any given minute will produce *roughly* the same amount of words, and an average of total words to total minutes gives a very good estimate of the speed of the typist. TTR is quite different. While the number of words increases linearly (just as with the minutes for the typist), the frequency of introduced types slowly decreases. Eventually, a speaker (or writer) has simply used up the range of words available and every new production is merely a repetition of words used before. Consequently, after just one word, the writer or speaker appears to be demonstrating

[1] This example, it needs to be noted, is by itself not ideal. LD scores for any given sentence may vary greatly and, as such, measuring a text of less than, say 50 words, cannot tell us very much about the lexical ability of the writer/speaker.

maximum diversity, but each subsequent production generally lessens this apparent diversity.

TTR history can be traced back to the 1940s when quantitative indices for both writing and speech began development (Johnson, 1944). Considering types and tokens as a meaningful ratio, however, is generally attributed to Templin who noted the children in her experiment produced " ... approximately one different word for slightly over two words uttered" (Templin, 1957, p. 115). While Templin only made a cursory note of ratios, it was left to Miller (1981) to examine the data she produced with mean TTR scores. It is through this analysis that Miller produced what Malvern et al. (2004) believe to be LD's most spurious conclusion: " ... If a normal hearing child's TTR is significantly below 0.5 we can be reasonably certain that the sparseness of vocabulary ... is probably indicative of language specific deficiency" (Miller, 1981, p. 41)[2].

Though Templin's data and Miller's analysis became seminal works in LD, the conclusions Miller drew are obviously troubling. If, for example, we consider that Shakespeare's most "diverse" work (Macbeth) barely raises above a TTR of 0.2 and Joyce's *Ulysses*, often considered the greatest work of literature, actually falls below 0.2, then, according to Miller, both works are "probably indicative of language specific deficiency." Clearly, Miller did not recognize that text length greatly confounds TTR. Such knowledge was, in fact, published as early as 1944 when Chotlos wrote "Usually type-token ratios are not directly comparable unless they are based on the same number of tokens for each individual." More detailed evidence of text length as a function of TTR

[2] Miller (1991) acknowledged the inherent flaw of TTR, although the measures he subsequently used can also be problematic.

was also published shortly after Templin's work (for example, Carrol, 1964; Guiraud, 1960). The failure of Miller to notice these works, coupled with his inviting 0.5 thresholds have lead to a quarter century of research based on highly inaccurate standards. For example, Stickler (1987), Layton and Savino (1990), and McEvoy and Dodd (1992) all form their conclusions having considered Templin's data and Miller's derived values as norms. Indeed, moving into the 21st century, TTR's inviting simplicity and relatively long history means it still permeates current language analysis, cropping up without any warning of its problems in studies as recent as Singh (2001) and Avent and Austermann (2003). More worrying still are some of the conclusions derived from Templin's apparent "norms." Ertmer, Strong, and Sadagopan (2003) repeatedly cite Templin's 0.5 norms when assessing the speech development of a young child with cochlear implants, concluding that the child's speech and language skills were different from that of a normally developing child, and that the subject had restrictive vocabulary and/or was verbally less able to express herself than a hearing child of the same age. Studies such as these suggest that TTR is not merely erroneous in what it asserts, but in some cases could even be quite dangerous.

Is TTR Ever Useful?

By the end of 1990s, a plentiful supply of researchers had supplied an overwhelming amount of evidence that raw TTR scores were a flawed measure of LD

(Richards & Malvern, 1997; Tweedie & Baayen, 1998). Such findings, however, do *not* mean that TTR is without any use at all, and it is certainly worth showing how TTR can be used for a number of purposes – including being an important indicator of LD.

Youmans (1990), for instance, suggests LD *differences* by comparing individual TTR curves over a variety of literature. Youmans avoids the problem of TTR by utilizing the pattern of the curves formed: While all TTR curves fall, they fall at differing rates; the greater the rate of fall, the lower the LD. From such analysis, Youmans was able to conclude that Longfellow's LD was greater than that of Hemmingway's – who in turn displayed greater LD than Shakespeare's *Macbeth* (in basic English) and the Bible's *Genesis* respectively. While comparing curves is a strong indicator of LD, such mapping does not provide an adequate *measure* of LD: which is to say, there is no numerically represented *score* of LD.

That said, a graphic representation of curves does allow for a meaningful visual comparison of texts which can tell us a lot about the texts being compared. In figure 1, for example, we see a Youmans-like comparison of four narrative texts: *Twenty Thousand Leagues Beneath the Sea*, *Huckleberry Finn*, *A Tale of Two Cities*, and *The Red Badge of Courage*. *A Tale of Two Cities* is well known for its repetitive opening sequence:

It was the best of times, it was the worst of times,
it was the age of wisdom, it was the age of foolishness,
it was the epoch of belief, it was the epoch of incredulity,
it was the season of Light, it was the season of Darkness,

it was the spring of hope, it was the winter of despair,
we had everything before us, we had nothing before us,
we were all going direct to Heaven, we were all going direct
the other way--in short, the period was so far like the present
period, that some of its noisiest authorities insisted on its
being received, for good or for evil, in the superlative degree
of comparison only.

Not surprisingly, therefore, figure 1 shows this text to drop to a very low TTR of 0.35 within its first 50 words. Only *Huckleberry Finn* drops as low, but that text requires over 300 tokens before doing so.

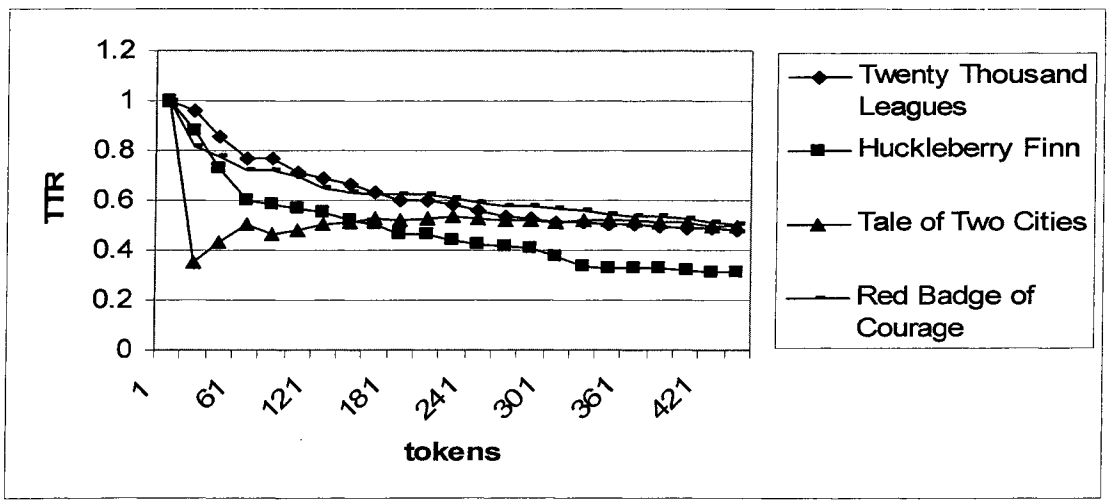

Figure 1. TTR Comparison of Four Texts Showing Differing LD Development Across the Chapter.

After 450 tokens, *Twenty Thousand Leagues*, *Tale of Two Cities*, and *Red Badge of Courage* have converged at a similar TTR: 0.49, considerably higher than *Huckleberry Finn's*. However, the route to this 0.49 is visibly different. *Twenty Thousand Leagues* indicates slightly greater diversity in its earlier stages, *Red Badge of Courage* is a little more repetitive early on, and (as we have seen) *Tale of Two Cities* soon loses its repetitive style and quickly becomes quite diverse.

Such TTR curves suggest that *Huckleberry Finn* is the least diverse text; not surprising as the story is written in the style of simple, spoken language of an uneducated 19th century boy of the south. The diversity of the other three texts, however, is difficult to distinguish as their curves soon merge.

While TTR scores do, eventually, form a long sweeping curve, close analysis of texts such as *Tale of Two Cities* clearly shows that there are areas of the "curve" which have peaks and valleys. These fluctuations can best be seen when TTRs are calculated after *each* type rather than in much larger incidence (Templin's baseline of 50 is a common division, as is Tweedie and Baayen's incidence of 20). A stretch of new types causes TTRs to rise, forming peaks; however, as the text progresses, the incidence of new types decreases and subsequent peaks are almost always smaller and smoother (see figure 2). Eventually, peaks and valleys become almost imperceptible (see figure 3). If TTRs are recorded at an interval of 20 tokens rather than just single intervals, the peaks and valleys would be completely eliminated (see figure 4).

Such peaks and valleys were neatly exploited by Youmans (1991). Youmans used newly occurring lexical items appearing in a moving window to plot what he called a Vocabulary-Management Profile (VMP). Youmans claimed that the VMP could be used

to illuminate textual features such as pragmatic moves (763). While such instances of TTR application are undoubtedly both useful and interesting, suggestions of features and traits are not a *measure* of LD, and far from the *reliable measure* we seek to establish here.

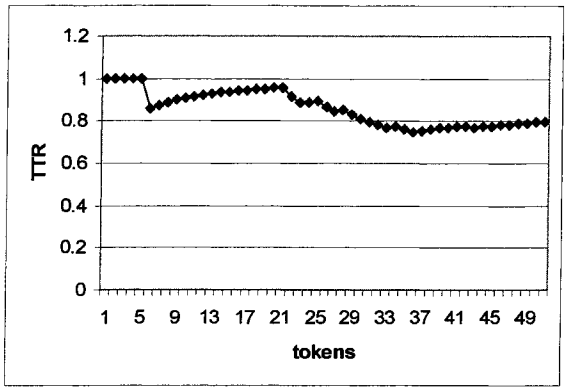

Figure 2. A Series of Peaks and Valleys Formed by TTR Analysis from the First 50 Tokens of Bill Clinton's Inaugural Address.

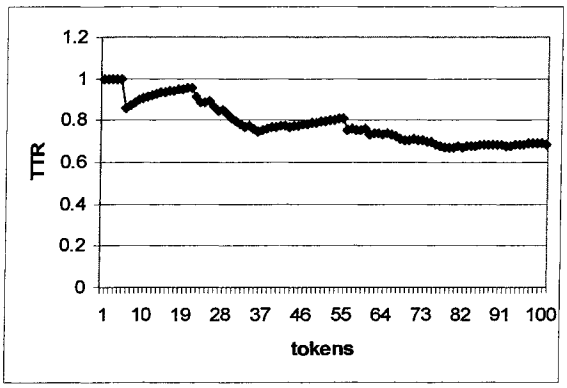

Figure 3. A Series of Peaks and Valleys Formed by TTR Analysis from the First 100 Tokens of Bill Clinton's Inaugural Address.

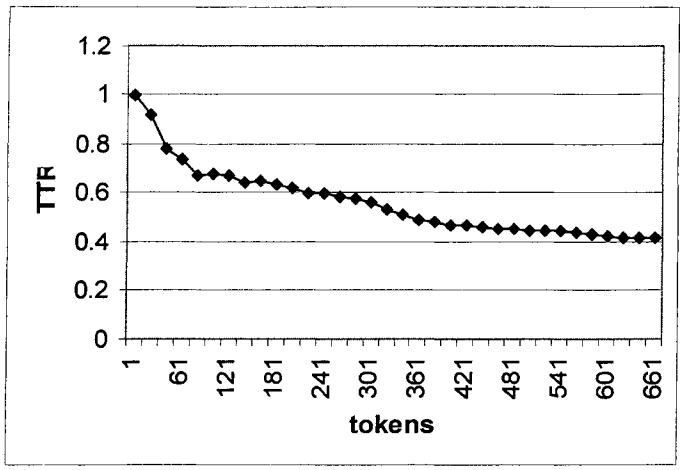

Figure 4. Bill Clinton's Inaugural Address with TTRs Recorded at Intervals of 20 Tokens.

What Are Receptive Measures?

Self-assessment tests, pseudo-word tests, and multiple choice questions form a very different approach to measuring vocabulary breadth. A simple approach for such tests would be to sample words randomly from a dictionary and then have subjects indicate whether or not they knew the words (Daller et al., 2003). At a more complex level, subjects may be given a single word and a list of alternative words. Subjects are then asked to indicate which of the alternative words is closest in meaning to the target word. University examinations such as the GRE, along with many English as a second language tests use just such an approach; however, the problems with such assessments

as a measure of LD are many. Primarily, such tests attempt to measure a subject's vocabulary *range*. Thus, they cannot be used to measure vocabulary deployment whether at spoken or written level. That is to say, they are measures of *receptive* vocabulary, not *productive* vocabulary.

But even as a measure of vocabulary availability, such assessments have received a wide range of criticism. Concerning self-assessment tests, Zechmeister, D'Anna, Hall, Paus, and Smith (1993) report a significant difference between the number of words subjects say they know and actually *do* know. Pseudo word tests attempt to reduce this effect by asking subjects to identify only real words from the non-real words (Meara & Buxton, 1987). Unfortunately, rejecting non-real words and identifying real words are quite different skills, leading to results that poorly indicate vocabulary range (Beekmans, Eyckmans, Janssens, Dufranne, & Van de Velde, 2001). Meara and Buxton (1987), Meara and Jones (1998), Meara (1992) and Read (1993) also offer evidence that non-real words are easier to notice and marking a word as *known* does not mean a subject can use that word correctly. In fact, the very premise of the test (knowing a word) leads away from issues of diversity of use and into questions of what it means to *know a word* (Daller et al., 2003; Verhallen & Schoonen, 1993; Vermeer, 1992)

Just as receptive and qualitative research must be distinguished from the LD measures sought in this study, so does the measure of spoken language known as *mean length of utterance* (MLU). The MLU counts the number of words uttered and averages by utterances. MLU has become a well-established indicator of age and language ability (Brown, 1973; Grela, 2001; Wells, 1985) and is theoretically sound because as normally developing children grow, the mean length of their utterances increase. The problem,

however, is not the usefulness of MLU, nor its accuracy, it is simply that MLU is not a measure of LD per se: MLU makes no account for types or range of vocabulary deployment in its metric. Thus MLU can legitimately be used as a measure of "language development," but is not effective as a measure of "lexical diversity."

Even if we could consider MLU as a *factor* of LD, we would have to recognize its shortcomings. Firstly, MLU is only used for spoken texts (although a similar measure for written texts, Fleisch-Kincaid, is also available). Secondly, the correlation of MLU to age covers only formative years: people do not keep making longer and longer utterances all their lives. Thus, while MLU is affective, its range is extremely limited and subjects of study seldom reach past infancy[3]. For these reasons, MLU is not considered further in this study.

Can the Problem of TTR Be Corrected?

We can now turn to various approaches of *correcting* the inherent flaw of TTR. We can place the many alternative measures, each of which has been offered as a suitable replacement for TTR, into four broad categories of LD: token-deinfinitising (Guiraud, 1954), frequency-addressing (Daller et al., 2003), distribution-addressing (Yule, 1944), and curve fitting (Malvern et al., 1997). We shall now deal with each of these approaches in turn, beginning with *token-deinfinitising.*

[3] Miller (1991) is something of an exception with an upper range of 13-years-old.

The inherent flaw of TTR, as stated above, is that types are finite and tokens are infinite. Thus, many theorists have addressed the problem of TTR by making finite, or *deinfinitizing*, the infinite. The simplest approach in this line of thinking is to report only the number of different words in a text (NDW) and just ignore the total number of words the text contains. Thus the infinite possibilities are reduced to zero. NDW posits that the greater the number of types, the greater the diversity (Ratner & Silvermann, 2000). However, just because the number of tokens has been taken out of the visible equation, it does not mean the number of tokens no longer plays a significant role. Any variance in NDW will depend on how many tokens are available to prize out the types, this applies whether or not the text length is actually reported: simply put, the longer the text, the more *different words* there are likely to be. Such criticism of NDW is common in the literature (Daller et al., 2003; Hess, Haug, & Landry, 1989; Malvern et al., 2004), yet this has not stopped some researchers from including it in their studies (Klee, 1992; Ratner & Silvermann, 2000).

Like TTR, however, NDW is not without its uses, as it serves to indicate the breadth of the used vocabulary (Malvern et al., 2004; Stokes & Fletcher, 2000). NDW has also served to suggest some notable findings. Klee (1992), for example, studied children whose ages ranged from 24 to 50 months. NDW not only correlated with age, it also managed to distinguish normally developing children from children with specific language impairment. Wright, Silverman, & Newhoff (2003) also offered some support of NDW, arguing that in child language studies it is the "preferred measure." They also argue that NDW is less sensitive to text size variations than TTR: the latter having to cope with the linear increase of tokens, whereas the former simply rises more and more

slowly. That said, Wright et al.'s overall conclusion is that NDW is not a good measure of LD as it is only stable with truncated texts, and that truncated texts lead to erroneous analysis.

A second approach, within this category, is to make the number of tokens finite by setting a time limit to the data collection. Disney and Roskens (1966) are the earliest examples we could find of such a method. They used 10-minute chunks of conversation from which to gleam transcript data. In recent years, time chunking has become quite popular as a standardization for TTR (Avent & Austerman, 2003; Geers, Spehar, & Sedey, 2002). The theory appears to be that if subjects are all allowed the same amount of time then differences in the number of tokens become a factor of productivity. This utilizing of productivity, the advocates argue, makes the method stronger (Geers et al., 2002). Avent and Austerman (2003) use such thinking to study the effects of reciprocal scaffolding treatment (RST) against discourse group treatment on a patient suffering from aphasia. Their positive conclusions for RST were largely based on the TTR (60 minute time chunks) findings.

Using time allotments to standardize TTRs may be common, but criticism of such methods is even more so: As age and language ability increase, so does the quantity of lexical output (Wells, 1985). Thus, if language ability were improving, productivity would increase, and TTRs would go down. If TTRs went up, it could only mean that productivity was decreasing. In Avent and Austerman's (2003) study, the subject produced a mean token count of 419 for RST, whereas he reached 999 tokens during discourse group treatment. Austerman and Avent note this difference, but claim that as the TTR was higher for RST that this context was the more powerful. Consequently,

Austerman and Avent's conclusions can be brought into question. Interestingly, this possible error stems from what appears to be an erroneous reading of Malvern and Richards (1997). The former cite the latter as justification for the time method they deploy; however the latter's only comments on time chunks reads "The problem [with time chunks] is that lexical diversity and volubility are confounded: children receive higher scores simply because they talk more": Malvern and Richards, therefore, can hardly be said to be issuing a ringing endorsement of the time-chunk approach.

Another attempt to standardize TTR through limitation of tokens is premised on the following *correct* assumption: if all texts are the same length then TTR is an accurate measure of diversity. This being the case, many researchers have gone on to conclude that the next logical step is to make all texts the same length. To this end Besnier (1988) uses the first 500 tokens of a text, Biber (1988) the first 400 words, Phillips (1973) used the first 300 tokens, whereas Bucks et al. (2000) interrogate their subjects until they have prized out a nice, round 1000 tokens. Singh (2001) used TTR as one her LD measures for differentiating gender groups and concluded that the results did not correlate significantly with text length. Singh's conclusions are, no doubt, quite plausible; however, as all subjects in her group were kept talking until they produced between 1000 and 1200 words, it is not surprising that there was no correlation to text length (there being no great difference between text lengths to correlate to).

Such restricting or expanding of data as that described above is necessary for a uniform analysis of TTR; however, many problems occur as a result of it. Firstly, if we consider that a natural text may be many hundreds or many thousands of words in length (and given Youmans' peeks and valleys illustrated above), an analysis of just the first 400

tokens leaves the vast majority of text discarded. Secondly, researchers seldom choose a uniform number of tokens to analyze, so comparing findings becomes difficult or impossible. Thirdly, even when texts have the same number of tokens, we are not necessarily comparing like with like. For example, consider a situation where Text A has 500 tokens, Text B has 600 tokens, and Text C has 10, 000 tokens. A TTR analysis, using the same number of tokens, would presumably select 500 tokens as the maximum amount of data (that is, the length of the shortest text). Such analysis would, therefore, consider *the whole* of Text A, the vast *majority* of Text B, but barely the introduction of Text C. Considering the highly repetitive nature of certain texts or passages of texts (for example, *A Tale of Two Cities*) we may end up not representing these texts at all accurately.

A more sophisticated standardization of TTR is known as mean segmental type-token ratio (MSTTR). MSTTR has a history that can be traced back as far as Johnson (1944), and, as Malvern et al. (2004) report, MSTTR has appeared in a wide range of investigations: for example, in students' L1 writing (Mann, 1944), in studies on schizophrenia (Manschreck, Maher, & Ader, 1981), in aphasiology (Wachal & Spreen, 1973), and in second language learning (Meara, 1978). MSTTR has also been used to help date previously undated work from ancient Greek Playwrights (Smith & Kelly, 2002) although in this case the LD measure is referred to as "partitioning," a method the authors believed they had discovered.

Within the research mentioned here, MSTTR's findings appear to have been successful. For example, the measure was able to discriminate aphasiac patients from

control groups (Malvern et al., 2004) and was also used to posit some serious questions as to the dates of redrafted works from Aristophanes, Euripides, and Terence (Smith & Kelly, 2002).

MSTTR divides texts into sections of equal size, say 100 tokens, and discards any remaining data. The TTR for each section is then recorded, and the mean of each section forms the final score. Section sizes are generally decided by the length of the smallest available text. Thus in Malvern and Richards (2000), a MSTTR (30) was used. On other occasions, outlying small samples are discarded and a "round figure" such as 100 is preferred (e.g., Carpenter & Hersh, 1985). At the high end, Chotlos (1944), having texts that each exceeded 3,000 words, was able to use MSTTR (1000). Bowker and Pearson (2002) also recommend 1,000 word chunks[4] in their introductory book on working with corpora.

For all its popularity, however, MSTTR is not without its problems. Firstly, the non-standardization of section size means that results cannot be compared. That is, there is no way to translate a MSTTR (100) into a MSTTR (1000), or *visa versa*. Malvern at al. (2004) also criticize MSTTR for discarding some data, even while they acknowledge that comparatively little data is left unused. A third problem lies in the fact that tokens may be often repeated, but may fall in different sections, and therefore would give higher diversity scores for lower MSTTRs (Malvern et al., 2004). And a fourth problem is simply getting data when it does not exist: Bowker and Pearson (2002), for example,

[4] Bowker and Pearson refer to MSTTR as "standardized TTR." Jarvis (2002) refers to same method as "split TTR."

invite their readers to slice texts into 1,000 word chunks, but they do not explain how to get 1,000 word chunks from texts that are less than 1,000 words long.

While each of these problems forms fair criticism, the major problem with MSTTR is a theoretical one. As MSTTR is reliable regardless of its length, provided all lengths are equal, the best length to choose is the shortest length (which allows for the least amount of discarded data). As such, given Text A of 30 words, Text B of 100 words, and Text C of 500 words, a MSTTR(30) would be the only size capable of including all texts. In fact, there is no reason why even smaller MSTTRs could not be used. MSTTR(1), for example, would include all possible text sizes and allow for no discarding of data. Unfortunately, a MSTTR(1) would always produce the same result, a TTR of 1.0. Smaller MSTTRs, therefore, cover all texts and most data, but produce very little variance, and consequently, very little discrimination, which means they are do not supply a "sensitive" score of lexical diversity. Larger MSTTRs, on the other hand, cover very few texts and discard a great deal of data, though they do produce clear variance leading to highly sensitive scores. We can argue, therefore, that as the accuracy of MSTTR *increases*, so does its ability to discriminate LD *decrease*. As such, we cannot consider MSTTR a sensitive measure of LD. And by the same token, as the sensitivity of MSTTR *increases*, so does its reliability *decrease*. As such, we cannot consider MSTTR a reliable measure of LD

If sample size cannot correct for the flaw inherent in TTR, then maybe mathematical *correction* of token count can (Carrol, 1964; Guiraud, 1960; Herdan, 1960).

Over the years, numerous such attempts have been made with each study acknowledging the TTR curve and formulating an equation to compensate for that curve. Unfortunately, as we shall see, none of these formulas passes close inspection.

Guiraud (1960) proposed root TTR (RTTR) and Carroll (1964) proposed corrected TTR (CTTR). Both believed the decline in TTR could be factored out by applying the square root of the total tokens. Guiraud's measure divided types (V) by the square root of tokens (N). Carroll produced much the same result although he first divided tokens by two before finding the square root.

$$RTTR = V/\sqrt{N} \quad (1)$$

$$CTTR = V/\sqrt{2N} \quad (2)$$

Daller et al. (2003) argued that such measures are superior to raw TTR as the manipulations of token count lessens the impact of falling TTR, at least for the first few hundred tokens. Daller et al. also report that researchers such as Vermeer (2000) and Broeder, Extra, & Van Hout (1993) have made some positive findings with these metrics. Other testing, however, showed that both measures were significantly confounded by text length (Hess et al., 1986; Ménard, 1983; Richards, 1997).

"Linearizing" approaches can be dated as far back as 1960, when Herdan replaced the "correcting" square root with the "correcting" log. A veritable flow of ever more complex linearizing soon followed.

Herdan (1960):	$H = LogTTR = (LogV/LogN)$	(3)
Summer (1966):	$S = LogLogV/LogLogN$	(4)
Mass (1972):	$a^2 = (LogN-LogV)/Log^2 N$	(5)
Dugast (1978):	$U = (Log^2 N)/(LogN-LogV)$	(6)
Tuldava (1993):	$T = LogLogN/(LogLog((N/V)+A))$ [5]	(7)

The inviting aspect of LogTTR measures, in whichever form, is that there does appear to be a relationship between TTR curve and logarithmic curve (see figure 5).

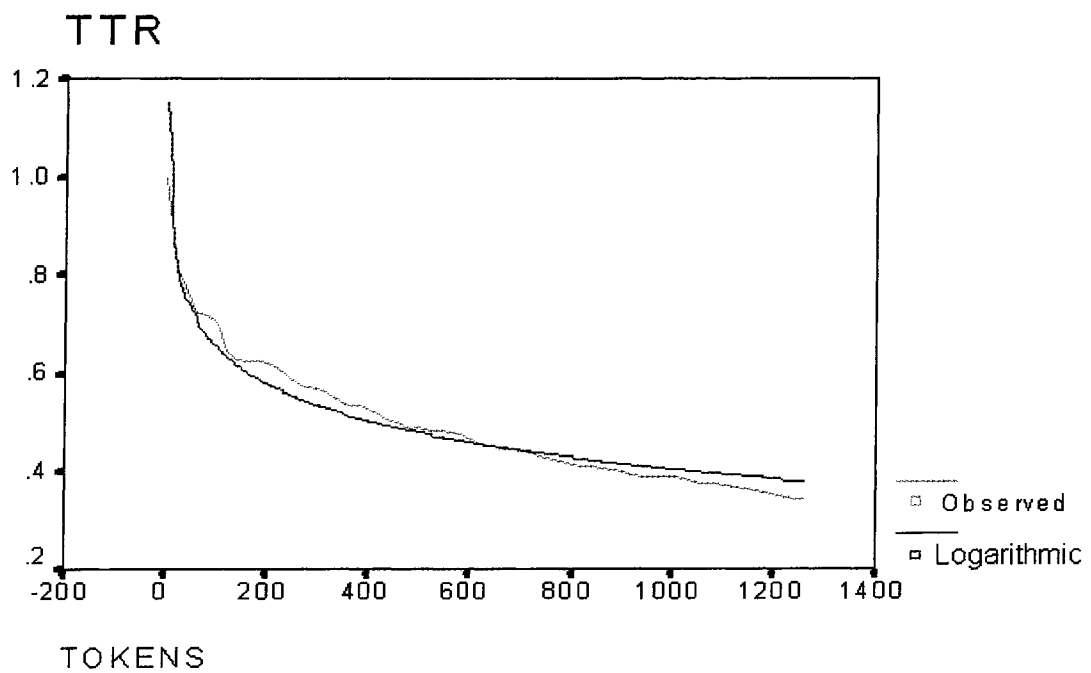

Figure 5. The TTR Curve Produced from the Opening Chapter of *Red Badge of Courage* Against the Logarithmic Curve Estimation.

[5] The value of A depends on the genre of the text

Unfortunately, appearances can be deceptive, and once again, extensive testing has invalidated LogTTRs (Malvern et al., 2004). Tweedie and Baayen (1998), for example, tested LogTTR and U (and a number of other related measures) over texts ranging from 24,246 tokens to 116,534 and found all to be functions of text length. Hess et al. (1986) also included LogTTR in their study language samples from children aged 3;0 to 5;11 and they too found no evidence that such measures were independent of text length. Each measure fails because it assumes there is a constant relationship between N and V. But as we have seen, V has a ceiling whereas N does not, so mathematical corrections can merely reduce but not eliminate the problem of type to token ratios.

Can LD Be Corrected by Assessing the Quality of Types?

Thus far, the correction approach has centered on adjustment of token count. Some researchers, however, feel that correcting for *quality* of types is more justified than correcting for *quantity* of tokens (Daller et al. 2003). Accounting for the quality of types rather than taking every type as equal is often referred to a lexical richness (LR) rather than lexical diversity (Vermeer, 2000), although as Malvern et al. (2004) point out, the terms are generally seen as being synonymous (for example, Arnaud, 1984; Menard, 1983; Wimmer & Altmann, 1999). To stem any unnecessary confusion in this study, we shall term supplementary, qualitative measures as lexical richness (LR), where LR is contained within, rather than is distinct from, LD.

Daller et al. (2003) attribute the superiority of LR as a measure of LD to an illustration made by Meara and Bell (2001). The latter compare Sentence A: "the man saw the woman," to Sentence B: "the bishop observed the actress." The type-token ratio for these sentences is equal; however, Sentence B contains less frequently used tokens and, therefore, demonstrates greater lexical ability. Those with greater lexical ability, the theory continues, will be more likely to use rarer words, whereas everyone will always use more common words. As such, by removing frequent words from the equation, the richness of the deployed vocabulary remains.

The origin of LR as a measure of LD stems from Arnaud (1984). Arnaud compared essays produced by native French speakers with essays written by non-native speakers. Excluding the 1522 basic words which the French Ministry of Education supposes all pupils of lycee entrance level should know, Arnaud counted the number of types that remained in the essays. Arnaud supposed that words outside of the 1522 list could be called advanced words and, on this premise, was able to posit a formula that distinguishes proficient language users.

Linnarud (1986) conducted a similar study, this time comparing 21 native Swedish speakers to 42 non-native speakers. Linnarud's richness tokens were calculated from lists of presumed known words produced by Swedish schools. Linnarud's assumption was that students who produced words outside the list would be demonstrating greater proficiency. The result of this study, as with Arnaud's, was a distinction between native speakers' work and non-native speakers'.

More studies on LR have used other lexical frequency lists (Laufer, 1995, Laufer & Nation, 1995). The principle here is much the same as with Arnaud and Linnarud.

Laufer's "beyond 2000," for example, takes the number of words *not* in a frequency list as a percentage of the overall number of words in the text. The conclusion is, the higher the resulting percentage, the greater the lexical knowledge.

Arnaud's, Linnarud's and Laufer's developments have all produced useful analyses. However, once again, there are many problems with their approach. Firstly, Arnaud only analyzed the opening 180 words of each text – the length of the shortest text produced. As we have seen above, this may not allow like to be compared with like. Similarly, Laufer's results do not appear to have circumvented the text length issue (Daller et al., 2003). Secondly, as Meara and Bell (2001) point out, low level learners and young children are unlikely to have much knowledge of rare words. As such, it is unclear how useful frequency profiles would be to anyone outside of a proficient speaker.

A third criticism questions the efficacy of frequency lists: As anyone involved with teaching English to students of other languages knows well, the problem for students is often *not* the use of *rare* words, it is the non-use of *common* words. For example, native French speakers may use *descend* where *go down* would be far more appropriate; multi-part verbs, on the other hand, are difficult for students of English, whereas a Latin based cognates are often far easier (though often not correct). As such, frequency lists account for rarer English words as used by native speakers, and measures using these lists discount those words as being *not* a significant measure of linguistic knowledge. In fact, these common words are often difficult to learn and difficult to use, and for that very reason, eliminating them from the analysis would not be helpful. Indeed, it could be very misleading.

Related to the third point is an issue of appropriateness. *Homo-sapian* is a less frequently used word than is, say, *man*. However, if one student were to write: "I met an interesting homo-sapian last night," whereas another were to write that she merely met a "man," a measure of LR would judge the former to have demonstrated greater knowledge, sophistication and/or proficiency. Thus, as these LR measures seem to present as many problems as they avoid, and as none of these measures have demonstrated that they are not text length sensitive, it is time to move on to other attempts to find a reliable measure of LD.

Can LD Be Corrected by Assessing the Frequency of Types?

A third approach to correcting for TTR is to consider the frequency or distribution of types. For example, one text may contain exactly the same number of types and tokens as another text; however, the number of tokens within each type may differ. The following two sentences reflect this issue of diversity, and although any given sentence may be negligible compared to an entire text, it needs to be understood that these sentences are representative of the distribution that could occur across an extended text:

(a) The tall boy and the tall girls like the clever boy.

Types = 7

Tokens = 11

(b) The tall, clever boys the small boys like, like tall girls.

Types = 7

Tokens = 11

Sentence A contains three tokens of the type *the*, whereas Sentence B contains only two. Sentence B, however, contains four types that have token frequencies higher than one: *the, big, boys*, and *like,* whereas Sentence A has only three such types: *the*, *big*, and *boy*. Thus, the total token count (N) and total type count (V) are the same, but the frequency of tokens to a specific type (r) is different. Many researchers have sought LD reliability through utilizing this additional element (Honore, 1979; Orlov, 1983; Yule, 1944). The theory underlying their encouragement is, while type count slows dramatically as token count increases, frequencies of tokens within types continues to rise. Thus, if a formula that characterizes frequency distribution can be tapped, perhaps reliability can be produced.

To this end, Honore (1979) produced the LD measure known as R by considering log scores with the number of types that had a frequency of one: *hapax legomena.*

Michea (1969) and Sichel (1975) considered types with a frequency of two: *hapax dislegomena,* from which were produced LD measures of M and S respectively. And Yule (1944), considered the frequency of each and every type to produce a measure known as K.

Each of these measures has been extensively tested, most thoroughly by Tweedie and Baayen (1998) who consider such measures to be useful for indicating style as long as they are used with caution. However, Tweedie and Baayen's main conclusion is that the notion that such measures are text length independent is simply "invalid." Tweedie and Baayen's protestations notwithstanding, Smith and Kelly (2002) used the LD measures of Z and K to study the chronological development of ancient author styles. Smith and Kelly cite Tweedie and Baayen (1998) so, clearly, the former were aware of the latter's paper. However, the former cite the latter as having sanctioned the measures whereas, in fact, Tweedie and Baayen's distrust of the measures could hardly have been made more clear: they write "Our conclusion will be, therefore, that it is extremely hazardous to use lexical 'constants' to compare texts of different length."

While Tweedie and Baayen's conclusions appear to negate the claims of measures such as Z and K, two caveats need to be considered. Firstly, Tweedie and Baayen used extremely long texts (24, 246 tokens was the shortest) to run their tests, thus these measures tested at shorter lengths may not show the same problems. Secondly, Tweedie and Baayen used only literary texts, thus it remains unknown how well such measures operate on other genres.

What Is "Curve Fitting" and Can it Correct for the Problem of LD?

A very different approach to measuring LD is curve fitting (Malvern & Richards, 1997; Jarvis, 2002). Malvern and Richards note that the entire TTR curve is what characterizes a text, yet all TTR measures to date have merely looked at single points within the curve or, more often, the very final point of the curve. Thus the object in curve fitting is to produce a formula that offers a theoretical parameter which closely fits the empirical TTR curve. This parameter would then stand as a measure of LD.

There are two clear advantages to such a measure. Firstly, the measure uses all the data available. Secondly, and more importantly for the sake of reliability, a parameter that encompasses the entire, empirically produced TTR curve is theoretically independent of text length: inasmuch as the theoretical curve matches the empirical curve, however long the empirical curve may be, the very match negates any issue of text length. The point where the theoretical curve ends its close fit with the empirical curve marks the limits of the parameter in terms of its independence of text length.

The parameter measure produced by Malvern and Richards (1997) is known as D; however a later adaptation of the measure (Malvern et al., 2004; McKee, Malvern, & Richards, 2000) is also known as D. For the sake of clarity, therefore, I shall refer to the initial parameter as D^a, and the later version as D^b. The distinction is important as the measures are produced in quite different ways and yield quite different results.

D^a is a simplified version of a formula first proposed by Sichel (1986). Sichel used two parameters (*b* and *c*), but Malvern (1991), having determined that the c

parameter was a better fit for shorter texts and the b parameter was a better fit for longer texts, reduced the formula to one parameter. As the inherent problem of TTR is most noticeable with shorter texts, and as Malvern's primary goal is measuring the lexical diversity of short, children's transcripts, Sichel's c parameter was adapted into a D parameter whereas the b parameter was discarded. The resulting formula was a more simple, single parameter that would be most appropriate for shorter versions. The resultant equation is as follows.

$$TTR = (2/D^a\ N)\ [(1 + D^aN)^{1/2} - 1] \quad (8)$$

Thus, TTR is the text's type-token ratio, N is the token count, and D^a is the parameter. The final value of LD is determined by adjusting the value of D^a until the equation converges on the value of the TTR.

The most complete testing of D^a (and other adapted curve fitting approaches) available in the literature is Jarvis (2002). Jarvis acknowledges that such parameter-approaches are time consuming, complex, and require many divergent skills, and that these factors have probably contributed significantly to the lack of research in such methods of solving LD. The relatively few available measures meant that Jarvis took four existing measures that could be adapted to curve fitting techniques: Herdan's index, Guiraud's index, the Uber index, and Zipf's Z, and to these he added D^a and a sample of TTR variations.

For material, Jarvis used essays produced by 140 Finnish and 70 Finland-Swedish students of English. An additional 66 native English speakers were also added to the

study. The produced texts were approximately 300 tokens in length. Jarvis divided the texts into two versions: whole texts and content word only texts. To produce empirical TTR curves against which to fit parameters, Jarvis adapted a "smoothing" technique for TTR scores calculated at intervals of 20 tokens: the interval range stems from Baayen and Tweedie (1998). The smoothing technique, endorsed by Malvern and Richards (1997) takes the two TTR scores on either side of the interval range and averages them to form a final TTR. Such smoothing neutralizes peeks and troughs that are characteristic of TTR curves formed one token at a time (see for example, Youmans, 1991).

Jarvis used the least-squares curve-fitting procedure to evaluate the theoretical LD measures against the smoothed, empirical TTR curve. This procedure produces mean distances from each of the selected points on the TTR curve to their corresponding points on the theoretical curves. The statistical significance of the distance was measured using the chi-square distribution. The results showed that the indices D^a and U had virtually identical success rates at curve-fitting for whole texts (98.19% and 97.83% respectively); however, for content word only texts, U produced 98.15% good fits, whereas D^a produced 90.77%. Other measures, except for Z, were considered "especially poor" for predicting curves on either variety of text. Z was highly successful with whole texts (96.74%) but was very poor with content word only texts (76%). Z also had the problem of producing unanalyzable scores as its formula could not produce a Z-score for some type/token counts.

Jarvis concluded that U was the best overall measure of LD, although D^a was certainly good for whole texts, but that neither model could predict all empirical curves and that both parameters left many texts as outliers. Jarvis also noted that analyses of

whole texts rather than content word only texts, and non-randomized texts over randomized texts, offer the best LD results. Jarvis's main concern with his findings was the limitations of his corpora. He urged further research over more genres and of texts of considerably greater length.

Despite good initial showings for D^a, as Jarvis (2002) confirms, Malvern and Richards soon replaced D^a with a hardier version: D^b (McKee et al., 2000). D^b remains a curve fitting approach but its calculation has become substantially more complex leading to its being subsumed into an application, *vocd* (McWhinney, 2000). Dissatisfaction with D^a and its simple curve fitting procedures stemmed from the fact that TTR "curves" (in their natural form) are anything but smooth[6], and consequently cannot be expected to fit a theoretical curve. While curves can be "smoothed," as we saw with Jarvis (2002), and while peeks and troughs are interesting in and of themselves, as we saw with Youmans (1991), the D^b approach sees the marriage of a set of *Ideal* curves (which are necessarily stable) with an empirical curve smoothed through thousands of random text samplings. D^b, then, fits the empirically derived curve to the closest matching ideal curve, the parameter of which is the D^b score given as the measure of lexical diversity.

Naturally, the operation to attain D^b is no simple formula but requires sophisticated software. The *vocd* program incorporating the 'D^b parameter measure' is just such software. vocd is written in C and is available in UNIX, PC, and Macintosh

[6] Malvern et al. (2004) describe empirical curves as having the tendency to "wobble about somewhat."

formats. The program requires all analyzed files to be recorded in 'Codes for the Human Analysis of Transcripts' (CHAT): a system developed by Brian MacWhinney (2000). The *vocd* application is available for download from (http://childes.psyc.cmu.edu) as part of the Child Language Data Exchange System (CHILDES) - a corpus is comprised of child and adult conversation in a variety of settings.

In total, we can describe D^b as the result of the following *vocd* procedures:

1. Random samples of text are sampled without replacement.
2. Sample sizes are from 35-50 tokens (N).
3. 100 samples of each token size is taken and a mean is produced.
4. A D-score is calculated (based on the D^a formula) for a TTR of each N.
5. An average D-score is calculated for each of the prior D scores.
6. The whole procedure is repeated 3 times and a final average D forms the score D^b.

The choice for these particular procedures and parameters clearly requires some further explanation, and as the development of D^b marks a significant shift in the approach to calculating LD measures, a detailed explanation of the vocd procedure is certainly warranted.

The choice for random sampling stems from Malvern et al.'s (2004) belief that naturally occurring texts tend to see vocabulary items clustering together. For this reason,

Malvern et al. argue that taking strings of text and measuring TTRs at given points along N (the length of the text), invites the possibility of peeks or troughs in the TTR curve. Malvern et al. do concede that sequential sampling does preserve the authenticity of the textual structure (Jarvis, 2002; Tweedie & Baayen, 1998), and they also acknowledge that frequency formula go some way to compensating for clustering effects; however, as frequency count formula have failed to gain reliable measures of LD (Jarvis, 2002), Malvern et al. opt for random selection of vocabulary items believing such an approach would account for frequency likelihoods while avoiding the textual clustering described above. In addition, the averaging of multiple samples would produce smoothed, empirically based TTRs leading to better fitting of the empirical TTR curves to the ideal curves.

Malvern et al.'s (2004) random sampling is *without replacement*. That is, the items they select from a text, once selected, are not eligible to be sampled again. Malvern et al.'s reason for this is that shorter texts may have a type that occurs only once (V_1). In the *with replacement* model, it would be possible for a V_1 to become a V_2 or higher – thus measuring a diversity that does not exist in the text.

The interval point for recording TTRs in Tweedie and Baayen (1998) and Jarvis (2002) was N/20. Thus in a text 400 words long the first point would be the 20^{th} word, and in a text 4000 words long, the first point would be the 200^{th} token. Malvern et al. point out that such a method sees each text measured at different points, thus confusing comparison. Instead, they prefer a standard range of 16 points (35-50 tokens in length). The relatively small choice is suitable for children's transcripts (which may be quite

short), while the random sampling means that tokens from anywhere in the text may be selected.

As any single selection of random tokens may not produce a representative TTR, Malvern et al. (2004) choose to repeat the sample 100 times, and then find the mean TTR to ensure a representative score is reached. Malvern et al. claim that the choice of 100 cycles was empirically derived and was the result of a marriage of accuracy with processing time.

A D-score having been derived from the D^a formula for the averaged TTR of each N, the D-scores are then averaged to form yet another score of D. This whole process is repeated three times over so that the final D^b is the result of the mean of three D-scores. Again, Malvern et al. claim that these parameters result from much empirical testing, and that the final D-score, while stochastic, represents the best and most reliable measure of LD available to researchers (McKee et al., 2000).

McKee et al.'s (2000) claim of reliability for D^b stem from tests conducted on a spoken genre produced by children aged from 27 to 33 *months* (McKee et al., 2000). These texts were divided using the split half method so that *even* words were collected in one file, and *odd* words in another files. These two versions, along with the complete version, gave McKee et al. three texts to compare: *odd tokens*, *even tokens*, and *whole text*. The average D^b for each version was recorded with the results showing that there was no statistically significant difference between the versions. Thus McKee et al. were able to claim that sample size did not significantly affect D scores.

Since its inception in 1997, D^a, and now D^b, has gradually grown in popularity with researchers. Silverman and Bernstein Ratner (2000) use D^b in a study that compared stuttering children's language to that of similar aged non-impaired subjects. The study found that D^b could differentiate the two groups, but that raw TTR scores were unable to. Harris Wright, Silverman, and Newhoff (2003) also compared a number of LD measures, concluding that D^b was "a promising tool." Owen and Leonard (2002) used D^b to compare the LD of children with specific language impairment (SLI) to that of normally developing children of similar ages, also concluding that D^b was a useful tool.

In the field of applied linguistics, the most noticeable application of D^b stems from Malvern et al. (2004). This work involved a study of a French oral examination by native English speakers using the transcripts collected from both teachers and students. Although other LD measures such as MSTTR were included, the main measure of LD interest was D^b. In order to support the results suggested by D^b scores, Malvern et al. also incorporated non-LD measures into their findings such as teachers' assessment of *range of vocabulary*, *complexity of structures*, *content*, *accuracy*, and *pronunciation.*

The examination transcripts used for Malvern et al.'s (2004) study came from the British GCSE for French (thus the typical age for subjects was 16 years). One important feature of this examination is that the examiners simultaneously conduct interviews as they score the candidates. Furthermore, the examiner of the students is typically the students' own teacher. Malvern et al. acknowledge that such procedures lead to results that may not be able to be generalized, but as such examinations exist on a national level, the investigation may reveal important results.

Malvern et al.'s (2004) interest in this investigation stems from issues of oral interview techniques and assessment. He and Young (1998), for example, questioned whether oral assessments resemble natural conversation. Van Lier (1989) points out, the "power differential" between interviewer and interviewee and the elicitation demands of the situation confound the naturalness of the interchange. Young and Milanovic (1992), posit that the testers contribution is structured towards the goals of the exercise, whereas the testee's contribution is more reactive, again limiting the naturalness of the conversation. Malvern et al. also add their own concerns by acknowledging that the interviewer is bound to have greater control of topic and turn-taking, thus reducing the effectiveness of the interchange to reflect true conversational ability (Johnson & Tyler, 1998; Lazaraton, 1992; Moder & Halleck, 1998).

The results of Malvern et al.'s study provide interesting findings. Firstly, as was predicted, D^b scores correlated significantly with MSTTR, and (though to a lesser extent) also correlated with *the number of different words* in the transcripts. At the same time, there was no significant correlation between D^b and total number of words in transcripts. Malvern et al. interpret these results as further evidence of the validity of D. Secondly, and more surprising, is that D^b showed no significant correlation with the teachers' rating of *Range of Vocabulary*. Naturally, this immediately brings into question the validity of the teachers' rating which, of course, were only the impressions gained from the interview. Malvern et al. conclude that the teachers' ratings are almost certainly a result of halo effects, although they do acknowledge the difficulty of rating students and that the naturalness of scoring subjects can easily become based on more salient elements such as apparent richness of vocabulary and use of less-common terms. A third major

finding of this study was that D^b scores were actually higher for students than for teachers. Malvern et al. conclude that this is probably due to teachers oversimplifying their lexis for the benefit of the students. This finding, however, was not supported by individual analysis of teacher to student interaction. Instead, Malvern et al.'s results indicated that teachers were gearing their lexical diversity at a class level. As such, there was no indication of teachers accommodating individual students by, for example, simplifying their language for weaker students, but rather that teachers had a predefined notion of the ability of their classes and geared their speech to the general level of the group.

Malvern et al.'s (2004) conclusions from this study may offer important issues for assessment of foreign language students. Firstly, it appears that rating students while interviewing them, whether analytically or holistically, may invite halo effected results. Secondly, while some accommodation at a class rather than individual level may be beneficial for certain students, an interviewer who has no prior knowledge of the students' abilities may provide a more even playing field through which to assess a student's proficiency.

Despite such success, D^b is not without its distracters, nor without its problems. For example, Owen and Leonard (2002), while broadly supporting D^b, were not completely convinced of the measure's reliability. In their study, they noted that D^bs of 500 word samples were significantly different to D^b samples of 250 words (t (77) = 8.02, $p < 0.01$). The discrepancy in that study may result from any number of reasons. Firstly,

McKee et al.'s (2000) test versions did not consist of authentic texts, as only odd or even tokens were used, whereas Owen and Leonard's study used a sample of the first half of the data but did not compare the whole versions to the second half. It is arguable that one or both of these methods is not conducive to testing the reliability of D^b. Secondly, Duran et al. (2004) note that results from extensive testing of D^b confirm that the measure is only reliable with texts of from 50 to "a few hundred" tokens. Malvern et al. (2004) even question the validity of 50 as a lower end score. It is arguable, therefore, that Owen and Leonard's larger texts fall outside Duran et al.'s upper limit of "a few hundred tokens."

If D^b's own creators claim that the measure is only reliable over such small transcripts then D^b's wider application to other genres and greater text length is obviously limited. Further, even within the narrow limits of test length that Malvern et al. and Duran et al. have adopted, D^b has still only be tested on very low diversity transcripts (mostly very young children's transcripts). In order to know whether D^b is a sensitive measure it would be good to see evidence of highly distributed LD rather than the consistently low LD scores which are indicative of children's speech.

D^b also has some theoretical problems stemming from its operationalization that need to be addressed. Jarvis (2002) is skeptical of D^b's random sampling techniques. He notes that such sampling treats coherent textual structure as if it were no more than individual particles like a "bucketful of differently colored corn kernels." While Jarvis concedes that the sampling technique does produce probabilistic TTR curves he is extremely doubtful that such curves actually reflect the text from which they are taken. Jarvis's point has some validity as a random selection of 50 words from, say, War and Peace, could, no doubt, supply enough lexical items to produce text from a great variety

of other recognizable genres. The point Jarvis is making, however, is that LD measures, to be authentic measures of a text, need to be formed from a sequential analysis of the lexicon.

Even setting aside Jarvis's concerns, the random sampling procedure of *vocd* presents other problems. Essentially, the more types a text has, the less the likelihood of any particular type being selected. For example, a type in a text of 100 types (where each type has a frequency of, say, 2) has a 50% chance of selection given 50 random selections; whereas a type in a text of 200 types (with the same frequencies) has a 25% chance of selection. As the number of types increases, so does the likelihood of its selection decrease, meaning that a selection of a repeated type becomes less and less likely. Selecting diverse types increases the numerator, thus decreasing the TTR, thus increasing the D-score. To overcome this, the random selection vocd operates needs to be increased from its current maximum of 50 selections; however, increasing the selection decreases the ability of D^b to analyze short transcripts – its primary purpose. Should the selection limit be increased to, say, 100, the shortest analyzable transcripts may be as much as 200 tokens: longer than most transcripts in the McKee et al. (2000) test that "validated" D^b in the first place. vocd is, therefore, locked into a similar kind of problem as MSTTR. Namely, an increase in accuracy brings with it a similar decrease in sensitivity. With the D^b measure as it is, we can predict that longer texts (having greater likelihood to produce greater numbers of types) will result in a D^b measure significantly with text length, with a more marked effect occurring as the text length increases.[7]

[7] This hypothesis is confirmed in the results of Chapter 3.

A related issue, causing more problems for D^b, is what we might call the two-equal-books phenomenon. Consider Text A, which is the entire text of the *Red Badge of Courage*. Now consider Text B, which is Text A with a second copy of the book tagged on at the end. If we can define Text A as having a theoretical LD value of X, we can arguably define Text B as having the same value, because it is merely the same book produced twice. Of course, TTR would report Text A as having a much lower LD than Text B as the latter would have twice as many tokens, but exactly the number of types. D^a, and other curve fitting approaches, would be able to approximate the TTR curve as such methods operate sequentially by following the development of the text. D^b, on the other hand, samples randomly and would, therefore, fall into the same problem as TTR – not least because D^b operates by calculating multiple TTRs. Random sampling, therefore, may not be the best approach to measuring LD.

Where Next for LD?

None of the LD measures mentioned here are without their uses though all, equally clearly, provide researchers with fundamental problems of reliability. To overcome these problems, some LD measures have recommended or enforced rigid criteria to their use: extending their reliability at the expense of limiting their use. A relatively simple example is limiting texts to exactly the same lengths – as was proposed

for TTR. However, as we have seen, discarding information from the analysis for the sake of convenience is hardly a good way to measure it.

More complex adjustments to texts for the sake of LD have also been considered. Daller et al. (2003), for example, argue that function words should be excluded from analyses as they "contaminate" measures with their repetitiveness telling us nothing about the proficiency of the speaker. However, excluding function words does not mean that these words are not part of the lexicon. Indeed, it is arguable that certain genres display greater levels of function words than others. Biber (1988), for example, believes that function words play a significant role in textual analysis. And Jarvis (2002) analyzed texts using both *with* and *without* function words and found that the best measures (U and D^a) operated almost equally whether function words were included or excluded.

Malvern et al. (2004), in their latest version of D^b, claim the stripping of inflections from words improves the reliability of D^b. The process, referred to as, "morphemicisation" significantly (and not surprisingly) reduced D^bs thus helping to bunch up results which, naturally, lead to less differences across scores.

Singh (2001), also removes certain phrases from her analysis of male/female speech; a study which claims 90% accuracy in determining the gender of the speaker. Phrases such as "you know" and yes/no responses were deemed "empty speech" that would negatively affect measures of lexical richness. The choice to remove such expressions is questionable, especially considering that researchers such as Freed and Greenwood (1996) base their *negative* findings of differences between male/female speech by looking at just such expressions as "you know."

Excluding lexical information from an analysis, whether a whole word or parts of words, is not a productive first step to establishing a reliable measure of LD. A reliable method would be able to calculate LD scores for texts regardless of their content – especially as it is their content that is being measured. A reliable measure of LD would also be able to measure short texts and long texts, as well as texts of various genres. Unfortunately, as we have seen, none of the measures thus far presented seem able to perform well unless rigid borders to their operation are imposed. Such a claim is not unreasonable given the wide variety of testing that has already been conducted (Jarvis, 2002; Tweedie & Baayen, 1998); however, as these tests comprised limited examples of LD measures, on texts of limited sizes and genres, we feel it prudent to examine the widest possible range of measures on the widest possible range of material. In the next chapter, therefore, we will test each of the main measures of LD described here against a wide set of requirements (various genres and various length). In doing so, we hope to establish clear limitations as to the usefulness of each measure as a reliable gauge of lexical diversity.

Chapter 3: Assessing Complex Measures of LD.

The testing of LD measures has almost as rich a history as does the development of those measures in the first place. And just as no agreement has been reached as to which LD measure is best, so too has no agreement as to the testing of those measures been established. Some research, for example, concentrates on written texts (Rietveld & Van Hout, 1993; Tweedie & Baayen, 1998), whereas others concentrate on spoken (Bucks et al., 2000; Wright, Silverman, & Newhoff, 2003). Some deal with relatively short texts of around 300 tokens (Jarvis, 2002), whereas others deal with extremely long texts ranging from 24, 246 to 116,534 tokens (Tweedie & Baayen, 1998). Some investigate LD measures from published works (Rietveld & Van Hout, 1993; Tweedie & Baayen, 1998), whereas others use student essays (Jarvis, 2002). And some use children as their subjects (Malvern et al., 1997; Vermeer, 2000), whereas others use adults (Wright, Silverman, & Newhoff, 2003).

Just as the material forming the corpus for testing is not agreed upon, neither is the method for testing reliability: McKee et al. (2000), for example, use the split half method for testing data, whereas Owen and Leonard (2002) tested whole texts against the text's first 250 words. Tweedie and Baayen (1998), on the other hand, plotted scores from ever increasing sample sizes, whereas Hess, Sefton, and Landry (1986) divided their entire transcripts into various segments – each increasing by a factor of 50 words.

With no established range of material or method of establishing reliability, the goal of this chapter becomes two-fold: to establish a corpus with suitably diverse material

for testing existing LD measures, and to establish a method for testing existing LD measures against this corpus. We believe that the corpus and tests related in this chapter are formed from a wide enough range of material, and a broader enough method of investigation to satisfy most critics.

That said, it is natural to suppose that whichever methods or material are chosen for whichever reasons, there will be criticisms from some quarters. For example, though most studies that test LD measures draw their material from just one genre (Jarvis, 2002; McKee et al., 2000; Tweedie & Baayen, 1998) this study uses twenty-three. But even this large sweep of language registers fails to include certain types of communication such as task-based activities or scientific manuscripts. The reasons for these limits (and others) were generally attributable to time constraints; however, we do acknowledge that any findings reported here may differ for material from genres not part of this study.

How Did We Select Material for the Test Corpus?

For Malvern et al. (2004), the goal of a sound measure of lexical diversity is reliability across both speech and writing in the form of independence from text length. The key words here are "speech," "writing," and "independence from text length." If such a measure is to be established, therefore, it should be tested against a corpus of both spoken and written material, and tested so that the length of the text should play as small a part as possible.

Beginning, then, with the selection of corpora containing written and spoken texts, we chose to use Biber's (1988) selection of texts from the Lancaster-Oslo-Bergen corpus (Johansson, Stig, Leech, & Goodluck, 1978), containing 15 written registers, and the London-Lund Corpora LLC corpus (Svartvik & Quirk, 1980), containing 6 spoken registers. The genres contained in these corpora can be seen in Table 1.

Table 1

The 21 Registers Taken from the LOB and LLC Corpora

Corpus	Register
Lancaster-Oslo-Bergen Corpus	Press reportage; editorials; press reviews; religion; skills and hobbies; popular lore; biographies; official documents; academic prose; general fiction; mystery fiction; science fiction; adventure fiction; romantic fiction; humor
London-Lund corpus	Face-to-face conversation; telephone conversation; public conversations, debates, and interviews; broadcast; spontaneous speeches; planned speeches

Source: Biber (1988)

The LOB corpus contains printed material from 1961. Each of the 500 texts is comprised of approximately 2,000 words. The LLC corpus contains 87 transcripts of

spoken English. The texts are approximately 5,000 words in length. These corpora have previously been used together for a variety of studies comparing spoken and written language (Biber, 1988; Louwerse, McCarthy, Graesser, & McNamara, 2004), so they are particularly appropriate for this study. Biber added a further two registers to the written texts supplied by the LOB corpus: personal letters and professional letters. In Louwerse et al. (2004), these registers were also included although different texts were used. Biber did not specify why these two registers were added, although it can be supposed that their purpose was to extend the written dimension of his investigation. In this study, however, a primary reason for investigation is not to disambiguate written genres from spoken (as was the purpose for both Biber and Louwerse et al.), but rather to test LD measures against a variety of genres so as to establish the breadth of performance. LD measures are particularly relevant for educational purposes, so Biber's additional registers were replaced by two new genres: spoken educational and written educational. The spoken educational was taken from the Michigan Corpus of Academic Spoken English (MICASE) (Swales & Malczewski, 2001). This corpus includes transcripts from over 200 hours of conversations in an academic setting. The written education is taken from *Glencoe Science* (Biggs, Daniel, Feather, Leach Snyder, & Zike, 2003), a grade 6-8 science text book.

What Was the Method Used for Analysing the Texts?

In previous studies of LD reliability, texts have generally been divided up into smaller pieces (Chotlos, 1944; Hess et al., 1986; Owen & Leonard, 2002). This is not the only method, however. Meara and Bell (2001), for example, compared two texts written about one week apart in order to test reliabilities. For the large number of texts in this study, however, obtaining texts that were written or spoken on separate occasions would have been too time-consuming. We therefore adopted the former method of dividing texts into smaller pieces for analyses.

Splitting up a text into smaller pieces can also be applied in a number of ways. McKee et al.'s (2000) "split-half" method provides two alternative texts (the even numbered words and the odd numbered words) to be tested against the original whole text. As Malvern et al. (2004) explain, such a method has the advantage of eliminating theme diversity: A text may have ten themes in its first half, but twenty themes in its second half. The greater diversity of themes may lead to greater LD as more themes imply more lexical variation in order to relate these ideas. The split-half method eliminates the potential confound of a theme imbalance.

But the split-half method also has problems. The first problem is that there is little evidence to suggest that theme variation leads to LD scoring imbalance: If greater themes lead to greater diversity scores then longer texts would have higher TTRs scores, which they do not. In addition, it is extremely difficult to adjudicate what constitutes a shift in theme, so the issue may well be moot. In either case, it could be argued that any kind of

theme deviations, such as they are, may constitute intrinsic elements of the text: eliminating such themes, therefore, may be counter-intuitive.

A second problem with the split-half method is its appropriateness for testing the reliability of a measure such as D^b. D^b functions by sampling without replacement. As such, the distribution of lexicon within a text is a crucial factor: an imaginary text containing just four words (where two types have two tokens each) is not the same as a text where the two types have one token and three tokens respectively. When a *sample* is extracted using a non-replacement method, as *vocd* does, the significance of distribution becomes apparent. Given a *sample* of *three tokens* from each of the above two examples, it is possible for all three tokens to be selected from only one type in the *second text*, but a sample of three tokens from the *first text* would have to incorporate both types. The result, therefore, *could be* a very low D^b score for the second text, but would always be the same moderate score for the first text. While this may sound like hair-splitting, the point is that the distribution of tokens throughout the text is essential in the calculation of D^b. Malvern et al. (2004) acknowledge this point themselves, stating that the failure to take into account the frequency of repeated types is a limitation of raw TTR. But the split-half method, while nullifying any potential theme variation, also nullifies the original textual distribution along with ending any kind of syntactic authenticity, meaning that the produced samples only very poorly represent the original.

A third problem is that the split-half method only provides three versions for comparison where the two smaller versions may well fall *above* the critical size of "a few hundred tokens" (Duran, Malvern, Richards, & Chipere, 2004; Malvern et al, 2004). It

would be preferable, therefore, to adopt a method which allowed for analyses that incorporated both long and short texts.

Parallel sampling computes the mean of texts that are cut into smaller segments. It is a much more common practice, not least because it preserves the syntactic structure of the text (Hess et al., 1989; Hess et al., 1986; Jarvis, 2002; Owen & Leonard, 2002; Tweedie & Baayen, 1998). But this method does have its problems too. Malvern et al. (2004) quite rightly point out that parallel sections mean that repetitions of tokens in subsequent sections will be missed, thus artificially driving upwards any LD scores for shorter sections. While it is acknowledged, therefore, that parallel sectioning of texts is not perfect, it still presents fewer methodological problems and has a longer and more diverse history as the method of choice.

The *size* of the parallel sectioning of texts also has a diverse history. Hess et al. (1986) divided texts at the 50-word level, Owen and Leonard (2002) at 250, and Tweedie and Baayen (1998) at 2000. Malvern et al. (2004) argue that Hess et al.'s method was both more sophisticated and more reliable as it divided texts into groups of 50, 100, 150, and 200 word sections: the greater the sectioning, therefore, the wider the analysis. Hess et al. had only used texts with a maximum size of 361 tokens, however, and when dividing their texts into smaller sections they had often had to discard access tokens: that is, a 361 token text could be cut into three sections of 100 tokens but the remaining 61 tokens would be discarded. The mean of the three 100-token section, therefore, is not an analysis of the whole text. Malvern et al. (2004) criticize methods that regularly discard

data and so we preferred to adapt Hess et al.'s method so that all available text could be retained. The results of this adaptation can be seen in Table 2.

Table 2

The Texts Lengths Used in the Analysis and the Number of Sections for Each

Text lengths	Number of sections
2000	1
1000	2
666	3 (two texts of 667 tokens)
500	4
400	5
333	6 (two texts of 334 tokens)
286	6 (two texts of 285 tokens)
250	8
200	10
154	13 (two texts of 153 tokens)
100	20

The selection of these section sizes requires some explanation. For example, we decided against analyzing a section size of 50 tokens for two reasons: Firstly, 50 tokens in not an authentic "sample" for the purposes of measuring D^b. That is, *vocd* samples

from 35 to 50 tokens during its cycles. As such, for a text of 50 words, the entire text would be analyzed, rather than a *sample* of the text. This concern is echoed by Malvern et al. (2004) themselves, who caution against the use of small samples, writing that D^b is not guaranteed over texts of such short length. Thus, we decided upon a lower limit of analysis of 100 tokens: twice the Malvern et al. stated minimum requirement.

The upper limit of the analysis was largely an artifact of the LOB corpus. Biber (1988) held that texts of 2000 words, such as those in the LOB corpus, were sufficiently representative. We, therefore, had to make parallel sections between the lower limit of 100 tokens and the upper limit of 2000 tokens such that any trend generated as a function of text length would become apparent. While it is certainly true that trends may not become apparent until much longer texts are analyzed, the upper limit of 2000 words is still much larger than virtually all other LD investigations[1]. In either case, one of the main purposes of this test was to establish whether and where D^b becomes affected by text length. As its creators have hazarded an upper limit of "a few hundred tokens," it is clear that 2000 tokens is well beyond that point.

In order to satisfy Malvern et al.'s (2004) reasonable preference, that in testing LD reliability all tokens should be included, we divided texts into parallel sections which left no remainders. These sections, as with Hess et al. (1986), were separated at around the 50 token mark. Thus, 150 tokens would be the natural first choice for the next larger section following 100 tokens. 150 tokens, however, does not fit into 2,000 words, thus we

1 Tweedie and Baayen (1988) is the obvious exception. One of the books they used one book contained 116,534 tokens.

decided on 154 tokens as the next sample size, allowing that two of these texts would have 153 tokens to accommodate the 2000 word maximum.

How Did We Select Texts?

The essence of corpus analysis is that very large bodies of similar texts provide a microcosm of linguistic structures. Biber (1988), for example, used texts from the LOB and LLC corpora in order to search for examples of 67 linguistic features. In this study, however, we do not search *through* texts for examples (discarding other data), we instead use each and every word of the text. For this reason, a representative portion of the corpora are all that is needed. Such reasoning echoes Biber's own analysis of these corpora, as he also drastically cut out numerous texts citing his reasons for this reduction as a marriage of the constraints of time while retaining what he believed was an adequate representation of the corpora.

The parallel sectioning of texts in this study gave us 11 sections per text to analyze (see Table 2). Based on the time constraints cited above, we decided that 9 texts per genre, yielding 99 parallel sections per genre, would form a suitable representation of each genre. As we intended to run correlations of the derived LD scores against token length of parallel sections, our 99 points of reference would be over three times as many as the sample size of 30 observations which is generally acknowledged to be a minimum for correlations (Van Genderen & Lock, 1977). While we feel that this offers a good faith

gesture to represent each genre, we accept that eleven slices of 2000 word texts could be viewed as insufficient. That said, the number of texts in this study dwarfs other highly regarded analyses of LD measures in almost every respect, as is shown in Table 3.

Table 3

Comparison of this Study's Range of Material Against Previous Similar Studies

Study	texts	words	Genres	Mode
This study	207	414, 000	23	Spoken/written
Tweedie & Baayen (1998)	16	1,126,240	1	Written
Rietveld & Van Hout (1993)	3	89, 307	1	Written
Bucks et al. (2000)	24	24, 000	1	Spoken
Jarvis (2002)	276	82, 800	1	Written
Owen & Leonard (2002)	91	45, 500	1	Spoken
Harris Wright et al. (2003)	18	3600	1	Spoken
Daller et al. (2003)	42	8,067	1	Spoken
Silverman & Bernstein- Ratner (2002)	15	6,404	1	Spoken
McKee et al. (2000)	38	12, 008	1	Spoken

Neither the LOB nor the LLC texts are self-contained. That is, the texts do not comprise a beginning, a middle, and an end of an instance of speech or writing. Given that the texts were representative samples, Biber (1988) used this premise to justify

further cutting texts to suit his analysis. For example, he divided 5000 word texts from the LLC corpus into smaller sections arguing that such actions did not violate the authenticity of the data. Louwerse et al (2004) also manipulated these same texts, again following the standards set by Biber. For these reasons, we felt that cutting the texts to the 2000-word level for the purposes of testing the reliability of LD scores was suitably justified.

Cutting texts at a particular point for the purposes on analysis is a common practice (Bucks et al., 2000; Harris Wright et al., 2003; Hayes & Ahrens, 1988; Menard, 1983; Owen & Leonard, 2002). Given that the primary aim of this experiment is to test *the reliability of a measure,* reducing as far as possible any differences in text length was the appropriate course. It could be argued, of course, that cutting text at the 2,000-word mark affects the syntax of the final sentence. While this is true, both the parallel sectioning and the split-half method will result in further syntactic amputations and such losses, while regrettable, could not be avoided.

Based on these reasons, we selected the first nine texts of each genre as representative samples. When texts were below the 2000-word level, they were ignored and the next text was selected in its place. As described above, longer LLC texts were cut into 2000-word samples to increase the number of texts available. Even with these manipulations, some registers were still unable to produce nine texts, as can be seen in Table 4. For these reasons, additional texts were selected from two addition corpora: the Brown corpus and the Wellington Spoken Corpus (WSC). The Brown corpus is almost identical to the LOB corpus in its design except that it uses American texts. From the Brown corpus, three Science Fiction texts were added to make the complete nine texts.

From WSC, which mirrors the LLC corpus using New Zealand English, additional texts were taken to make up the short fall.

Table 4

The Number of Files for Each Genre and their Source

Register	LOB and LLC	Brown and WSC	Texts in genre
Broadcasts	8	1	9
Face to face conversations	9	0	9
Interviews	9	0	9
Prepared speeches	5	4	9
Spontaneous speeches	5	4	9
Telephone conversation	4	5	9
Press reportage	9	0	9
Editorials	9	0	9
press reviews	9	0	9
Religion	9	0	9
Skills and hobbies	9	0	9
Popular lore	9	0	9
Biographies	9	0	9
Official documents	9	0	9

(table continues)

Table 4 (cont'd)

Register	LOB and LLC	Brown and WSC	Texts in genre
Academic prose	9	0	9
General fiction	9	0	9
Mystery fiction	9	0	9
Science fiction	6	3	9
Adventure fiction	9	0	9
Romantic fiction	9	0	9
Humor	9	0	9
Educational spoken	9	0	9
Educational written	9	0	9

Which LD Measures Were Included in the Study?

Of primary interest in this study was the new measure of D^b. However, like Jarvis (2002) and Tweedie and Baayen (1998) we decided to use the available data to test a wide range of other lexical diversity measures. Included in this test, therefore, were the following measures, each of which was described in chapter 2: D^b, D^a, Yule's K, M, S, U, CTTR, RTTR, H, Somer's S, Richet's K, Mass's a^2, and W. TTR as a mean of each

sample was also included. The measure Z was initially used in this study but functional problems lead to its exclusion. The problem for Z was very similar to that reported by Jarvis (2002). Namely, Z is often unable to calculate a score for some texts because the curves formed by the Z formula do not always intersect with a number representing the total types. While we cannot, therefore, offer any claims as to the reliability of Z, its failure to offer many results (especially over shorter texts) does suggest that Z is not the most useful LD measure for researchers to pursue.

As we were concerned about the sampling method incorporated by D^b, and as Malvern et al. (2004) acknowledge that D^b is unlikely to be reliable past a few hundred tokens, we predicted that D^b would correlate with text length. As Owen and Leonard (2002) had reported text length correlations at the 500 word level, we predicted that D^b would be text length dependent at least from this point on. Jarvis (2002) had had some success with D^a and with U. We therefore predicted that these measures would be more likely to avoid text length dependency. With Jarvis (2002) however, the text lengths were less than 400 words so we predicted any success with these measures would be quite limited. Our primary hypothesis was that all measures would show a significant overall correlation to text length.

\

Results

Results were taken by organizing the genres into three categories: all genres together (AG), all written genres (WG), and all spoken genres (SG). Further results would to be taken by analyzing the genres individually. The results can be seen in Tables 5 and 6.

Table 5

Significance Correlations for the first 7 Measures.

	H	SS	RK	Mass	W	D[a]	TTR
All genres (AG)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Written genres (WG)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Spoken genres (SG)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Broadcasts	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Educational spoken	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Face to face	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Interviews	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Prepared speeches	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01	< 0.01
Spon. Speeches	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01	< 0.01

(table continues)

Table 5 (cont'd)

	H	SS	RK	Mass	W	D[a]	TTR
Tel. conversations	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01
Academic	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Adventure fiction	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Biographies	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Educational written	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01	< 0.01
General fiction	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01	< 0.01
Hobbies	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Humor	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Mystery fiction	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Official documents	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Popular lore	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Press editorials	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Press reportage	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01	< 0.01
Press reviews	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Religion	< 0.01	< 0.01	< 0.01	ns	< 0.01	< 0.01	< 0.01
Romantic fiction	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Science fiction	< 0.01	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01	< 0.01

Table 6

Significance Correlations for the second 7 Measures

	D[b]	YK	M	S	U	CTTR	RTTR
All genres (AG)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Written genres (WG)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Spoken genres (SG)	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Broadcasts	ns	ns	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Educational spoken	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Face to face	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Interviews	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Prepared speeches	< 0.05	< 0.05	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
Spon. speeches	< 0.01	< 0.05	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
Tel. conversations	< 0.05	< 0.05	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Academic	< 0.01	< 0.05	< 0.01	< 0.01	ns	< 0.01	< 0.01
Adventure fiction	ns	ns	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Biographies	ns	ns	< 0.01	< 0.01	ns	< 0.01	< 0.01
Educational written	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
General fiction	< 0.05	ns	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Hobbies	< 0.05	ns	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
Humor	< 0.05	< 0.05	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01

(table continues)

Table 6 (cont'd)

	D^b	YK	M	S	U	CTTR	RTTR
Mystery fiction	ns	ns	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01
Official documents	ns	ns	< 0.01	< 0.01	ns	< 0.01	< 0.01
Popular lore	< 0.01	< 0.05	< 0.01	< 0.01	ns	< 0.01	< 0.01
Press editorials	< 0.05	ns	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
Press reportage	< 0.01	ns	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
Press reviews	< 0.05	< 0.05	< 0.01	< 0.01	ns	< 0.01	< 0.01
Religion	ns	ns	< 0.01	< 0.01	ns	< 0.01	< 0.01
Romantic fiction	< 0.05	ns	< 0.01	< 0.01	< 0.05	< 0.01	< 0.01
Science fiction	< 0.05	< 0.05	< 0.01	< 0.01	< 0.01	< 0.01	< 0.01

We hypothesized that all of the measures would correlate with text length. We also hypothesized that D^b would be particularly susceptible to inaccuracy as the text length grew. We further hypothesized that D^a, D^b, and U would be the most effective measures and that D^b would be more effective over spoken texts. As predicted, all of the 14 measures proved to be text length dependant: i.e. none were reliable. A Pearson's correlation showed that text length was significantly correlated with the AG category, ($r(2276) = .31, p < 0.01$), the WG category ($r(2276) = .19, p < 0.01$), and the SG category ($r(2276) = .36, p < 0.01$).

At the individual genre level, the best performing measure was predicted to be D^b. In fact, Yule's K was the most effective measure, managing to avoid correlation with text length for 11 of the 26 genres. Mass's a^2 was the second most effective, managing to avoid correlation with 9 genres. The equal third most successful measures were two of those predicted to be more successful: U and D^b. These measures avoided text length dependency with just six of the 23 genres. It should also be noted that neither U nor D^b was successful with any genre where Yule's K failed. Each of the other measures, including D^a, significantly correlated with text length for every genre ($p < 0.01$).

Malvern et al (2004) argued that D^b was reliable for spoken texts of up to a few hundred tokens. By systematically reducing the larger text slices, we found that D^bs became more reliable as the text lengths decreased. This was as predicted. The best performance for D^b was between 100 and 500 tokens. Here, four of the seven spoken genres (Broadcasts, Prepared speeches, Telephone conversations, and Spontaneous speeches) did not correlate with text length. Overall, however, D^b maintained text length dependency.

All of the LD measures had the greatest difficulty with spoken genres. Here, avoiding becoming a factor of text length was only achieved with the Broadcasts' register, and this only by Yule's K and D^b. With written genres, on the other hand, Biographies, Official documents, and Religion were all resistant to text length dependency with four different measures: D^b, Yule's K, U and Mass's a^2. We can therefore argue that successful measures were better able to cope with written rather than spoken genres.

The next step was to look more closely at the results using the most successful measure (Yule's K) to see if any distinctions between the scores were apparent. We were particularly interested with the 11 genres for which K had found no correlation. We began by checking Biber (1988) and Louwerse et al., (2004) i.e., two studies which used largely the same material, to see if there was any similarly in results. Both studies used factor analyses, deriving six dimensions. These dimensions split the texts into such categories as "involved versus informational production" (Biber, 1988) and "spoken versus written" (Louwerse et al. 2004). Despite having 12 overall dimensions available, none of Biber's six dimensions, nor Louwerse et al.'s six dimensions, showed any similarity to the genre divergence in this study. Although these studies were much more diverse in their investigation of genre (Biber using 67 lexical features and Louwerse et al. using 250 measures of cohesion and textual difficulty) it is perhaps surprising that no similarities emerged here.

We then turned to mean K-scores to see if the differentiating between correlations and non-correlations could be explained there. Of the spoken genres, Broadcasts received the highest overall average K at 129.06 leading to speculation that high K-scores were most susceptible to text length dependency; however, genres that show no correlation to text length include Adventure fiction, General fiction, and Press reportage despite having mean Ks below 106, the same mean scores as Press reviews and Humor which do correlate with text length. As such, neither overall high nor overall low K-scores explained the results.

Looking within the registers to the nine texts themselves, we found that individual texts did tend to correlate with text length, with only a few exceptions: Religion and

Official documents each had two texts that did not correlate with text length, and Broadcasts had one. But overall, the tendency, even here, was overwhelmingly towards text length correlation.

What appears to be the reason for no overall correlation with some registers was the lexical diversity range that some genres produced. As can be seen in Table 7, eight of the highest nine standard deviation scores belong to registers that did not correlate with text length, whereas six of the lowest seven standard deviation scores belong to registers which did correlate with text length.

This suggests that many genres do not have typical LD scores, but rather that their LD scores can be extremely diverse. For example, the range of Ks for Broadcasts was 129.86 whereas the second highest range for spoken registers was Telephone conversations at only 90.18. As such, it is probably range diversity that leads to a lack of correlations to text length, rather than the measures themselves actually avoiding the text length problem.

In sum, none of the measures explored so far used can be described as reliable measures that are independent of text length. All measures are particularly weak for spoken genres; however Yule's K was effective for 66% of the written genres.

Table 7

Shows the Standard Deviation Scores for Genres with High Scores Indicating a Tendency to be not Correlated with Text Length

Genres	*SD*	Correlated	Genres	*SD*	Correlated
Broadcasts	32.26	No	Hobbies	27.99	No
Interviews	11.53	Yes	Humor	18.63	Yes
Face to face	15.15	Yes	Mystery fiction	29.35	No
Educational spoken	17.29	Yes	Official documents	32.78	No
Academic	30.37	Yes	Popular lore	10.28	Yes
Spontaneous speeches	16.1	Yes	Press editorials	16.49	No
Telephone conversations	19.72	Yes	Press reportage	14.24	No
Prepared speeches	19.61	Yes	Romantic fiction	45.05	No
Adventure fiction	21.09	No	Press reviews	13.73	Yes
Biographies	22.63	No	Religion	38.5	No
Educational written	18.07	Yes	Science fiction	14.91	Yes
General fiction	16.4	No			

Discussion

The results of this study support those of many previous studies investigating the reliability of existing LD measures (Jarvis, 2002; Tweedie & Baayen, 1988). Namely, any existing LD measure is susceptible to text length, though there may be limited success at particular text length ranges with certain specific texts. The results of this study showed clear correlations with text length in each of the major categories: AG, WG, and SG. However, some individual genres (such as Broadcasts and Religion) demonstrated a great range in potential diversity of lexical usage. Where this is the case, the more reliable LD measures (such as Yule's K) do not show correlations.

These results fill a useful gap in the literature because they cover a text length range that is significantly shorter than Tweedie and Baayen (1988) and significantly longer than Jarvis (2002). These results also include substantially more LD measures than were used in either of the previous studies. More importantly, this study, covered 23 genres, and provided evidence that LD measures may operate differently on different genre. Future research involving the testing of lexical measures should ensure that a diverse range of genres are used as strong evidence of reliability with one measure for one genre may not be the same for another genre.

While extending our knowledge of LD measures, this study did not cover every eventuality. First, there are other genres that could be investigated such as the speech of very young children. Second, this study only examined English language genres, so there is no guarantee that the same results would occur for other languages. We also

acknowledge that this study did not strictly cover the home turf of measures, such as D^b, which have thus far predominantly been used to measure the LD of extremely young children. These children would have to produce a significantly large amount of dialogue to form a sensitive test of D^b, and as Malvern et al. (2004) prefer scenarios where topic shifts do not occur, it is extremely hard to see how D^b would be tested for reliability. This study also used unlemmatized tokens, which may have affected the results for some measures. It is possible that some measures may work better when only stem forms are used, or perhaps when function words are removed from the analysis.

Overall, the results of this study suggest that any existing LD measure must be used with caution. This study also suggests that a new measure of LD or a new way of looking at LD needs to be considered. If none of the existing formula or approaches can overcome the text length dependency problem, then meaningful comparisons across text length or genre remain in question. A reliable measure of LD must be able to avoid text length dependency, function across both major modes of communication, and give reliable results across a wide range of genre.

The next chapter introduces such a measure: the measure of textual and lexical diversity (MTLD) – will be introduced. After explaining the design and theory of this method, the measure will be submitted to the same gamut of testing as was conducted in the present chapter.

Chapter 4: The measure of textual lexical diversity (MTLD): A new approach to measuring LD

In Chapter 3 we saw that some measures of LD (for example, RootTTR, CTTR, and H) attempted to correct for the problem of text length dependency by trying to reduce the impact of the value of tokens. We saw that some approaches were based on frequency distributions (for example, Z and K), and we also saw that the latest approach to measuring LD blends curve fitting with sampling (D^b). In testing these measures, we saw that none were satisfactorily able to give reliable scores across linguistic modes or genres. We did see that some genres (such as Broadcasts) provided a wide range of LD scores which gave the appearance that some LD measures were not correlating to text length. However, closer investigation showed that this only occurred when the most successful measures analyzed the most diverse genres. We concluded that while Yule's K performed best in these tests, none of the available measures of LD were significantly reliable or sensitive. We therefore argued that a new approach to measuring LD was required. This new approach would need to produce a measure that was not only reliable but also returned results that successfully differentiated between high and low diversity across modes, genres, and individual texts. These two aspects (reliability and sensitivity) form the essential criteria for a useful measure of lexical diversity.

In Chapter 4, we shall look more closely at our understanding of lexical diversity: examining how the concept itself may be better understood for the purposes of a useful measure. We shall argue that maintaining the text structure rather than sampling the text

provides a more authentic measure of diversity, and we shall argue that the legitimacy of a measure decreases to the extent that a text is altered to suit the measure applied to it. Following this, we shall introduce a new measure of LD: the measure of textual diversity (MTLD). This measure, we argue, satisfies the reliability and sensitivity issues we believe to be essential. We shall also see that this measure satisfies our legitimacy test, as no alteration or modification of text is necessary. We shall then examine the results of this new measure as it is tested across the same range of material as used in chapter 3. These results, we shall argue, establish the credibility of MTLD as a reliable and sensitive measure of LD.

How Is 'Textual' Lexical Diversity Distinct From Lexical Diversity?

As we saw in Chapters 1 and 2, lexical diversity is by no means an agreed upon term. Some researchers do not distinguish LD from vocabulary richness (Arnaud, 1994), some see rarity of words as being an important element that must be brought into the calculation (Daller et al., 2003), and others see vocabulary richness as a multidimensional feature including components such as lexical variation, lexical sophistication, lexical density, and number of errors (Read, 2000). The lack of agreement as to what to call the measures we are here investigating and what to include in these measure is perhaps not so surprising in light of the more recent focus on establishing reliability (as is noted by Jarvis, 2000). Indeed, the debate as to what an LD measure should be measuring and

what the results of this measure mean might be considered moot given that a method for measuring LD has proven to be so elusive (Jarvis, 2002).

For some measures, and some testing of measures, texts and tokens require preparation to facilitate the accuracy of the measure. For instance, Jarvis (2002) lemmatizes tokens, arguing that inflected variations confound knowledge of tokens with knowledge of grammar. Jarvis's approach is far from unique, following in the wake Vermeer (2000), Laufer (1991), and McClure (1991). Despite this, Jarvis criticizes McKee et al. (2000) for a sampling technique that renders the text less than authentic, suggesting that the authenticity of text may not be that important after all. Another form of modification is provided by Daller et al. (2003), who use an advanced form of RootTTR where types are rated for rarity (see Chapter 2). In their calculations, only "advanced" vocabulary was used in the numerator as the authors argue that basic words provide no information as to the size of the subjects' lexicon. Indeed, Daller et al. argue that including non-advanced words would "contaminate" their measure. The denominator in their study, however, includes *all* tokens used, both basic and advanced. Malvern et al. (2004) have also come to believe that altering the lexical elements of the text is beneficial for the accuracy of their measure. They note that D^b is more accurate when tokens have been lemmatized, and argued that lexical diversity, for the most part, is whatever the researcher decides to call it, and however the researcher decides to define a word.

Whatever the justification for altering the components of a text, it is self evident that the more a text is changed from its original form, the less the original form is actually measured. It is also fair to argue that if the difference made by these changes were not significant then the researchers in question would not go to the trouble of making them.

Consequently, whatever it is that is being measured, it is not the text, and, as we have seen in Chapter 3, when the text itself is the subject of analysis, the measure fails to score reliably. We therefore argue, that in addition to reliability and sensitivity, a third criteria for a good measure of LD is that the text is left unaltered. Having said this, we do not feel that lemmatizing a text is without its uses. As Jarvis (2002) comments, there is a difference between lexical knowledge and grammatical knowledge, and researchers may well wish to distinguish this. For this reason, we basically agree with Malvern et al. (2004) - that lexical diversity, for the most part, is whatever the researcher decides to call it; however, we believe that there is a fundamental difference between lemmatizing a text so as to increase the reliability of the measure as opposed to lemmatizing a text to aid a researcher with a research question.

Despite the numerous linguistic features that comprise it, a text is far more than mere words and sentences strung together. A text also includes a structure, and this structure binds together the textual components so as to allow a reader or listener to form a coherent mental representation (Van Dijk & Kintsch, 1983). As such, to analyze a text, we need to consider the text as both the words it contains and the structure through which it is formed. We assume that the whole is more than the sum of its parts, and argue that any other kind of analysis at least risks losing significant elements of text.

We have decided to call a measure of lexical diversity that treats the text as a sequential whole, rather than a randomly ordered concatenation of discrete items, a *measure of textual lexical diversity* (MTLD). Measures of LD such as D^b, K, and TTR

consider the items of a text without reference to its syntax, yet they often choose to consider inflections, frequencies, and rarity. As such, they are measuring the parts of the text without reference to how those parts function in the text, which means that they are not measuring "the text" so much as they are a basket of "colored corn kernels" as Jarvis (2002) calls it. Of course, these measures do not claim to be measures of "textual diversity." However, a key assumption of LD measures is that they are measuring the text, so there may be reason for concern.

The consequence of such analyses can be illustrated by considering a shuffled text. For measures such as D^b, K, and TTR, a text may have its components ordered and reordered endlessly without ever altering the LD score it produces. If we consider textual diversity, however, any changes to the structure of the text may well affect the diversity score produced. For example, consider the following sequences:

(1) The the boy boy big small hit.

(2) The big boy hit the small boy.

Both sequences contain the same number and types of lexical items and, therefore, would receive the same score of *lexical* diversity (for any of the LD measures so far considered). However, the same texts *sequentially* analyzed, will form different TTR *patterns* (Youmans, 1991). Text A, for example, drops from 1.0 to 0.5 within the first two tokens.

It then rises, falls, and rises again. Text B, on the other hand, is quite steady for the first four tokens before slowly drifting downwards[1].

While this example is only at the sentence level, larger texts can exhibit major differences (as was evidenced by Youmans, 1991). The point is that texts are much more than the words that comprise them, as anyone who compares the sentence *the dog chased the cat* to the sentence *the cat chased the dog* will know.

For a further example, consider the following texts which are comprised of the indicated novels:

(3) *the Red Badge of Courage.*

(4) *the Red Badge of Courage* together with another copy of the *Red Badge of Courage.*

Virtually all existing LD measures (with the exception of, for instance total number of words in the text) would consider Text 4 to be far *less* diverse than text 3, even though exactly the same lexical items appear in both. The reason is that text 4 contains the same types but twice the overall tokens. Lexically, therefore, considering text 4 as *less* diverse is justifiable. However, from a *textual* point of view, the diversity is considered to be exactly the same. Considering two copies of the same book far less diverse than one copy of the book is not very helpful when judging material such as literature. This is to say, if

[1] The distortions will obviously differ considerably with larger texts.

LD is to be considered a useful *indicator* of difficulty or style, how can two copies of the same book be significantly less difficult or less stylish than one copy? Further, the logical conclusion from such an analysis is that, according to tradition LD measures, the complete works of Shakespeare are far less diverse than any individual work of Shakespeare; and if every novel ever written could be analyzed, then the diversity would be very low indeed. Whatever the justifications of such an outcome, the question is one of how helpful such an analysis could be?

For these reasons, we prefer to consider LD at the textual level and consider the following four criteria to be desirable for an overall measure of LD:

1. A measure of LD should be independent of text length.
2. That a measure of LD needs to produce sensitive scores. That is, where there is a wide range between low LD scores of texts and high LD texts.
3. That a measure of LD should be computed without need for lemmatization.
4. For a measure of LD to take into account text structure, it should also be able to solve a text sequentially.

At this point, a measure will be introduced that should cover the problems we have described while maintaining the strengths of earlier measures.

What Is the Measure of Textual Lexical Diversity (MTLD)?

Chapter 2 introduced MSTTR, a measure that divides texts into equal sizes, often 100 tokens, and discards any remaining tokens.[2]. MSTTR is a theoretically good measure of LD; however, as the reliability of this measure increases, the sensitivity decreases; and as the sensitivity increases, the reliability decreases. MTLD is based on MSTTR, just as D^b is based on Sichel (1986), but MTLD overcomes MSTTR's inherent problem by replacing segmented strips of tokens with segments of a given TTR score: the theory being that not all texts have a length of X but virtually all texts do have a TTR of X. As such, we do not attempt to capture a *size* of texts (in tokens) that is common to all texts (or virtually all texts) because no size is common. Instead, we capture a moment in the decline of TTR that it is common to all texts (or virtually all texts). Therefore, instead of seeing what the TTR of the whole text is, we see how many times a common TTR can fit into a single text.

Understanding how this theory works can be better understood if we first consider what we know about TTR. We know, for example, that all texts begin with a TTR of 1.0, and then gradually fall towards 0.0; however, we also know that the rate of fall is not constant. We also know that the rate of decline in TTR scores is far greater in the early stages than in the latter stages. That is to say, TTRs generally fall extremely quickly before they start to fall extremely slowly. A single token in the early stages of a text may

[2] MSTTR divides texts into equal sizes, often 100 tokens, and discards any left overs. The MSTTR score is the mean of the TTRs for each segment. Chapter 3 discusses this measure in detail.

cause the TTR score to fall by as much as 0.5; however, it could take hundreds of even thousands of tokens to move a TTR score the significantly lesser 0.05 if the text is substantially long. Therefore, the object becomes one of locating that point between these two extremes; the point at which TTRs are falling constantly enough, and consistently enough, for an analysis to take place. The point used in the initial experiments conducted here is a TTR of 0.71[3]. Each section of the text that consists of a TTR of 0.71 is deemed to be a "factor." By counting how many of these factors a text has, and dividing that number into the total number of tokens in the text, we derive a value that generally falls between 70 and 150. As with D, the higher the value, the greater the LD. The MTLD values are also roughly comparable with the scores calculated by D, although we believe that MTLD does not suffer from the text length dependency described in Chapter 3. With MTLD, based on the tests listed below, we believe we have a useful measure of LD that satisfies the four requirements for a reliable measure of LD that we have listed above.

How Is MTLD Calculated?

Just as D^b is a complex measure that requires the software of VOCD (McWhinney, 2000), we too have had to create software to run MTLD. The program, written in Visual Basic 6.0 and currently operating on Windows operating systems, calculates a MTLD score (see figure 6) through a number of operations. Firstly, the

[3] See Factor size of 0.71 (below) for more details.

program separates the text into individual unlemmatized token instances, one token at a time.[4] A TTR score is calculated each time a new type is found. When the TTR score reaches the factor value (default 0.71) the text is cut and a count of the tokens in that factor is recorded. Having cut the text at its factor size, MTLD then resets its TTR value at 1.0 and the process is repeated. MTLD will continue to produce factors until no tokens are left in the text. Any text that does not comprise a full factor is assigned a value based on the TTR level reached, which is to say, a "remainder score."

Once this initial cycle is complete, forming a first MTLD score, the whole text is reanalyzed in reverse order. This reanalysis provides a further score of MTLD. MTLD scores are formed by dividing the number of factors (and part factor) into the total length of the text. The final score for textual diversity is given as an MTLD score which is the mean of the forward MTLD score and the reverse MTLD score. A higher score of MTLD indicates greater lexical diversity.

Some features of the MTLD, such as the estimation score for remainders, the factor size of 0.71, and the reverse analysis, serve to optimize the reliability and sensitivity of the MTLD score without altering the words within the text or breaking the sequence or the authenticity of the text. These final criteria were the result of calibrations over test data, and were arrived at in much the same way as the final, selected criteria for D^b. D^b, it will be remembered from Chapter 2, selected its default sample sizes and number of recursions through testing a variety of alternatives. Malvern et al. (2004) concluded that the selected defaults for D^b allowed for the greatest marriage of accuracy,

[4] The tokens remain unlemmatized to maintain the authenticity of the text; however, as mentioned above, some may wish to lemmatize tokens for their own research purposes.

reliability, and processing time. Even greater recursion and sampling size did not lead to significantly better results for D^b, or, where there may have been better results, processing became prohibitive. MTLD follows this same line of thinking. That is, while some aspects of the MTLD are not theoretically etched in stone, initial testing has shown that they do offer the greatest marriage of accuracy, reliability, and processing time which, we will argue, has lead to reliable and sensitive LD scores while preserving textual authenticity. We must stress, however, that whether the calibrations used in these experiments are optimal is far from decided. No doubt a significant amount of tweaking will ultimately need to occur. As such, the results garnered from the analyses in this dissertation should be viewed as being an initial test of the MTLD tool, and that subsequent testing may yield even more accurate results.

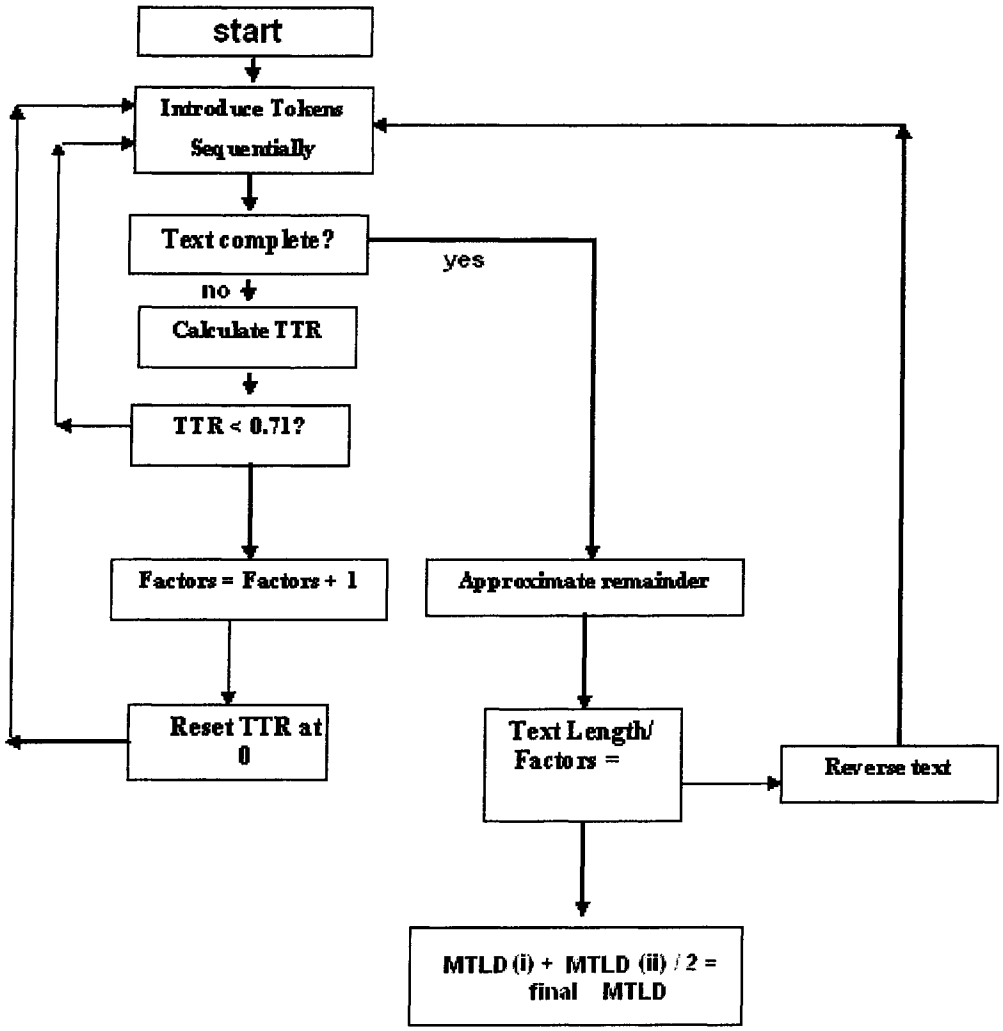

Figure 6. Flowchart for MTLD.

How Are Incomplete Factors Scored?

Incomplete factors are calculated according to the following formula:

$$IFS = (1/RS)/(1-FS)*100 \quad (9)$$

IFS is the *incomplete factor score* which is added to the number of factors. RS is the *remainder score:* the TTR at the point where the analysis of the text ends. FS is the *factor score*, which in these analyses is set at 0.71. We shall discuss the factor size of 0.71 below; however, since one of the reasons for such a score stems from the issue of incomplete factors, it is this issue which must be addressed first.

A text, quite obviously, will rarely end at exactly a completed factor. There will nearly always be some remaining text left after the MTLD analysis is complete. For example, suppose that the first factor of a text has been truncated and forms 81 words; the second factor has also been truncated and forms 120 words. After a further search of the remaining 43 words, a TTR of only .79 has been reached – this is not a complete factor, so the question becomes how to treat it, or score it?

MSTTR (see Chapter 2) ignores such extra data as it has no way to translate smaller MSTTR segments into larger ones. An MSTTR (1000), for example, which is dealing with a text of 1500 words would look at the first 1000 tokens and then ignores the final third of the text. Were a MSTTR (500) to be computed for this final third, the score would be significantly higher than the MSTTR (1000) owing to the inherent problem of

TTR. Thus, no mean can be found for MSTTR and the extra data must be discarded. In this typical example, therefore, we are left with a LD score that is only measuring two-thirds of the text.

MTLD is different; it does not ignore any data. Incomplete factors are included in the analysis as they are an intrinsic part of the text, and as they often form extremely large parts of the texts – meaning that their exclusion would be rendering the LD score a measure of something other than the object text.

In calculating a score for remainder factors, the MTLD software first of all *smoothes* the final TTR score (for example, see Jarvis, 2002; Tweedie & Baayen, 1998). The process of *smoothing* scores, as explained in Chapter 2, is the recognition that blips occur in TTR measurements. As any one measure may be somewhat higher or lower than its surrounding area, the MTLD software takes an average of the final three TTR scores. The factor size attributed to the MTLD is then calculated based on the how far the TTR score has progressed towards the designated factor TTR. For instance, our above hypothetical example found two complete factors (see Table 8). For Factor 1, a TTR of 0.70 was reached, this is below 0.71 and so the text was cut at this point. For Factor 2, a TTR of 0.698 was reached, this is also below 0.71 and so the text was cut once more. The remaining text actually reached 0.79; however, the last three TTRs averaged to .80 and so for Factor 3 we have a TTR that is 68.97% of a factor, or, a factor of 0.69. Consequently, such a text would be comprised of 244 tokens, which at a default 0.71 MTLD gives it 2.69 factors, thus resulting in an initial MTLD score of 244/2.69 = 90.71.

Table 8

Calculating the Remainder Score for a Hypothetical Text

Factor	TTR	Factor count	Number of tokens
1	.70	1	81
2	.698	2	120
3	.80	0. 69	43

The process of calculating textual remainders was introduced to increase the reliability of MTLD. Theoretically, it stands to reason that including a score for textual remainders aids the authenticity of a textual measure; however, we must once more state that whether greater or lesser weight should be given to this remainder is a matter for many subsequent tests. The calculation of this remainder has also helped to determine the default factor size of 0.71, the derivation of which we can now explain in greater depth.

Why Is a Factor Size of TTR 0.71 Used?

A factor size of 0.71 has been the primary TTR score used in calibrating the MTLD tool. Thus far, such a factor size has supplied both reliable and sensitive results. By 'reliable' (as with the tests in Chapter 3) we mean that a smaller sections of a text can

be separated from the overall text and, once reconstituted, a significantly similar combined score is reached. By 'sensitive' we mean that there are significant differences between compared texts, such that, for example, *moderately low* diversity is distinct from *low* diversity. Though a factor size of 0.71 has regularly provided such output; factor sizes between 0.65 to 0.73 have also offered highly similar results. Empirical testing of 0.71 against a variety of other potential TTR factors (from 0.65 to 0.73) has shown strong correlations ($r = 0.82$ to $0.94, p < 0.01$). This evidence suggests that far more testing of MTLD is needed before an optimal factor size can be decided. However, as the strong correlations do not suggest that a factor size of 0.71 is an inappropriate size, and as an initial test of MTLD's reliability is the first priority in this study, we have decided to persist with this size for the present study.

The similar results derived from a TTR range of 0.65 to 0.73 is not so surprising when we consider what falls on either side of these numbers. Firstly, when considering the higher side (> 0.75), we must remember that TTRs fluctuate greatly over their initial stages: a single clause may cause a TTR from a short text to lurch up or down by considerable margins. As such, chunking texts above 0.75 invites the possibilities that a non-characteristic chunk of text is taken as a factor. That is to say, there is the danger that it is not the text per se that is being chunked but rather a non-representative syntactical constraint, or a rhetorical move. Which is to say, there is a danger that the LD of the grammatical structure of English is being measured rather than the author's use of lexicon. For example, consider possibilities such as "that that one is not as good as this one ..." which contains ten tokens but a mere six types. The upshot of this is that high factor values (such as 0.80) are likely to inaccurately increase MTLD scores (see Figure 7

for an example). On the other hand, an advantage of higher TTR scores as the default factor size is that much smaller remainders (in terms of tokens) would be likely to occur. The trade-off, however, is prohibitive as what is gained from the lower remainders is lost by the inaccuracies of the factoring.

On the lower side of the optimal range we have 0.65. Such a factor score is inviting because it falls immediately below a critical value of 0.66 – a factor size which represent the lowest likely factor size for any probable given sequence above two words. For example, while a section of text may have two consecutively identical words (*that that*, or *had had*) it is extremely unlikely to have three. Consequently, the third word of any sequence guarantees that the TTR is at least 0.667. Unfortunately, the lower the TTR factor score, the more tokens the factor requires. That is, lowering a TTR from any one point to another point requires more and more tokens for each point. Thus, to move a TTR from 0.69 to 0.68 would generally require more tokens than would be required to move the TTR score from 0.70 to 0.69. This results in two main problems. Firstly, there is a lowering in the number of factors: a low TTR factor score simply burns up many tokens, and the more tokens that are burnt, the less tokens there are available to characterize the text. Secondly, a low TTR score is likely to lead to a large remainder text. As any remainder text must be estimated, the accuracy of the measure is affected. It must be remembered that the remainder factor size is calculated by how far towards the target factor size the TTR has reached. Thus, if the target factor size was 0.70 then a TTR of 0.85 would be 50% of the factor. However, as explained above, the number of tokens required to reach from 1.00 to 0.85 is far fewer than would be the number of tokens required to reach from 0.85 to 0.70. Thus, a high TTR for a remainder should actually

over-compensate the factor score, whereas a low TTR remainder score should under-compensates. In short, the lower the default TTR score, the greater the likelihood of inaccuracies.

Such hypotheses were confirmed by empirical testing (see Figure 7). The scores calculated for Figure 7 were calculated using the texts described in chapter 3. We subtracted the MTLD score for 100 token segments from the MTLD score for the 2000 (complete) token segment. The 100 token sections provide the mean score of 20 MTLDs, whereas the 2000 token section provides for just one. If over- or under-compensation consistently appears, it will be most obvious in a comparison across these two extremes. This indeed was the case, with a factor size of 0.6 consistently scoring considerably lower normalized scores for 100 token sections than for 2000 token sections. At the other end of the extreme, a factor size of 0.85 consistently gave 2000 word sections higher scores than 100 word sections.

As can be seen from Figure 7, the Broadcasts genre (for example) appears to prefer a factor size of 0.69. This number is based on the mean differences between 100 and 2000 section sizes across all nine texts of the genre. For 0.69, broadcasts recorded a mean difference of just 2.46, whereas either side of this factor score 0.68 and 0.70 reported mean differences of 3.24 and 3.05 respectively. Such small differences are not significant, however, and the additional fact that three of the nine texts returned their best results for a factor score of 0.71 precludes a rush to judgment as to what a preferred factor size should be. The purpose of these results, then, is merely to show that while a range of factor sizes may provide stable results for MTLD, there are also clearly areas outside of this range that will not suffice. A general factor size of less than 0.65 or greater

than 0.73 certainly seems inappropriate. Having said this, highly diverse genres such as Science Fiction actually perform best with a factor size as high as 0.75. If a factor size of 0.67 is applied to such texts a large inaccuracy between sections of 100 tokens and tokens appears.

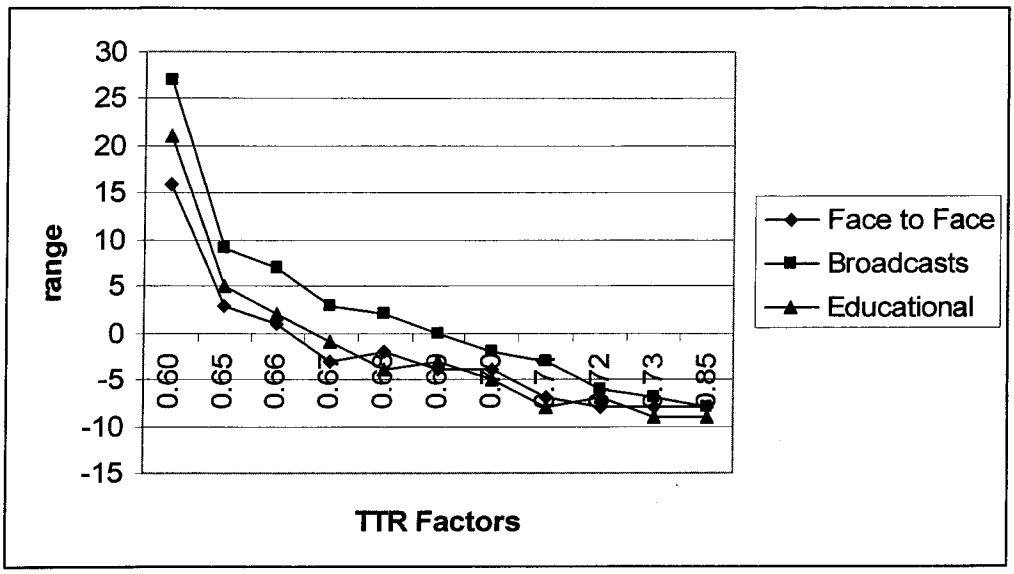

Figure 7. A range of TTR Factor Sizes Against Degree of Difference Scores for Three Different Genres

As a consequence of such testing, 0.71 was selected as the default factor size for the initial testing conducted here. 0.71 may not be a factor size that suits any one genre, nor may it turn out to be the mean of a variety of genres; however, as the results produced below suggest, 0.71 does seem to provide a fairly stable factor size that produces reliable

results. Ultimately, of course, a default factor size needs to be founded on stronger empirical evidence; however, once again we stress the importance of testing the measure for reliability and sensitivity before we move on to questions of honing the calibrations.

Why Is There a Minimal Factor Size?

The MTLD program restricts factors from containing less than 10 words. This restriction exists because, as explained above, TTRs have their greatest volatility within their initial stage. For example, given a clause that begins "That that situation is ...," the TTR moves through a range of 1.0, 0.5, 0.66, .75. From this point the TTR would generally rise to around 0.9 before beginning to fall. Typically, however, TTR scores will continue to fluctuate during their initial stages with repetitions of types causing the TTR score to fall sharply. TTR scores, it should be remembered, do not form a smooth hyperbolic curve, at least not within the first 100 words of a typical text. Instead (see Chapter 2) there are sharp drops at the beginning of a text, followed by regular rises. It is not until a steady score of around 0.71 that a noticeable continuous curve starts to form. However, even then, there are sometimes areas of the curve which "wobble" (Malvern et al, 2004). This may occur because of context changes or rhetorical stylistics. With very long texts, of course, the curve *would* eventually appear much smoother. Factors of less than ten words, therefore, are not allowed as they may only represent a brief syntactical or rhetorical, textual blip.

Why Is There a Reverse Order Analysis of the Text?

While the textual analysis described above performed well, an additional reverse analysis of the text was found to significantly improve the accuracy of the measure and was therefore added to the overall MTLD score. A natural initial objection to such an analysis is that we do not *read* text backwards so why should we *analyze* it backwards? Such objections are clearly valid; however, a number of good reasons prevail for such a method. First, as the syntactic and lexical order of the text is preserved in such an analysis, as it is 'the text' and its structure rather than 'the meaning of the text' that is being analyzed, and as the reliability of the measure is ensured by such a procedure, the reverse analysis was deemed acceptable. Second, while the TTR of the text is calculated in reverse order, the factor itself has the same TTR score whether calculated backwards or forwards – it is, after-all, the same textual element with the same number of types and tokens.

The necessity of this additional analysis can be better understood if we consider the *forward* analysis and, in particular, the cut off point of a factor. When a factor is complete, the text is cut at that particular point. As such, the very next word in the forward cycle begins another factor and that factor begins once more at a TTR of 1.00. Naturally, where the analysis begins will dictate which word will be the final word of a factor. Thus factors, and the remainder factor, are dictated from the very first word of the analysis. Such factoring only becomes an issue if we wish to compare different sections of the text. Consider Text A, which has, say, 1000 tokens, and an analysis of such a text

where the first half (Text A1) is compared to the second half (Text A2). Factoring through Text A2, we would begin our analysis at the first token, which is the 501st token of Text A. The 501st token is extremely unlikely to be the first token of any factor in an analysis of Text A. Thus we see that a single analysis of a text gives us only one view of the text. Unfortunately, as all texts are linearly arranged, it is virtually impossible to have any other view which encompasses the entire text. By reversing the text, however, there is a virtual assurance that all words fall into at least two factors in the analysis, while, at the same time, the sequence of the text is preserved.

Of course, beginning the analysis from any token in the text apart from the first word token would also serve to offer an alternative MTLD score; however, starting another *forward* analysis from a different point would not allow the *whole* text to be analyzed sequentially. For example, starting a second forward analysis from the 35th token would mean that the first 34 tokens were not accounted for. These tokens could be treated as a second remainder; however, as the remainder scores are estimations it is not productive to increase the potential for inaccuracies.

We acknowledge that the reverse analysis will remain a controversial element of the MTLD score. However, as its inclusions provides results that are sensitive and reliable (see below), we shall continue to use it as part of the overall MTLD score.

What is the Reliability of MTLD?

MTLD was tested under exactly the same circumstances as were the LD measures described in Chapter 3: Namely, there were 23 registers, each consisting of 9 texts, each text being 2000 words in length, and each text being divided into 11 sections (100 tokens – 2000 tokens). A Pearson's correlation found that all existing LD measures significantly correlated with text length ($p < 0.01$) under the three primary conditions: all genres (AG), spoken genres (SG), and written genres (WG). In testing MTLD under these same circumstances, no significant correlations to text length were found for any of the conditions: be they individual or group.

In the WG category (consisting of 1584 items), MTLD significantly correlated with all other LD measures ($p < 0.01$). More significant, however, was that MTLD correlated very highly with the best performing measures from Chapter 3: Yule's K ($r = -.71, p < 0.01$), and Mass a^2 ($r = -.82, p < 0.01$). MTLD also performed extremely well against U ($r = .79, p < 0.01$) which Jarvis (2002) reported as the best performing LD measure. The highest correlation, however, was with D^b ($r = 0.81, p < 0.01$). Against measures with a less sensitive history of LD reliability, MTLD correlation scores were slightly lower (for example, D^a $r = .56, p < 0.01$; TTR $r = .29, p < 0.01$). That these correlations should be slightly lower is to be expected given that these measures are much more prone to the effects of text length (see Chapter 3).

In the SG category, MTLD's results were also impressive. Once again, MTLD correlated significantly to all other LD measures, but, unlike all other LD measures, it did

not correlate with text length. The highest correlation was to U ($r = .80, p < 0.01$), and the lowest complex LD measure was to D^b (r (693) = .11, $p < 0.01$). All other measures correlated at highly significant levels ($p < 0.01$). As would be expected, TTR did correlate to MTLD but not remotely as significantly as other measures ($r = .09, p < 0.05$). With D^b, for example, the TTR correlation was $r = 0.18$; $p < 0.01$; and with U it was 0.47, $P < 0.01$. TTR's low correlation with MTLD and its relatively low correlation to D^b compared with its much higher correlation with other LD measures supplies further evidence that TTR is strongly affected by text length, and that all other complex LD measures are themselves somewhat affected.

For the AG category, the story remained the same. MTLD correlated significantly to all complex LD measures (a^2; r = (2277) -.86; $p < 0.01$; D^b; r = (2277) .70; $p < 0.01$); somewhat lower with TTR (r = (2277) .33; $p < 0.01$; and not at all to number of tokens.

At the individual genre level, the best performing previous measure of LD was Yule's K. This was able to avoid correlations with text length for 11 or the 23 genres, and was 66% effective for written genres. MTLD, however, managed to avoid correlations with text length for all spoken and written genres.

This collection of results suggests that our first requirement for a reliable measure of LD has been met. That is, MTLD does not significantly correlate with text length.

Is MTLD Sensitive?

Our second requirement for a useful measure of LD was that MTLD should be sensitive. We believe the sensitivity of the measure can be suitably demonstrated by comparing MTLD results to the results of the two extreme forms of mean-segmental-type-token-ratios (MSTTR) used in this dissertation: MSTTR (100) and MSTTR (2000). In Chapter 2 we showed that the smaller the section size of MSTTR, the less it was affected by Heap's Law; however, the smaller the MSTTR, the less the opportunity for divergence in TTR scores. Thus, MSTTRs increase in reliability as they decrease in sensitivity, and visa-versa.

Taking the *Broadcasts* register and the *Educational written* registers as representative genres, we can verify MTLD's degree of sensitivity by comparing MTLD scores to MSTTR scores at either extreme. Such a test provided the following results which, we believe, offers compelling evidence for the sensitivity of MTLD.

1. As MSTTR (100) is highly accurate, MTLD would need to correlate significantly with the rank ordering of the 18 texts comprising the two registers. A Spearman's rank order test showed that this indeed was the case (r (17) = .89, $p < 0.01$).

2. As MSTTR (2000) is highly *in*accurate, owing to the inherent problem of TTR, MTLD would need to have no correlation with the rank ordering of the 18 texts

comprising the two registers. A Spearman's rank order test showed that this too was the case.

3. For support of MSTTR hypothesis in points 1) and 2), we would hypothesize that MSTTR (100) would not correlate with MSTTR (2000) – as the former is accurate but the latter is not. Once more, this was the case: no significant correlation.

4. Unlike MSTTR, however, we hypothesized that MTLD rankings at the 100 token level *would* correlate with MTLD rankings at the 2000 word level because both measures are highly accurate, indeed, they should be reporting extremely similar scores. This hypothesis was also strongly confirmed: a Spearman's rank order test showing (r (17) = .97, $p < 0.01$). The slight difference of scores can be accounted for by the remainder estimations.

5. Even further support of MTLD sensitivity is provided by standard deviation scores derived from the normalized scores of the MSTTR and MTLD analyses. MSTTR(100) is not sensitive, so we would expect the standard deviation of MTLD scores to be significantly higher. This indeed is the case (2.36 and 12.79 respectively). With MSTTR (2000), an extremely large range is in play, so the MSTTR measure would be far more sensitive. Even here, however, MTLD is able to supply wider results (10.47 and 15.71 respectively). Such results offer compelling evidence that MTLD is a sensitive measure of LD as it outperforms MSTTR at both extremes.

For a useful measure of LD, we also noted that two other conditions would be desirable. Both of these conditions were indeed met. Firstly, MTLD did not require any preparation in order to produce reliable results, and secondly, the text was analyzed sequentially and, therefore, the authenticity of the text was maintained. The addition of these two supplementary conditions mean that not only is MTLD a reliable and sensitive measure of lexical diversity, but also, that the lexical diversity measure produced here is actually representative of the text to which it was applied, and therefore a measure of textual diversity.

Is MTLD Affective Over a Wider Range Than Other LD Measures?

While MTLD appears to have shown itself to be an effective measure of LD over a range of 100 to 2,000 tokens, other LD measures may actually also be effective over narrower bands. Indeed, it is possible that some more established measures of LD are more effective over particular lengths of texts.

To address this issue, we used a one-way ANOVA polynomial test to track the linear trend of the LD scores over the ever increasing token-section size for the all categories condition of texts (see Chapter 3). This a priori test was supplemented with post hoc Bonferroni tests in order to see where in the section size progression non-significant differences in output could be identified.

Based on the correlation results from Chapter 3 and previous findings (Jarvis, 2002; Malvern et al., 2004), as well as the results on MTLD from this chapter, we predicted that MTLD would show no significant linear progression. We also predicted that all other measures would show a significant linear progression, but that the more successful measures from Chapter 3 (Yule's K, U, D^b) would be less marked, whereas non-complex measures such as TTR would be more marked. The results (see Table 9) very much confirmed this.

As can be seen in Table 9, MTLD displays no significant linear progression. All other measures progress consistently and significantly. As predicted, Yule's K was the least marked, with a^2, D^b, and U all having relatively low, though highly significant, F values. Interestingly, Robet's K had the highest F value suggesting that its relation to text length may be even stronger than that of TTR.

Table 9

ANOVA Results for Contrasts by Measure Across all 2277 Section Sizes of Texts

Measure	*F*	*P*
MTLD	0.2	Ns
Yule's K	5.99	< 0.01
Mass a^2	11.19	< 0.01
D^b	13.26	< 0.01

(table continues)

Table 9 (cont'd)

Measure	*F*	*P*
U	20.5	< 0.01
Somer's S	79.65	< 0.01
D[a]	103.76	< 0.01
M	140.81	< 0.01
S	158.81	< 0.01
H	263.37	< 0.01
W	528.4	< 0.01
CTTR	731.29	< 0.01
RTTR	731.31	< 0.01
TTR	740.2	< 0.01
Robet's K	941.76	< 0.01

The post hoc Bonferroni served to identify where in the 100-2000 word range a significant progression appeared (see Table 10).

Table 10

The Four Best Performing Optimal LD Measure Ranges

	range 1	range 2	range 3	range 4
MTLD	100-2000	-	-	-
Yule's K	100-500	154-666	250-1000	400-2000
Mass a^2	100-333	154-2000	-	-
D^b	100-1000	-	-	-
U	154-1000	200-2000	-	-
Somer's S	154-286	200-400	250-500	333-666
D^a	100-333	154-400	200-500	333-666
M	154-250	200-333	250-400	286-500
S	154-250	200-333	250-400	333-500
H	200-286	250-333	286-400	333-500
W	250-333	none	none	none
CTTR	250-333	none	none	none
RTTR	250-333	none	none	none
TTR	none	none	none	none
Robet's K	none	none	none	none

There are two immediate conclusions to be drawn from Table 10. First, MTLD appears to be an affective measure across all ranges in this study: under no condition (section size) was there a significant difference between MTLD scores. Second, the measure D^b appears to have outperformed even its creators' range of effectiveness: that is, texts from 100 to 1000 words may be able to be meaningfully compared. The second observation, however, is a very generous interpretation of the results and requires further explanation.

D^b displays no significant differences between 100 and 400 tokens; however, between 100 and 500 tokens the difference is significant ($p < 0.01$). D^b again shows no significant difference between 200-666 tokens, but at 1000 tokens the difference is significant ($p < 0.01$). Between 400-1000 words there is, once more, no significant difference; however between 400 and 2000 tokens the difference is significant ($p < 0.01$). Table 10 offers a tentative acceptance of the 100-1000 word range as there is no significant difference between the D^b score for 400 tokens and those from 100 to 1000 tokens. However, this only suggest that texts of 400 tokens might be compared to texts of lower or greater token size; it does not mean that texts of 100 tokens can be meaningfully compared to texts of 1000 tokens (as, for example, MTLD appears able to do).

For a better idea of which range of text sizes can be meaningfully compared, we have to look at Table 11. As many measures have already demonstrated the narrowness of the effectiveness (see Table 10), and as MTLD appears effective across all ranges,

Table 11 only looks at the effective ranges for the best performing LD measures from Chapter 3.

Table 11

The Four Best Performing Ranges for the Best Performing LD Measures from Chapter 3

	range 1	range 2	Range 3	range 4
Yule's K	100-500	154-666	250-1000	400-2000
Mass a^2	100-154	154-333	200-666	250-2000
D^b	100-400	200-500	250-666	400-1000
U	154-250	200-500	254-1000	286-2000
D^a	100-200	154-286	200-333	250-400

Table 11 is quite revealing. First of all, we see that Yule's K appears effective from as little as 100 tokens, and manages to cover quite large ranges of differences. Mass a^2 is also quite effective but at the lower end of token count (the difficult end to judge), a^2 only seems able to compare 100 to 154 tokens. This suggests that Yule's K would be a better measure over shorter texts than a^2. U is also quite effective, but does not start to perform reliably until the 154 token size. Consequently, for shorter texts such as child speech, U might not be the best LD measure. Finally, we see some support for Malvern et al. (2004). Both D^a and D^b perform reliably for shorter text lengths; however, D^b manages

to compare, without significant difference, texts from 100 to 400 words – very much in keeping with the prediction of Malvern et al. (2004).

The ranges described here, namely those that avoid significant differences, we shall refer to as *stable*. By *stable*, we mean that researchers may be able to meaningfully compare texts that fall within these areas without undue concern over text length. Naturally, far greater testing will increase confidence in such a claim; however, the results of the linear analysis conducted here certainly offer an initial hope that ranges of LD effectiveness can be established.

What Percentage of the LD Scores Can Be Accounted for by Text Length?

We have suggested above the possible ranges of stability for the various LD measures. Such limits are supported by the variance that can be accounted for by text length (see Table 12 and Table 13).

As Table 12 shows, in this simple linear regression, none of the MTLD scores over the full range of texts can be explained by text length. This is compelling evidence once more that MTLD is not affected by the number of words being analyzed, at least for the ranges investigated here. The variance also strongly supports previous claims that Yule's K, Mass a^2, U and D^b all perform well and in a similar band, even though all variances are significant ($p < 0.01$). Rubet's K (a log adjusted LD score) joins the TTR variants as the LD measure that has the most variance explained by text length (67%).

Table 12

Variance Explained by Text Length (100-2000 Tokens) for all LD Measures

	DF	R^2	*F*	*P*
MTLD	2276	0	0.01	ns
Yule's K	2276	0.02	47.76	< 0.01
Mass a^2	2276	0.02	51.61	< 0.01
U	2276	0.03	76.73	< 0.01
D^b	2276	0.05	115.23	< 0.01
Somer's S	2276	0.2	558.74	< 0.01
D^a	2276	0.3	977.55	< 0.01
M	2276	0.32	1054.99	< 0.01
S	2276	0.32	1084.81	< 0.01
H	2276	0.43	1687.59	< 0.01
W	2276	0.44	1797.59	< 0.01
TTR	2276	0.59	3285.43	< 0.01
CTTR	2276	0.67	4598.83	< 0.01
RTTR	2276	0.67	4599.38	< 0.01
Rubet's K	2276	0.67	4622.4	< 0.01

In Table 13, we see variance explained by text length for the same measures over the same texts but only for a range of 100-500 tokens. Again, the shorter ranges have always been the harder area for LD measures to account for (Malvern et al., 2004). Measures such as the Ds claim to be reliable over such expanses even if they are not so over longer texts (McKee et al., 2000). Table 13 goes some way to supporting these

claims as only 2% of the variance can be explained by text length for D^b, albeit significant ($p < 0.01$). Once again, however, it is MTLD that resists any accounting for by text length, and Yule's K that is the best performing alternative measure.

Table 13

Variance Explained by Text Length (100-500 Tokens) for all LD Measures

	DF	R^2	*f*	*P*
MTLD	1655	0	0.3	ns
Yule's K	1655	0.01	15.86	< 0.01
D^b	1655	0.02	26.81	< 0.01
Mass a^2	1655	0.03	57.73	< 0.01
U	1655	0.06	103.02	< 0.01
D^a	1655	0.12	230.32	< 0.01
Somer's S	1655	0.15	287.33	< 0.01
M	1655	0.25	557.26	< 0.01
S	1655	0.27	618.35	< 0.01
H	1655	0.35	878.48	< 0.01
TTR	1655	0.59	2409.38	< 0.01
Rubet's K	1655	0.63	2871.9	< 0.01
W	1655	0.65	3013.33	< 0.01
RTTR	1655	0.67	3334.85	< 0.01
CTTR	1655	0.67	3336.49	< 0.01

Summary

MTLD appears to have made a promising start as a measure of LD. It has shown no correlation to text length over distances of 100-2000 tokens. It has shown no significant difference between results over the 11 interval scores. A regression of MTLD scores shows that none of the variance can be explained by text length. Despite this, however, MTLD maintains high correlations with all other major LD scores, suggesting that MTLD may be measuring lexical diversity and lexical diversity alone.

This is not to say that MTLD is without its potential problems or issues, however, and some of these problems and issues need to be highlighted. Firstly, we do not yet know the limitations of MTLD: is it reliable under 100 tokens or beyond 2000 tokens? It seems fairly safe to predict that MTLD will be effective with longer texts. However, the challenge, as ever with LD, is for greater accuracy over shorter texts. Second the factor measure of 0.71 remains somewhat arbitrary and greater theoretical and empirical evidence is needed to support it. A third issue questions the segmentation of texts into factors. Inasmuch as such partitioning occurs, has MTLD violated its own demands for textual authenticity? That is, a segmented text may remain in lexical and syntactical order, but textually it is now in pieces and, therefore, not actually the same as it was originally. Finally, while MTLD seems to have been tested successfully here, it has yet to be applied independently to any corpus with useful results. Further, while D^b, U and a^2 may have performed relatively poorly in Chapter 3, the evidence from Chapter 4 suggests that they may, after-all, prove to be useful over shorter distances. We must also remember that the D measures and U have a history of producing compelling results

(Jarvis, 2002; Malvern et al., 2004) and so writing off these measures is, as yet, premature.

In Chapter 5, we will seek to respond to some of these issues by testing MTLD further, and applying the measure to various texts in order to see how and if MTLD can be a useful tool for lexical and textual research. Specifically, the focus will be on a corpus of texts of relatively low LD with a token range from 100-400 words. Such conditions, based on the evidence thus far collected, suggest optimal conditions for a measure such as D^b, U, a^2 and Yule's K.

Chapter 5: The application of lexical diversity measures to a corpus of native and non-native speakers.

In Chapter 3 we supplied evidence that all existing LD measures correlated to text length. In Chapter 4 we saw that MTLD appears to avoid the confound of text length correlation. We also saw in Chapter 4 that measures appeared to avoid significant differences between certain ranges, the areas we regarded as *stable* (see Table 11). The measure with the largest stable area was MTLD (100 to 2000 words), although U and D^b also showed quite a wide range of effectiveness (100 to 500 and 100 to 400 words respectively).

In Chapter 5, LD measures were tested further by using a corpus of texts of relatively low LD and relatively low text length (token range 100-400 words). Such conditions, based on the evidence from Chapters 3 and 4, and on evidence supplied in the field of lexical diversity analysis (Daller et al. 2003; Jarvis, 2002; Malvern et al, 2004) are particularly difficult conditions for LD measures to supply useful data. Such conditions, however, are also the domain in which the measure D^b was specifically designed. The purposes of these tests, therefore, are to determine which LD measures inform us of lexical variation trends across low LD texts, and the extent to which they can inform us.

What Materials Were Used?

In Chapter 3 we tested LD measures across a wide corpus of texts, all of which were supplied by native-speakers (NS). For a reliable measure of LD to be established, that measure needs also to be able to analyze texts composed by non-native speakers (NNS). A measure that can reliably compare both classes of texts would be even more useful, because the assessment of a NNS's composition to a NS's composition might indicate the progress of the student towards proficiency in the target language. The corpus by Jarvis (2002), therefore, provides a perfect opportunity for just such a study. This corpus comprises 276 primary and secondary student essays. Finnish-speaking Finnish students (hereafter Finnish) produced 4 groups of 35 students each for a total of 140 essays; 70 further essays were written by Swedish-speaking Finnish primary and secondary students (hereafter Swedish) comprising two groups, and 66 essays were written by primary and secondary school native speakers (NS) of English comprising three groups (see Table 14). All essays were narratives written immediately after the students had seen the Chaplin silent movie *Modern Times*.

Jarvis (2002) used this corpus primarily to test D^a against other more established measures (see Chapter 2). Unfortunately, some of the texts were so short that they caused problems in the analysis. As we explained in Chapter 3, the LD measure Z was unable to calculate a score for some short texts. This same problem[1] exists for D^b. Further, as

[1] As explained in Chapters 2 and 3, Malvern et al. (2004) make no guarantee of the accuracy of D^b for shorter texts.

MTLD has only been tested for texts with a minimum of 100 tokens, it would be inappropriate to include shorter texts in this study. A natural consequence of this is that any conclusions drawn from the analyses conducted here may not generalize to shorter essays.

Table 14

Details of Subjects Used in this Study

Group	First language	Grade	Years English	Years Swedish	Years Finish	Number of subjects
s1	Finnish	5	2	0	-	27
s2	Finnish	7	4	0	-	29
s3	Finnish	9	6	2	-	31
s4	Finnish	9	2	6	-	35
s5	Swedish	7	2	-	4	22
s6	Swedish	9	4	-	6	35
s7	American	5	11	0	0	17
s8	American	7	13	0	0	22
s9	American	9	15	0	0	20

Which Measures Were Included in This Analysis?

Alongside MTLD, this study also includes the four best performing LD measures from chapter 4: Yule's K (K), Dugast's U (U), MASS's a^2 (a^2), and D^b. We also added D^a to the study because Jarvis (2002) had found the measure to be largely successful. We shall refer to these LD measures as "complex measures" to distinguish then from Raw TTR and other measures such as *grade level* and *number of years of English*, which are also added to this study for comparative purposes.

Which Questions Are to Be Addressed in This Analysis?

Three questions will be addressed in this chapter.

Question 1: Can convergent validity be established between MTLD and other complex LD measures?

Question 2: Do the differences in results that MTLD and other LD measures produce help us to better understand textual differences of NS and NNS students?

Question 3: Do the differences in results that MTLD and other LD measures produce help indicate the grade of the student writer?

It must be stressed that we do not offer or expect definitive answers to any of these questions in this chapter. The first and most pressing reason for conducting such analyses is the testing of the LD measures over shorter texts. It should be remembered that LD measures have a history of being most unreliable and most prone to the effects of text length over shorter texts (Malvern et al., 2004). Any indications from these analyses, therefore, will be a useful indicator of the usefulness of the various complex LD measures.

Question 1: Can convergent validity be established between MTLD and other complex LD measures?

In Chapter 4, convergent validity was established between MTLD and other complex measures in a corpus of NS texts ranging from 100 to 2000 tokens. This corpus, being considerably shorter and containing the compositions of NNSs presented a tougher challenge for MTLD – shorter texts being traditionally harder to analyze. At the same time, this same corpus provides a presumably easier challenge for measures such as D^b - D^b being designed to work best over texts of this length. Despite the difference in corpus, however, we predict that MTLD will correlate highly to all other measures of LD.

As we mentioned in Chapter 3, it will be no surprise should complete text length and all LD measures (including MTLD) correlate significantly. In this study, we are *not* comparing text length as a whole with partitioned text lengths as we did in Chapters 3 and 4. Instead we are only measuring the complete texts. For complete texts, greater length is *likely* to have greater diversity, just as it would come as no surprise if a student paper of greater length received a better grade: we can assume that those who write more generally know more about the subject, and those who know more and write more will use a wider range of vocabulary. For this reason, it should be remembered that correlations here between LD measure and text length do not, in and of themselves, indicate that any measure is necessarily text length dependent.

Results

A series of Pearson correlations confirmed that there were significant correlations among all the complex LD measures ($p < 0.01$). The highest overall correlations could be ascribed to U, followed by D^a (see Table 15). U correlated particularly highly with Mass a^2 ($r = 0.97$): significantly different ($p < 0.01$) to the highest D^a correlation which was with U ($r = 0.95$). In terms of correlations between LD measures and *number of tokens*, U was significantly lower ($p < 0.01$) than D^a. The lower U correlation to *number of tokens* suggests that U is less affected by text length, or, at least, more able to negate it, a claim which is somewhat supported by the evidence of Chapter 3 and the claims of Jarvis

(2002). As we can see from Table 16, however, there was generally no significant difference for U and D^a when these were correlated with other complex measures.

Table 15

A Pearson's 2-tail Correlation of the LD Measures Used to Assess the 236 Texts of 100 Tokens or Greater

	MTLD	D^b	D^a	U	a^2	K	Tokens	TTR
MTLD	-	0.86	0.81	0.78	0.74	0.72	0.24	0.34
D^b	0.86	-	0.9	0.87	0.86	0.86	0.3	0.35
D^a	0.81	0.9	-	0.95	0.93	0.72	0.33	0.4
U	0.78	0.87	0.95	-	0.97	0.73	0.1	0.61
a^2	0.74	0.86	0.93	0.97	-	0.75	0.17	0.55
K	0.72	0.86	0.72	0.73	0.75	-	0.18	0.37
Tokens	0.24	0.3	0.33	0.1	0.17	0.18	-	0.68
TTR	0.34	0.35	0.4	0.61	0.55	0.37	0.68	-

TTR produced the lowest overall correlations. Its highest correlation was with *number of tokens* in text ($r = 0.68$). However, and with the exception of its correlation to U, there was a significant difference between the TTR correlations with text length and its correlations to other complex LD scores ($p < 0.01$). Such results would not come as a

surprise, because TTR has been well established as a close function of text length (see Chapters' 3 and 4).

MTLD's correlations with the most consistent measures from Chapter 3 were consistently high: Yule's K ($r = -.72, p = 0.01$); Dugast's U ($r = .78, p < 0.01$); Mass a^2 ($r = -.74, p < 0.01$); D^a ($r = .81, p < 0.01$); D^b ($r = .86, p < 0.01$).

Table 16

Correlations for U and D^a Against Other Measures

Measure	U	D^a	Difference between *r* values
D^b	0.87	0.9	ns
K	0.73	0.72	ns
MTLD	0.78	0.81	ns
Tokens	0.1	0.33	< 0.01
TTR	0.61	0.4	< 0.01
Mass a^2	0.97	0.93	< 0.01

If we consider U and Mass a^2 to have the highest overall correlations then we must further consider that D^a outperformed D^b in this analysis: the correlations between D^a and both U and Mass a^2 were significantly higher than those of D^b ($p < 0.01$).

Conclusion

The strong correlations for U and D^a go a long way to supporting the conclusions of Jarvis (2002) who credited U and D^a as the best available measures of LD. The high correlations in this study for D^a, significantly higher than D^b, also supports the claims of Jarvis (2002) who argued that the newer version of D is not necessarily an improvement. That said, with all the complex LD measures offering high correlations in this initial analysis, there are few claims that can be strongly made except that MTLD is performing equally as well as more established measures.

All of the complex measures (except MTLD) performed relatively poorly in the tests conducted in Chapter 3, but they seem to have performed well in the initial test conducted here. We might speculate at this point that a basket of LD measures may be needed to produce the best interpretation of results. An alternative conclusion at this time is that tests of this nature are best interpreted by LD measures such as U and D^a. These results do tell us that MTLD is at least capturing a similar construct to K, U, D^a and Mass a^2, and very similar construct with D^b.

Thus, the answer to our first question *Can convergent validity be established between MTLD and other complex LD measures?* is yes, it can. Where the measures differ, however, becomes our next area of investigation.

Question 2: Do the differences in results that MTLD and other LD measures produce help us to better understand textual differences produced by NSs and NNSs?

A good test of LD measures can be made by comparing the essays of NSs to those of NNSs. The differences between these groups are well documented and sufficiently numerous that formulating a hypothesis of lower LD for NNSs is quite reasonable (Leki, 1993; Raimes, 2001). In the first place, NNSs especially of the grade levels used in this study, can be assumed to have less vocabulary items available. From a technical standpoint, we can also claim that NNSs are likely to exhibit lower LDs because they tend to develop their work less rigorously than NSs, NNSs, for example, take less time to plan their work, and are also less likely to scan their work with a view to improvement or development (Raimes, 2001). Instead, NNSs tend to correct only surface errors rather than develop ideas. Leki (1993) also argues that NNS are more concerned with surface form than they are with linking and developing themes, the result being a focus on grammaticality rather than diffusion of discussion of ideas.

Splitting the essays into groups of native speakers and non-native speakers, an ANOVA was used to see which LD measure best predicted differences between the measures (see Table 17).

Table 17

ANOVA Results for Measures Predicting Differences for NSs and NNSs.

Measures	*F*	η^2	*p*
Years of English	1917.94	0.88	< 0.01
U	14.48	0.05	< 0.01
Mass a[2]	13.39	0.05	< 0.01
D[a]	12.68	0.05	< 0.01
Grade	10.67	0.04	< 0.01
D[b]	9.34	0.04	0.02
MTLD	7.95	0.03	0.05
TTR	4.29	0.02	0.04
Tokens	1.05	< 0.01	ns
K	0.39	< 0.01	ns

Although offering significant results, LD measures predicting differences between NS and NNS were low (see Table 17). The best complex LD predictor of difference was the measure U (F (236) =14.48, $\eta^2 = 0.05$, $p < 0.01$); however, even TTR was able to predict differences (F (236) =4.29, $\eta^2 = 0.02$, $p = 0.04$). In fact, virtually all LD measures produced very similar results, thus the most significant finding that can be reported here was that Yule's K, the best performing measure of Chapter 3, was not significant in this analysis. The best performing non-LD measure in this analysis was the *number of years*

studying English ($F = 1917.94$, $\eta^2 = 0.88$, $p < 0.01$). This large effect is not surprising given the large difference between the native and non-native speakers: the native speakers had a minimum of 11 years of English whereas the NNS had a maximum of 6 years.

Conclusion

With the exception of Yule's K, any of the complex LD measures (including MTLD) are able to predict differences between native speakers and non-native speakers under the circumstances described here.[2] The number of words was non-significant in this test, so we have some evidence to suggest that non-native students of English wrote as much as native students. The presumable lack of linguistic skills, therefore, may not hinder productivity.[3] If this is the case, then we can hypothesize that LD measures may also not be substantially different for students of similar ages. However, we do not yet know whether this is the case for all ages in this study, or just for certain groups. For example, NS 9th graders may produce higher LD scores than NNS because of such composing skills as those described above. NS 7th graders with less access to composition

[2] The failure of Yule's K in this test, coupled with its success in Chapter 3 suggests that Yule's K may be an excellent measure of LD only over texts of lengths in excess of a few hundred words

[3] Though it must be remembered that these tests included only the texts that were a minimum of 100 tokens.

knowledge may well be producing similar LD results to their NNS counterparts. With this in mind, we need to turn to a more fine-grained analysis of the data. In this supplementary analysis we compared the LD results for only 9th graders (9GO) and for only 7th graders (7GO).

Question 2a: Do the differences in results that MTLD and other LD measures produce help us to better understand textual differences produced by *9th grade* NS and NNS?

Considering the 9th graders first, the 9GO group consisted of Finnish dominants with 6 years of English, Finnish dominants with 2 years of English, Swedish dominants with 4 years of English, and NS who we considered to have an average of 15 years of English.

In this test, the highest overall results are produced by MTLD ($F = 6.93$, $\eta^2 = 0.15$, $p < 0.01$), with D^b, U, Mass a^2, and D^a once again all very close behind (see Table 18). The eta squared results, although higher than the previous test, were only low to moderate. Yule's K was significant here, but was the poorest overall performing complex LD measure, only a marginally better predictor than TTR.

For non-LD measures, *number of words* is once again not significant. In fact, the lowest overall mean number of words was produced by the native speakers (see Table

19). Standard deviation for total number of words also failed to differentiate the groups in a linear pattern. This suggests that 9th graders, regardless of whether they are writing in first or second languages produce a similar amount of work. On the other hand, all LD measures increase (in terms of diversity) with rising number of years. This suggests that for 9th graders, output is stable whereas LD increases with experience of the target language.

Table 18

ANOVA Results for Measures Against 122 9th Graders

	F	η^2	*p*
MTLD	6.93	0.15	< 0.01
D[b]	6.15	0.14	< 0.01
U	6.01	0.13	< 0.01
Mass a[2]	5.87	0.13	< 0.01
D[a]	5.85	0.13	< 0.01
K	3.72	0.09	0.01
TTR	3.32	0.08	0.02
Tokens	1.19	0.03	Ns

Table 19

Mean and Standard Deviation Differences Measures Against 121 9th Graders

Years English		MTLD	D^b	Mean words	K	U	TTR	Mass	D^a
2	M	39.32	39.41	244.89	243.62	35.23	0.43	287.91	37.89
	SD	6.74	9.31	100.65	54.51	4.21	0.07	35.61	9.82
	N	35	35	35	35	35	35	35	35
4	M	46.80	46.57	274.74	212.09	37.36	0.43	270.95	44.57
	SD	8.34	8.95	77.65	46.56	4.26	0.06	30.5	11.29
	N	35	35	35	35	35	35	35	35
6	M	46	47.03	254.45	203.05	38.94	0.46	259.66	48
	SD	8.57	9.32	105.2	43.73	4.06	0.07	28.37	12.2
	N	31	31	31	31	31	31	31	31
15	M	49.37	50.15	226.7	217.37	39.78	0.48	255.9	48.9
	SD	12.95	13.27	101.05	69.02	5.37	0.07	35.73	12.93
	N	20	20	20	20	20	20	20	20

A post hoc Bonferroni analysis of results shows that while all of the LD measures predict some differences between the groups, only MTLD and D^b were able to separate 9th graders with 2 years of English from all other groups of 9th graders (see Table 20). The measures U, Mass a^2, and D^a were able to tell 9th graders of 2 years English from 9th

graders of 6 years English, and also 9th graders of 2 years English from 9th graders of 15 years English, but were unable to differentiate those with 2 years from those with 4 years – arguably the most difficult of the distinctions. The measure K was only able to differentiate those with 2 years from those with 6, while measures such as number of words and TTR failed to detect any differences.

Table 20

MTLD and D^b Differences Between 9th Graders with 2 Years of English and Other Groups of 9th Graders Totaling 121 Subjects

Measure	Years English	*F*	*P*	Measure	Years English	*F*	*P*
MTLD	4	2.13	< 0.01	D^b	4	2.38	0.02
	6	2.2	0.02		6	2.46	0.01
	15	2.5	< 0.01		15	2.79	< 0.01

Having found some evidence that 9th graders of 2 years English produce lower LD scores than students of more years experience, we checked whether any complex LD measure could distinguish students of 11,13, and 15 years of English. In these categories, none of the measures could detect a difference. This offers some evidence that major LD differences may be more likely to occur prior to 4 years of English language learning.

Question 2b: Do the differences in results that MTLD and other LD measures produce help us to better understand textual differences produced by 7^{th} *grade* NS and NNS?

The 7GO group consisted of Swedish dominants with 2 years of English, Finnish dominants with 4 years of English, and NSs whom, owing to their grade, we considered to have an average of 13 years of English. The results for 7GO were quite different from those of 9GO. While the measure of *number of words* was not significant for 9^{th} graders, it was the strongest predictor for 7^{th} graders ($F = 9.24$, $\eta^2 = .21$, $p < 0.01$) (see Table 21). Closer inspection of the results, however, shows that the *number of words* measure is not a simple linear progression: the 7^{th} graders of 2 years of English actually produced significantly more words than their counterparts with 4 years of English (see Table 22). As such, we must assume that the *number of words* measure is not helpful here. Each of the complex LD measures, on the other hand, suggested a gradual increase of LD (in terms of diversity) across the *number of years* band. MTLD, the strongest indictor of differences for the 9GO, was one of the weakest for 7GO, albeit still significant ($F = 3.60$, $\eta^2 = 0.09$, $p = 0.03$). D^a suggested the greatest differences, once again suggesting slightly stronger relations than D^b. In fact, in post hoc Bonferroni tests, only D^a could distinguish all three groups of students (see Table 23). Yule's K, once again, was the only complex measure whose results were not significant.

Table 21

ANOVA Results of Measures for 71 7th Graders in Predicting Years of English

	F	η^2	*P*
Tokens	9.24	0.21	< 0.01
D^a	8.12	0.19	< 0.01
U	7.1	0.17	< 0.01
Mass a^2	6.5	0.16	< 0.01
D^b	6.33	0.15	< 0.01
TTR	4.02	0.10	0.02
MTLD	3.6	0.09	0.03
K	2.42	0.06	ns

Table 22

Mean and Standard Deviation Differences Measures Against 71 9th Graders

Years		MTLD	D[b]	Mean words	K	U	TTR	Mass	D[a]
2	M	34.72	34.09	227.5	276.25	32.26	0.41	314.45	31.36
	SD	7.16	8.35	88.88	58.12	3.95	0.05	38.54	10
	N	22	22	22	22	22	22	22	22
4	M	39.02	38.19	184.39	239.79	34.37	0.45	296.43	35.14
	SD	8.82	9.06	55.32	64.09	4.6	0.05	42.49	10.48
	N	28	28	28	28	28	28	28	28
13	M	41.3	44.15	285.59	240.39	37.15	0.43	272.75	44.36
	SD	8.57	10.86	102.95	70.33	4.31	0.07	32.56	12.67
	N	22.	22	22	22	22	22	22	22
Total	M	38.41	38.76	228.49	251.12	34.57	0.43	294.7	36.81
	SD	8.57	10.13	91.72	65.62	4.68	0.06	41.36	12.12
	N	72	72	72	72.	72	72	72	72

Table 23

Results Showing which Measures could Distinguish which Sub-Groups of 7th Graders on Years of English

Measures	Years of English	Years of English	*F*	*p*
Words	4	13	23.36	< .01
D^a	4	13	3.12	.01
D^a	2	13	3.33	.01
U	2	13	1.30	.01
Mass a^2	13	2	11.57	.02
D^b	2	13	2.84	.02
TTR	4	13	.016	.01
MTLD	2	13	2.48	.03

Conclusions

The 9GO and 7GO results suggest that the amount of textual output neither depends on grade level nor on number of years of English. On the other hand, most complex LD measures, particularly MTLD, D^a and D^b were able to distinguish

differences. These differences were particularly significant between those subjects with two years of English and those with more. Such results allow us to suggest that LD differences are considerable up to about the level of four years of English, and, thereafter, grow only slowly. The pedagogical implications of this data suggest that beginning NNSs may not need to work on writing longer texts so much as they do on developing what it is that they write. Such a claim largely supports those of Raimes (2001) and Leki (1993); however, without far greater analyses, such claims remain tentative.

The answer to our second research question, therefore - *do the differences in results that MTLD and other LD measures produce help us to better understand textual differences of 7th and 9th grade NS and NNS students?* – appears to be yes: LD measures suggest that differences between NS and NNS do exist, especially prior to the 4 years of English level[4]. We shall return to this question later to see what those differences may be.

Question 3: Do the differences in results that LD measures produce help to indicate the grade of the student writer?

Three groups of differently aged NNSs each had 2 years of English: Finnish language 5th graders (F5), Swedish language 7th graders (S7), and Finnish language 9th graders (F9). We have claimed above the existence of differences between NS and NNS,

[4] However, MTLD is not alone in determining these results and other LD measures often appear to be at least as sensitive.

but argued that the most marked differences occur between students of two years English and those with more experience. In this test, the number of years of English is constant, but as grade increases steadily we would predict these results to show a small increase in LD across the grade level and a much more significant increase in terms of *number of words*. Though number of words was not a significant predictor of differences when subjects of all levels were compared, we would consider that subjects having only two years of English and presumably much lower vocabularies to work with, would produce greater amounts of work based on their cognitive development and abilities to express ideas.

As we can see from Table 24, the number of words does indeed rise steadily, but only MTLD shows a progressive increase in LD. The relatively strong eta squared score for MTLD (F (27) = 5.13, $\eta^2 = 0.11$, $p < 0.01$) is in line with the result for number of words (see Table 25). This, coupled with MTLD's ability to differentiate students with two years of English from all other groups (described above) is strong evidence that MTLD is a good indictor of grade level based on LD. Indeed, the only other measure that comes close to predicting differences is Mass a^2 whose results are *approaching* significance (F (27)= 2.94, $\eta^2 = 0.07$, $p = 0.06$).

Table 24

Mean and Standard Deviation Differences Measures Against 84 Subjects with two Years of English Learning

Grade		MTLD	D^b	Mean no. of words	K	U	TTR	Mass a^2	D^a
5	M	33.44	36.5	163.67	255.79	34.6	0.47	296.41	35.11
	SD	8.94	11.11	51.87	72.77	5.76	0.06	47.05	12.53
	N	27	27	27	27	27	27	27	27
7	M	34.72	34.09	227.5	276.25	32.26	0.41	314.45	31.36
	SD	7.16	8.35	88.88	58.12	3.95	0.05	38.54	10
	N	22	22	22	22	22	22	22	22
9	M	39.32	39.41	244.89	243.62	35.23	0.43	287.91	37.89
	SD	6.74	9.31	100.65	54.51	4.21	0.07	35.61	9.82
	N	35	35	35	35	35	35	35	35

Table 25

ANOVA Results for Measures Against 71 7th Graders to Predict Years of English

	F	*p*	*Eta*^{2}
Words	7.38	< 0.01	0.15
TTR	6.25	< 0.01	0.13
MTLD	5.13	< 0.01	0.11
Mass a^2	2.94	0.06	0.07
U	2.81	ns	0.06
Da	2.46	ns	0.06
Db	2.1	ns	0.05
K	1.88	ns	0.04

A post hoc Bonferroni test revealed that *number of words* could significantly differentiate between 5th and 7th graders and between 5th and 9th graders, but not between 7th and 9th graders (see Table 26). This suggests that younger students write less, but that the amount of written work plateaus. For LD, only MTLD as a complex measure was able to distinguish groups in terms of grade level. Table 26 shows that MTLD suggests differences between 5th and 9th graders and is approaching significance between 7th and 9th graders. Mass a^2, D^a, and U also approach significance in terms of differentiating 7th and 9th graders. This suggests, for students of two years of English, greater lexical diversity only begins to become established when students reach the 9th grade.

Table 26

A Post Hoc Analysis with Bonferroni Adjustment Across Groups of 5th, 7th and 9th Graders with 2 Years of English

Measure	Grade distinction	p
Tokens	5 and 7	0.03
	5 and 9	< 0.01
TTR	5 and 7	0.03
	5 and 9	0.05
MTLD	5 and 9	0.01
	7 and 9	0.09
Mass a^2	7 and 9	0.05
D^a	7 and 9	0.09
U	7 and 9	0.07

Conclusions

In this test, where *number of tokens* and LD perform very different tasks, only MTLD as a complex measure of LD was able to suitably differentiate groups of students. TTR also distinguished groups but as its results were much more akin to the results of *number of words* rather than LD measures, we once again have evidence that TTR and

token size are highly related. D^a once again outperformed D^b which again lends weight to Jarvis (2002) that perhaps Malvern and Richards were too quick to discard their original method. Yule's K, like D^b, did not yield significant results here, lending greater evidence that shorter texts are not the domain for this measure.

Our answer to question 3, therefore *Do the differences in results that MTLD and other LD measures produce help indicate the grade of the student writer?* is once again, yes. For NNSs, LD measures, particularly MTLD and *number of tokens* can significantly predict the grade of students.

In this chapter, we addressed three specific questions concerning lexical diversity. We can now address each individually with a short summary.

***Question 1**: Can convergent validity be established between MTLD and other complex LD measures?*

MTLD is similar to other complex measures of LD. Under different circumstances different measures perform differently, and it is possible that a basket of LD measures may be the best approach available for researchers. Of the measures used in this chapter, there is evidence to suggest that Yule's K is not a good measure of short texts (a few hundred words) or, alternatively, that it is not a good measure of texts written by NNSs. This may be because K depends not only on types and tokens, but on frequencies of types (see Chapter 2). It is possible that these shorter texts do not supply

the variance necessary for Yule's K to form a useful measure of LD here. There is evidence to suggest that D^a may be just as good as, and possibly even better than, D^b over data such as that presented here. The measures U and Mass a^2 performed adequately but are probably also better over longer texts or greater variance (as evidenced by Chapter 3). The measure MTLD is at least as good as any other LD measure and appears to be the least susceptible to the effects of text length.

***Question 2**: Do the differences in results that MTLD and other LD measures produce help us to better understand textual differences of NS and NNS students?*

While NNSs may well exhibit grammatical and phonological differences that can distinguish them from NSs, there is not an obvious difference in terms of raw productivity. Complex LD measures, however, particularly MTLD were able to suggest a strong distinction between those subjects with two years of English and those with more years of English.

***Question 3:** Do the differences in results that MTLD and other LD measures produce help indicate the grade of the student writer?*

Measures such as MTLD strongly indicated the grade-level of the students. It is feasible to extrapolate from this that students' LD scores may indicate the relevant grade or stage in the productive vocabulary development of a student.

Post Hoc Question

The answers to our three questions lead us to posit the following post hoc question which we will examine briefly before we go to Chapter 6: Can the differences of text length and LD reflected in these texts be explained by the contents of the texts in terms of parts of speech deployed?

For question 2 (*do the differences in results that MTLD and other LD measures produce help us to better understand textual differences for 7th and 9th grade NS and NNS students?*) our answer was that LD measures did suggest differences, especially prior to the 4 years of English level. Our post hoc question asks what those differences may be comprised of. For example, it might be hypothesized that content items such as adjectives or adverbs increase with age and ability. In fact, after analyzing all the texts in terms of POS, we found that the mean percentage of adjectives and adverbs does not significantly differ with development of age or grade (see Table 27). Instead, the change can be described in terms of proper nouns shifting to pronouns, with proper nouns decreasing at about the same rate that pronouns increase. We can therefore argue that younger and perhaps less skilled writers tend to make greater use of concrete items like proper nouns, whereas more skilled writers co-ordinate their writing with the more abstract pronouns.

Coordinating conjunctions such as "and" also appear to decrease as ability and experience in English increases. Graesser, McNamara, Louwerse and Cai (2004) claim that cohesive elements of texts such as coordinating conjunctions benefit less skilled readers, whereas they are a hindrance to more skilled readers. These data allow us to speculate that less skilled *writers* may also find cohesive conjunctions beneficial, whereas more skilled writers require them less.

Table 27

The Parts of Speech as a Percentage of Overall Tokens per Group

Years of English	Grade	proper nouns	Pronouns	Coordinating conjunctions
2	5	5.74	11.82	8.01
2	7	4.07	13.43	7.51
2	9	4.22	13.13	7.42
4	7	4.42	13.78	7.72
4	9	3.28	14.67	7.21
6	9	4.49	12.69	7.71
11	5	2.52	14.91	8.42
13	7	1.53	14.7	6.74
15	9	0.98	15.16	6.22

The primary function of this chapter was to further test the measure MTLD, and to apply LD measures to textual data in order to see how LD may inform teachers and researchers. MTLD performed at least as well as other LD measures and also helped to suggest many aspects of NS and NNS texts. The results of these tests, therefore, have lead to many questions which need to be addressed in future studies. The following questions form just some of the areas that may be investigated.

1. At what rate does LD increase with grade level, and at what stage does it plateau?

2. Does the increase and plateau of LD rates correlate to length of texts produced?

3. Does any other feature of language, such as cohesive devices, correlate to these developments?

4. Do these developments apply equally to NSs and NNSs?

While many measures will help to answer these questions, we believe that MTLD will form an important part of the analysis.

Chapter 6: Conclusion

What Was the Purpose of This Dissertation?

The purpose of this dissertation was to assess reliability and sensitivity of the most well-known measures of lexical diversity against a wide corpus of texts. Reliability of shorter length texts (100 to 2000 words and 100 to 500 words) would be the ultimate goal because these lengths have proven to be the most difficult to measure (Malvern et al., 2004). If none of the established measures proved to be reliable, then we would propose a new measure of lexical diversity (MTLD) which, we believed, *could* provide reliability. If MTLD's reliability was confirmed, then it was also important to establish the measure as being sensitive enough to be useful for researchers to analyze and compare texts and for teachers to assess texts.

Why Was This Research Necessary?

The need for reliability in LD measures has long been acknowledged (Malvern & Richards, 1997; Yule, 1944), not least because of its use in fields as diverse as *stylistics*, (Smith & Kelly, 2002), *neuropathology* (Bucks et al., 1999), *language acquisition*,

(Singh, 2001); and *forensic linguistics* (Colwell et al., 2002). Unfortunately, traditional measures of LD, such as TTR, have repeatedly been shown to provide questionable results, especially when texts of different lengths are compared (Daller et al., 2003; McKee et al., 2000; Tweedie & Baayen, 1998).

The lack of reliability from existing LD measures has left some researchers seriously questioning the strength of many studies' conclusions. Malvern et al. (2004), for example, list Le Normand and Cohen (1999), Ouellet et al. (2000), and Delaney-Black et al. (2000) as being prime candidates for publishing highly debatable conclusions as their work is based on misinformed understandings of the problem of LD. To these names we might add a plentiful supply of other papers such as Miller (1981) and Ertmer et al. (2003)[1]. As LD is used so often and relied on so much, from theory to teaching to patient treatment, it is necessary that the conclusions LD helps to form are based on the most reliable measures available. Malvern et al. (2004) make this exact case themselves: arguing the single sentence claim that "These things matter." We too believe that these things matter; and we further believe that what has been shown in this dissertation should go a long way to ensuring a future with a greater use of LD and a greater claim to the strength of the results produced by it.

[1] See Chapter 2 for details

What Did the Dissertation Show?

In Chapter 2 we outlined many approaches to measuring LD. These were correcting models such as Carrol's TTR and log measures of LD such as U and a^2. These were qualitative supplements, such as those proposed by Daller et al. (2003). These were frequency approaches such as Yule's K. And these were the most recent development of LD measures, the two curve fitting approaches of the measure D (Malvern et al., 2004).

Each of these measures has been tested on numerous occasions in the past (Jarvis, 2002; Owen & Leonard, 2002; Tweedie & Baayen, 1998), and none, outside from their own creators, have received resounding approval. This is not to say that any measure is without its uses, but it is to say that no existing LD measure has proven itself over the toughest terrain: returning reliable and sensitive results over short texts of varying register.

In Chapter 3 we carried out just such tests. The best known and most widely used measures of LD were tested against a corpus of 23 registers, consisting of 207 overall texts and comprising 414,000 words. The results of these tests showed that each of the measures correlated significantly to text length. This was the case for both spoken and written registers, as well as a combination of both. We concluded from these results that none of the existing measures of LD could reliably account for variations in text length under the specified testing conditions established in Chapter 3.

In Chapter 4 we introduced a new measure of LD called the measure of textual, lexical diversity (MTLD). We demonstrated that MTLD calculates LD through averaging consecutive sections of text, each with the same TTR value. We theorized that by fixing the factors to a constant TTR, counting the number of factors, and then normalizing by dividing the number of factors by the text length, we would go a long way to neutralizing the text length effect that has plagued LD measures. We tested this theory by measuring MTLD against the same corpus as was used in Chapter 3. We found that, as predicted, neither the written nor the spoken corpus of texts correlated with text length. We also found that a combination of the two corpora did not produce correlations with text length. From this we concluded that MTLD had provided the soundest evidence yet of LD reliability. We further tested MTLD against the 23 individual registers and yet again found no correlation with text length. Such results suggest that MTLD is able to produce reliable results over an extremely wide range of genres. As MTLD also correlated significantly with all other LD measures, we are further able to claim that MTLD is not only avoiding text length correlation, but that it is reporting much the same results, though presumably more accurate, than existing established measures of lexical diversity.

Having provided evidence that MTLD is unlikely to be text length dependent, we then tested MTLD for sensitivity. This sensitivity was established by comparing MTLD with MSTTR (100)[2] and MSTTR (2000)[3]. In this test, MTLD first proved just as reliable

[2] Specially arranged so as to provide optimal results with texts equally divisible by 100 words.
[3] A measure that is highly sensitive but not reliable (see Chapter 2).

as MSTTR (100) though more sensitive. Against the unreliable but very sensitive MSTTR (2000), MTLD provided greater diversity in its scoring while avoiding the unreliability of the MSTTR measure.

Next in Chapter 4, we used linear progressions to supply evidence for the optimal textual ranges for LD measures. We showed that measures such as D^b and Yule's K performed well, with neither supplying significantly different results over distances from 100 to 500 tokens when tested against all registers. MTLD, on the other hand, showed no significant differences over any distances (100-2000 tokens). We concluded from this that LD measures such as D^b and Yule's K may be of use when comparing texts within a few hundred tokens of each other in length, while measures of MTLD might measure not only shorter texts but a comparison of longer to shorter texts. MTLD even appears to be able to reliably compare texts that are around 2000 words different in length.

Finally in Chapter 4, we ran simple linear regressions to establish the amount of variance from LD measures which could be explained by text length. All measures, apart from MTLD, had significant variance explained by text length (from 2% to 67%). The combination of the linear progression analyses and the regression analyses lead us to believe that MTLD was probably the most reliable measure of LD; however, there was some evidence that measures such as Yule's K, U, Db, and Mass a2 may also be useful indicators of LD over shorter ranges.

In Chapter 5 we applied these better performing measures to just such a corpus with shorter ranges– a corpus consisting of complete essays of 100 to 500 words produced by NSs and NNSs students. Such a corpus not only provided a stern test for all LD measures, it also allowed us to see whether LD measures could inform researchers and teachers of the progress of language students by grade and years of study. Such a test would also supply further evidence of convergent validity if MTLD results could be adequately compared to the results of existing LD measures.

We reported that MTLD performed at least as well as other LD measures and often proved the most effective measure. On some occasions, other complex measures such as Da produced the most notable results. This leads us to conclude that LD analyses performed with a basket of measures may be the most useful approach for researchers.

What Is the Significance and Implications of These Findings?

If further research supports the findings related in this dissertation then the concerns of specialists in the field of lexical diversity over confoundings caused by text length may finally be allayed (for example, see Malvern et al, 2004; Yule, 1944). With a reliable and sensitive measure of LD, researchers will be able to compare texts across different lengths, genres, and modes in order to better establish aspects such as: rate of language acquisition, rate of speech development, degree of language decay and many others.

To give a practical example, researchers might develop a database of MTLD scores for written assignments by grade and years of students of English. Teachers could then assess their own students' development against these norms. If a student consistently scores under the average for LD, then maybe the student is in need of individual, greater attention to vocabulary development. On the other hand, if a student is consistently making normal LD scores then perhaps there are grounds for such a student spending their effort on areas other than vocabulary learning. In this small way, therefore, a reliable measure of LD may allow both teacher and student to spend their time more profitably.

What Future Research Might Be Conducted Using MTLD?

MTLD has made an impressive start as a tool for use in the assessing of LD. The reliability and sensitivity of the results across the wide variety of registers and text lengths analyzed here suggest that MTLD might come to be commonly used for any research purposes that have used any form of productive LD measurement. Naturally, a reliable LD measure will be most profitable where LD measures have previously been found to be most problematic. As Malvern et al. (2004) point out, the speech of young children commonly produces relatively few tokens and generally low LD. It stands to reason, therefore, that MTLD needs to address more texts in this area if it is to establish itself as the LD measure of choice.

In stylistics, Hoover (2003) claims that existing measures of LD fail to characterize authors or their writing style and suggests such an approach will never work as long as LD measures remain unreliable and assess texts with no regard for their form. MTLD, as we have seen in Chapter 4, assesses texts sequentially and, therefore, may be an approach to attributing authorship and adjudicating writing styles. Kelly and Smith (2002) appear to have had some success using a form of MSTTR in their assessment of ancient texts. Consequently, a more reliable and more sensitive measure of LD might be able to produce better results.

Other areas of interest that could be served well by MTLD include forensic linguistics. Colwell et al. (2002) argue that fabrication leads to a greater cognitive burden which may show itself through greater diversity in vocabulary use. The greater burden for greater LD theory is supported by Carpenter and Hersh (1985) where stress rather than lying was the cause. We can hypothesize, therefore, that subjects asked to repeatedly write an account of their activities may display greater diversity when fabricating their stories than when telling the truth. Such a study would be an interesting and useful avenue to pursue.

How Else Might the MTLD Technique Be Used?

MTLD allows researchers to attain a reliable measure despite the fact that one element in the equation (types) is finite, whereas the other is potentially infinite (tokens).

Lexical diversity, however, is not the only approach to textual assessment that has such a relationship between two elements. Parts of speech (POS) and phonetic production also display finite and infinite relations and, therefore, present us with interesting alternative applications to MTLD.

A measure of syntactic diversity (SD), for example, might begin with a parse of the text. Such a parse would yield around 50 POS tags, depending on the parser used. In such a case, the tags such as *nouns*, *adverbs*, *adjectives*, *interjections*, *wh-pronouns*, and *hedges* would serve as the types. The applications for SD are as diverse as those for LD. SD measures, for example, may point towards linguistic development. As we saw in Chapter 5, there is evidence that NNS may rely more on proper nouns than on other parts of speech. A reliable SD measure may, therefore, be a useful indicator of the grade or linguistic ability of a student, or even the degree of difficulty of a text from which the student is studying.

A similar development to SD would be phonetic diversity (PD). Researchers such as Oller (2003) have painstakingly analyzed the sounds produced by babies into several over-arching groups: groups which include sounds such as raspberries and grunts. While Oller's work focuses on attempting to represent these sounds through some form of IPA, PD might seek to measure the *diversity* of such sounds. For example, greater PD of baby noises may indicate earlier development of recognizable phonemes.

What Alternative Measures Might Be Used?

The results of Chapter 5 suggest that MTLD *alone* may not be enough to tell us everything about LD. At times, measures such as U, D^a, and D^b show differences between groups that MTLD does not account for. Such results lead us to believe that a "basket of measures" approach to measuring LD may also be necessary. Such a basket approach would mean establishing a computational tool that produced a number of LD scores. Naturally, which LD measures to include for which texts over which lengths will require further research.

We also feel that while MTLD appears to be a good measure of LD, there is always the chance that other measures may yet be developed that tell us more about lexical diversity. We would like to list here some possible further approaches to the measuring of LD that might be developed in the future.

Lexical Frequency Diversity. Lexical frequency diversity echoes the work of qualitative LD researchers (see, for example, Daller et al., 2003). The difference is that instead of excluding common words (such as function words), we instead computationally group all tokens together in terms of their frequency of use. By replacing words with their frequency of general occurrence, we may be able to establish a lexical diversity score which is more in line with lexical *difficulty*. [4]

[4] Such an approach naturally opens plenty of further questions such as vocabulary specialization: a lawyer looking at a law text, for example, would be likely to better understand certain terms than a engineer looking at the same text. Such issues only make the prospect of investigation that much more intriguing.

Word frequency scores may be accessed through databases such as Celex (Baayen, Piepenbrock, & van Rijn, 1993). Such a database also provides the log score of the word, allowing for simpler and greater comparison of words. If the actual words of a text are replaced by their log frequency scores then it is theoretically feasible that a method such as MTLD would provide reliable diversity scores of the lexicon used in terms of lexical difficulty – that is, if we hold that less used words are more difficult. For example, the frequency of words such as *and* and *the* is extremely high and therefore we might group them as of the same type.

Lexical diversity through remoteness and positioning. Current LD measures explain a text in terms of its diversity as a whole; however, such a measure does not account for a reader's ability to remember that any given type has occurred before. The word *ball*, for example, may be mentioned twice in a text, but its appearances may occur in two consecutive sentences; this is quite different from another word, for example *cat*, which may also appear twice but be placed 10 or more sentences apart. The *recency effect* suggests that the reader is more likely to remember a word the more recently it appears (Bjork & Whitten, 1974; Glanzer & Cunitz, 1966; Murdock, 1962). For this reason, it may be useful to measure diversity with an eye to the distance between the instances of types.[5] By the same token, we might also consider where in the sentence a word occurs. Words positioned at the beginning or the end of a sentence have greater reading time

[5] This approach too has its problems as types may be referred to pronominally.

(Haberlandt & Graesser, 1985). For this reason, we may consider lexical diversity with an eye to reading speed.

Such approaches go beyond the realm of merely avoiding text length correlation. However, these developments may all add to the usefulness of LD as a measure for researchers and teachers.

What Are the Limitations and Problems for MTLD as a Measure of LD?

MTLD's results thus far have been compelling, but in no way is the debate yet settled. There is a plentiful supply of questions and problems which must yet be answered. The following list, therefore, represents just some of the issues to be addressed in future research:

1. In the tests conducted here, MTLD analyzed texts split into sections from 100 to 2000 words. Linear progression analysis, regressions and correlations reported no relation to text length, but the question becomes, how low can MTLD go? In future research, we shall reanalyze, ever smaller, the texts used here to try to establish the lower limits of MTLD. Clearly, the lower the limit that can be established, the more useful the measure - if for no other reason than many areas of research examine extremely short texts: first language development for extremely young children being a prime example.

2. As we saw in chapter 4, many of the elements in the calculation of MTLD need further testing. For example, the factor size of 0.71 used in the tests conducted here is unlikely to be the optimal size for all analyses. The smoothing of the final three TTR scores for any given factor may also prove to be unnecessary or insufficient. Similarly, the reverse analysis of the text may prove to be unnecessary or insufficient. We may even find that inherent differences between certain modes of discourse (for example expository versus narrative) require different factor sizes.

Using different factor sizes for different modes means that comparisons between texts becomes far more difficult. However, other computation methods increase their accuracy by conducting analyses relevant to the mode. Latent semantic analysis (Landauer, Foltz, & Laham, 1998), for example, can measure the referential cohesion between two texts by judging the similarity of the words comprising the texts. These measures are based on a large corpus of typical words that might be found in such texts and, consequently, it is wise to use the corpus that is most closely suited to the texts one is analyzing. For the same reason, MTLD might come to have factor sizes that are relevant to typical LD scores for any given genre. Further research is, of course, greatly needed in this area.

Conclusion

LD has been used in a great many fields for the better part of a hundred years. Unfortunately, LD measurements have always been confounded by text length, rendering the results from many studies at least questionable. This dissertation tested all the established LD measures and found each to be correlated to text length. Some LD measures are undoubtedly more affected than others, but none have escaped the effect completely. This dissertation next introduced and tested a new measure of LD (MTLD). This measure was tested under the same circumstances as the existing measures and showed itself to be resistant to any correlation to text length over the lengths tested. We therefore submit, while emphasizing the need for greater testing, that MTLD will be an extremely useful measure to add to the arsenal of computational textual analysis.

References

Arnaud, P. J. L. (1984). The lexical richness of L2 written productions and the validity of vocabulary tests. In T. Culhane, C. Klein Bradley, & D. K. Stevenson (Eds.), *Practice and problems in language testing: papers from the International Symposium on Language Testing* (pp. 14-28). Colchester: University of Essex.

Avent, J.R, & Austermann, S (2003). Reciprocal scaffolding: a context for communication treatment in aphasia. *Aphasiology, 17*, 397-404

Baker, L. (1990). Review of: Parrot Easy Language Sample Analysis; Parrot Language Sample Utility. *Child Language Teaching and Therapy, 6*, 77-84.

Baker-Van Den Goorbergh, L. (1994). Computers and language analysis: theory and practice. *Child Language Teaching and Therapy, 10*, 329-48.

Beekmans, R., Eyckmans J., Janssens, V., Dufranne, M., & Van de Velde, H. (2001). Examining the Yes/No vocabulary test: some methodological issues in theory and practice. *Language Testing 18*, 235-74

Berstein Ratner N. (1988). Patterns of parental vocabulary selection in speech to very young children. *Journal of Child Language, 15*, 481-92.

Besnier, N. (1988). The Linguistic relationship of spoken and written Nukulaelae registers. *Language, 64*, 707-736.

Biber, D. (1987). A textual comparison of British and American Writing. *American Speech, 62*, 99-119.

Biber, D. (1988). *Linguistic features: algorithms and functions in Variation across speech and writing.* Cambridge: Cambridge University Press.

Bollard, P.M., Chute, P.M., Popp, A., & Parisier, S.C. (1999). Specific language growth in young children using the Clarion cochlear implant. *Annals of Otology, Rhinology and Laryngology, 177* (Suppl.), 119-123.

Bondy Bougere, M. (1969). Selected Factors in Oral Language Related to First-Grade Reading Achievement. *Reading Research Quarterly, 5*, 31-58.

Bowker, L & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora.* London and New York: Routledge.

Brown, R. (1973). *A first language: the early stages.* London: Allen & Unwin.

Broeder , P., Extra, G., & Van Hout, R. (1993). Richness and variety in the developing lexicon. In C. Perdue (Ed.), Adult language acquisition: cross-linguistic perspectives (Vol. I: Field methods, pp. 145-232). Cambridge: Cambridge University Press.

Brunet, E. (1978). *Vocabulaire de Jean Giraudoux: Structure et Evolution.* Geneva: Slatkine.

Bucks, R.S., Singh, S., Cuerden, J.M., & Wilcock, G.K. (2000). Analysis of spontaneous, conversational speech in dementia of Alzheimer type" Evaluation of an objective technique for analyzing lexical performance, *Aphasiology, 14*, 71-91.

Carrell, P.L., & Monroe, L.B. (1993). Learning styles and composition, *The modern Language Journal*, 77, 148-162.

Carroll, J. B. (1938). Diversity of vocabulary and the harmonic series law of word-frequency distribution. *Psychological Record, 2,* 379-86.

Carroll, J. B. (1964). *Language and thought.* Englewood Cliffs, NJ: Prentice-Hall.

Carpenter, R. H., & Hersh, R.E. (1985). A stylistic index of deteriorating military morale: using form in correspondence for intelligence purposes. *Language and Style, 18*, 185-91.

Chotlos, J.W. (1944). Studies in language behavior. IV. A statistical and comparative analysis of individual written language samples. *Psychological Monographs, 56*, 75-111.

Colwell, K., Hiscock, C.K., & Memon, A. (2002). *Interviewing techniques and the assessment of statement credibility. Applied Cognitive Psychology, 16*, 287-300.

Conner, C., Heiber, S., Arts, H.A., & Zwolen, T.A. (2000). Speech, vocabulary, and the education of children using cochlear implants: Oral or total communication? *Journal of Speech, Language, and Hearing Research, 43*, 1185-1204.

Cossette, A. (1994). *La Richesse Lexicale et sa Mesure.* Number 53 in Travaux de Linguistique Quantitative. Paris: Slatkine-Champion, Geneva.

Dahl, H. (1984). Lexical differences between working and resistance sessions in psychoanalysis, *Journal of Clinical Psychology, 40*, 733-737.

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). Lexical Richness in the spontaneous speech of Bilinguals, *Journal of Applied Linguistics,3,* 197-222.

Dickens, C. (1995). *A tale of Two Cities.* Longman: London.

Dizney, F. & Roskens, R.W. (1966). An investigation of certain qualitative aspects of verbalization of gifted children. *American Education Research Journal, 3* (3), 179-186.

Dugast, D. (1978). Sur quoi se fonde la notion d'etendue theoretique du vocabulaire? *Le Francais Moderne, 46*, 25-32.

Dugast, D. (1979). *Vocabulaire et Stylistque. I Theatre et Dialogue. Travaux de Linguistique Quantitative.* Geneva: Slatkine-Champion.

Duran, P., Malvern, D., Richards, B., & Chipere, N. (2004). Developmental trends in lexical diversity. *Applied Linguistics, 25*, 220-242.

Ertmer, D.J., Strong, L.M., & Sadagopan, N. (2002). Beginning to communicate after cochlear implantation: oral language development in a young child. *Journal of Speech, Language, and Hearing Research, 46*, 328-40.

Feldstein, S., Dohm, F.A., and Crown, C.L., (1993). Gender as the mediator in the perception of speech rate. *Bulletin of the Psychonomic Society, 8*, 25-32.

Flowerdew, L. (2003). A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly, 3,* 489-511

Freed, A.F., & Greenwood, A. (1996). Women, men and type of talk – what makes the difference. *Language in Society, 25*, 1-26.

Geers, A., Spehar, B., & Sedey, A. (2002). Use of speech by children from total communication problems who wear cochlear implants. *American Journal of Speech-Language Pathology*, 11, 50-8.

Grela, Bernard G. (2002). Lexical verb diversity in children with Down syndrome. *Clinical Linguistics & Phonetics, 14*, 251-263

Guiraud, H. (1954). *Les Caracteres Statistique de Vocabulaire.* Paris: Presses Universitaires de France.

Guiraud, P. (1960). *Problèmes et méthodes de la statistique linguistique.* Dordrecht: D. Reidel.

Harris Wright, H., Silverman, S.W., & Newhoff, M. (2003). Measures of lexical diversity in aphasia, *Aphasiology, 17*, 443-452.

He, A.W., & Young, R. (1998). Language proficiency interviews: a discourse approach. In R. Young and A.W. He (Eds), *Talking and testing: discourse approaches to the assessment of oral proficiency* (pp. 1-24). Amsterdam: John Benjamins.

Heaps, H.S. (1978). *Information retrieval – computational and theoretical aspects.* Academic Press.

Herdan, G. (1955). A new derivation and interpretation of Yule's characteristic K, *Zeitschrift fur Angewandte Mathematik und Physik.*

Herdan, G. (1960). *Quantitative Linguistics.* London: Butterworth.

Hess, C. W., Sefton, K.M. & Landry, R.G., (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech and Hearing Research, 29,* 129-34.

Hess, C. W., Haug, H.T., & Landry, R.G. (1989). The reliability of type-token ratios for the oral language of school age children. Journal *of Speech and Hearing Research, 32,* 536-40.

Honore, A. (1979). Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin, 7*, 172-177

Hoover, D. (2003). Another perspective on vocabulary richness. *Computers and Humanities, 37*, 151-178.

Hovy, E.H. (1993). Automated discourse generation using discourse structure relations. In Artificial Intelligence 63, Special Issue on Natural Language Processing.

International Computer Archive of Modern and Medieval English (2000). *Lancaster/Oslo/Bergen Corpus of British English* (CD-ROM).

Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing, 19*, 57-84.

Johnson, W. (1944). Studies in language behavior: I. A program of research. *Psychological Monographs*, 56, 1-15.

Johnson, M. & Tyler, A. (1998). Re-analysing the OPI: how much does it look like natural conversation? In R.Young & A.W. He (Eds), *Talking and Testing: discourse approaches to the assessment of oral proficiency* (pp. 28-51). Amsterdam: John Benjamins.

Kelly, D. (1997). Patterns in verb use by preschoolers with normal language and specific language impairment. *Applied Psycholinguistics, 18*, 199-218.

King, G. & Fletcher, P. (1993). Grammatical problems in school-age children with specific language impairment. *Clinical Linguistics and Phonetics, 7*, 339-352.

Kirby, J.R. (1988). Style, strategy, and skill in reading. In R.R. Schmeck (Ed.), *Learning Strategies and Learning Styles* (pp. 229-274). New York: Plenum.

Klee, T. (1992). Developmental and diagnostic characteristics of quantitative measures of childre's language production. *Topics in Language Disorders, 12,* 28-41.

Laufer, B. (1995). Beyond 2000. A measure of productive lexicon in a second language. In L. Eubank, L. Selinker, & M. Sharwood Smith (Eds): The Current State of Interlanguage. Studies in Honor of William E. Rutherford (265-272). Amsterdam/Philadelphia: John Benjamins.

Laufer, B. (2001). *Quantitative evaluation of vocabulary: How it can be done and what it is good for?* In Elder et al. (Eds): Experimenting with Uncertainty. Essays in honour of Alan Davies. Cambridge: Cambridge University Press.

Laufer, B. & Nation, P. (1995). Vocabulary size and use: lexical richness in L2 written production. *Applied Linguistics, 16*, 307-22.

Lawrence, G. (1984) *People types and tiger stripes: A practical guide to learning styles.* Gainsville, FL: Center for Applications of Psychological Type.

Layton, T.L. & Savino, M.A. (1987). Acquiring a communication systemby sign and speech in a child with Down Syndrome: a longitudinal investigation. *Child Language Teaching and Therapy,* 6, 59-76.

Lazaraton, A. (1992). The structural organization of a language interview: a conversation analytic perspective. *System, 20*, 373-86.

Lee, K., Cameron, C.A., Webster, S., Munro, K. Hunt, A. K., & Linton, M. J. (1995). Text cohesion in children's narrative writing. *Applied Psycholinguistics*, 16, 257-269

Linnarud, M. (1986). *Lexis in composition. A performance analysis of Sweedish learners' written English.* (Lund Studies in English 74). Malmo: Liber Forlag (CWK Gleerup).

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A.C. (2004).Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the twenty-sixth annual conference of the Cognitive Science Society* (pp. 843-848). Mahwah, NJ: Erlbaum.

MacWhinney B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd edn., Vol. 2: The database). Mahwah , NJ: Erlbaum.

Malvern, D.D. (1989). The type-token characteristic - an empirical investigation of a mathematical model for the type-token ratio. Unpublished working paper, Faculty of Education and Community Studies, University of Reading.

Malvern, D.D., & B.J. Richards. (1997). A new measure of lexical diversity. In Ryan, A. and A. Wray (Eds), Evolving Models of Language (pp. 58-71). Clevedon: Multilingual Matters.

Malvern, D.D. & B.J. Richards. (2000). Validation of a new measure of lexical diversity. In Beers, M., B. v.d. Bogaerde, G. Bol, J de Jong, & C. Rooijmans. (Eds), From Sound to Sentence: Studies on First Language Acquisition. Groningen: University of Groningen, Centre for Language and Cognition.

Malvern, D.D. Richards, B.J., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: Quantification and assessment.* Houndmills, Basingstoke, Hampshire: Palgrave Macmillan.

Mann, M.B. (1944). Studies in language behavior: III. The quantitative differentiation of samples of written language. *Psychological Monographs, 56,* 41-74.

Manschreck, T.C., Maher, B.A., & Ader, D.N. (1981). Formal thought disorder, the type token ratio, and disturbed voluntary movement in schizophrenia. *British Journal of Psychiatry, 139*, 7-15.

McEvoy, S. & Dodd, B. (1992). The communication abilities of 2- to 4-year-old twins. *European Journal of Disorders and Communication, 27*, 73-87.

McKee, G, Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literacy and Linguistic Computing, 15*, 323-337.

McCarthy, M. (1990). *Vocabulary.* Oxford: Oxford University Press.

McNamara, D.S. (2001). Reading both high and low coherence texts: Effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55,* 51-62.

McNamara, D.S., Kintsch, E., Songer, N.B., & Kintsch, W. (1996). Are good texts always better? Text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction, 14,* 1-43.

McNamara, D.S., & Kintsch, W. (1996). Learning from text: Effects of prior knowledge and text coherence. *Discourse Processes, 22,* 247-287.

Meara, P. (1978). Schizophrenic symptoms in foreign language learners. *UEA Papers in Linguistics, 7,* 22-49.

Meara, P., & Bell, H. (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 texts. *Prospect, 16*, 5-19.

Meara, P., & Buxton, B (1987). An alternative to multiple choice vocabulary tests. *Language Testing 4,* 142-51.

Meara, P., & G. Jones. (1998). Vocabulary size as a placement indicator. In P. Grunwell

(ed.): Applied Linguistics in society (pp. 80-87). London: Centre for Information on Language Teaching and Research (CILT).

Meara, P. (1992). *EFL Vocabulary Tests.* University of Wales, Swansea, UK: Centre for Applied Linguistics Studies.

Meara, P. (Ed.). (1983). *Vocabulary in a second language (1st ed.).* London: CILT.

Menard, N. (1983). *Mesure de la richesse lexicale. Theorie et verifications experimentales. Etudes stylometriques et sociolinguistiques,* Number 14 in Travaux de Linguistique Quantitative. Paris: Slatkine-Champion, Gneva, 1983.

Michea, R. (1969). Repetition et veriete dans l'emploi des mots. *Bulletin de la Societe de Linguistique de Paris*, 1-24.

Miller, J.F. (1981). Quantifying productive language disorders. In J.F. Miller (Ed.), *Research on child language disorders: a decade of progress* (pp. 211-20). Austin, TX: Pro-Ed.

Miller, J.F, (1991). Quantifying productive language disorders. In J. F. Miller (Ed.), *Research on Child Language Disorders: A Decade of Progress* (Austin: Pro-Ed).

Moder, C.L., & Halleck, G.B. (1998). Framing the language proficiency interviews as a speech event: native and non-native speakers' questions. In R.Young and A.W. He (Eds), *Talking and testing: discourse approaches to the assessment of oral proficiency* (pp. 117-46). Amsterdam: John Benjamins.

Nagy, W.E., (1998). *Teaching vocabulary to improve reading comprehension.* Newark, NJ: International Reading Association.

Orlov, Y.K. (1983). Ein model der haufigekeitsstruktur des vokabulars. In Studies on Zipf's Law (pp. 154-233). Bochum: Brockmeyer.

Osgood, C.E. (1960). Some effects of motivation on style of encoding. In T. Sebeok (Ed.), *Style in Language* (pp. 293-306). Cambridge, MA: MIT Press.

Owen, A.J., & Leonard, L.B. (2002). Lexical diversity in the spontaneous speech of children with specific language impairment: Application of D. *Journal of Speech and Hearing Research, 45*, 927-937.

Phillips, J.R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons, *Child Development, 44*, 182-185.

Ransdell, S., & Wengelin, Å (2003). Socioeconomic and sociolinguistic predictors of children's L2 and L1 writing quality. Arob@se, 1-2 , 22-29 http://www.arobase.to/somm.html

Ratner, N. & Silverman, S. (2000). Parental perceptions of children's communicative development at stuttering onset. *Journal of Speech, Language, and Hearing Research, 43*, 1252-1263.

Read, J. (1993). The development of a new measure of L2 vocabulary knowledge. *Language Testing 10*, 355-71

Reynes R., Martindale, C., & Dahl, H. (1984). Lexical differences between working and resistance sessions in psychoanalysis. *Journal of Clinical Psychology, 40*, 733-737.

Richards, B.J. (1987). Type token ratios: what do they really tell us? *Journal of Child Language 14*, 201-209.

Richards, B.J. & D.D. Malvern, 1997. *Quantifying Lexical Diversity in the study of language development. New Bulmershe Papers.* Reading: University of Reading.

Richards, B.J. & D.D. Malvern, 1998. A new research tool: Mathematical modeling in the measurement of vocabulary diversity (Award reference no. Rooo221995). Final Report to the Economic and Social Research Council, Swindon, UK.

Sanford, F. H. (1942). Speech and personality. *Psychological Bulletin, 39*, 811-45.

Sichel, H.S. (1975). On a distributive law for word frequencies. *Journal of the American Statistical Association, 70*, 542-7.

Sichel, H. S. (1986). Word frequency distributions and type-token characteristics. *Mathematical Scientist, 11,* 45-72.

Silverman, S., & Bernstein Ratner, N. (2000). Word frequency distributions and type-token characteristics. *Mathematical Scientist, 11*, 45-72.

Sing, S. (2001). A pilot study on gender differences in conversational speech on lexical richness measures. *Literary and Linguistic Computing,* 6, 251-264.

Skinner, B. F. (1937). *The distribution of associated words. Psychological Record, 1,* 71-76.

Smith, J.A., & Kelly, C. (2002). Stylistic constancy and change across literary corpora: using measures of lexical richness to date works. *Computers and the Humanities,* 36, 411-430.

Snow, C.E. (1972). Mothers' speech to children learning language. *Child Development, 43,* 549-565.

Snow, C. E. (1996). Change in child language and child linguists. In H. Coleman & L. Cameron (Eds.), Change and language (pp. 75-88). Clevedon: BAAL in association with Multilingual Matters.

Stokes, S.F., & Fletcher, P. (2000). Lexical diversity and productivity in Cantonese-speaking children with specific language impairment. *International Journal of Language and Communication Disorders, 35*, 527-41.

Stickler, K.R. (1987). *Guide to an analysis of language transcripts.* Eau Clair, WI: Thinking Publications.

Templin, M. (1957). *Certain language skills in children.* Minneapolis: University of Minneapolis Press.

Thordardottir, E.T., & Ellis Weismer, S. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *Internation Journal of Language & Communication Disorders, 36*, 221-244.

Tuldava, J. (1993). The statistical structure of a text and its readability. In L. Hrebicek and G. Altmann (Eds), *Quantitative text analysis* (pp. 215-27). Trier: Wissenschaftlicher Verlag Trier.

Tweedie, F.J., & Baayen, R.H. (1998). How variables may a constant be? Measures in lexical richness in perspective. *Computers and the Humanities, 32*, 323-352.

Van Lier, L. (1989). Reeling, writhing, drawling, stretching and fainting in coils: oral proficiency interviews as conversation. *TESOL Quarterly, 23*, 489-508.

Van Genderen, J.L., & B.F. Lock. 1977. *Testing Land-Use Map Accuracy.* Photogrammetric Engineering and Remote Sensing, 43: 1135-1137.

Vermeer, A. (1992) Exploring the Second Language Learner Lexicon. In L. Verhoeven & J.H.A.L. de Jong, (Eds) *The Construct of Language Proficiency.* Amsterdam: John Benjamins.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17*, 65-83.

Verhallen, M. & Schoonen, R. (1993). Lexical knowledge of monolingual and bilingual children. *Applied Linguistics, 14*, 344-364.

Watkins, R.V., Rice, M.L., & Moltz, C.C. (1993). Verb use by language-impaired and normally developing children. *First Language, 13*, 133-43.

Watkins R.V., Kelly, D.J., Harbers, H.M., & Hollis, W. (1995). Measuring children's lexical diversity: differentiating typical and impaired language learners. *Journal of Speech and Hearing Research, 38*, 1349-1355.

Weitzman, M. (1971). How useful is the logarithmic type-token ratio? *Journal of Linguistics 7,* 237-243

Wells, C.G. (1985). *Language development in the pre-school years.* Cambridge: Cambridge University Press.

Wimmer, G. & Altmann, G. (1999). Review article: on vocabulary richness. *Journal of Quantitative Linguistics, 6*, 1-9.

Wright, H.H., Silverman, S.S., & Newhoff, M. (2003). Measures of lexical diversity in aphasia, *Aphasiology, 17*, 443-452.

Youmans, G. (1990). Measuring lexical style and competence: The type-token vocabulary curve. *Style 24,* 584-99.

Youmans, G. (1991). A new tool for discourse analysis: the vocabulary management profile, *Language 67*, 763-789.

Young, R., & Milanovic, M. (1992). Discourse variations in oral proficiency interviews. *Studies in Second Language Acquistition, 14*, 403-24.

Yule, G.U. (1944). *The statistical study of literary vocabulary*. Cambridge: Cambridge University Press.

Zechmeister, E.B., D'Anna, C., Hall, J.W., Paus, C.H., & Smith, J.A. (1993). Metacognitive and other knowledge about the mental lexicon: Do we know how many words we know? *Applied Linguistics 14*, 188-206.