A STUDY OF ISSUES AND TECHNIQUES
FOR CREATING CORE VOCABULARY LISTS
FOR ENGLISH AS AN INTERNATIONAL LANGUAGE

BY

C. JOSEPH SORELL

A thesis
submitted to Victoria University of Wellington
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Victoria University of Wellington

2013

ABSTRACT

Core vocabulary lists have long been a tool used by language learners and instructors seeking to facilitate the initial stages of foreign language learning (Fries & Traver, 1960: 2). In the past, these lists were typically based on the intuitions of experienced educators. Even before the advent of computer technology in the mid-twentieth century, attempts were made to create such lists using objective methodologies. These efforts regularly fell short, however, and – in the end – had to be tweaked subjectively.

Now, in the 21st century, this is unfortunately still true, at least for those lists whose methodologies have been published. Given the present availability of sizable English-language corpora from around the world and affordable personal computers, this thesis seeks to fill this methodological gap by answering the research question: How can valid core vocabulary lists for English as an International Language be created?

A practical taxonomy is proposed based on Biber's (1988, 1995) multi-dimensional analysis of English texts. This taxonomy is based on correlated linguistic features and reasonably covers representative spoken and written texts in English.

The four-part main study assesses the variance in vocabulary data within each of the four key text types: interactive (face-to-face conversation), academic exposition, imaginative narrative, and general reported exposition. The variation in word types found at progressive intervals in corpora of various sizes is measured using the Dice coefficient, a coefficient originally used to measure species variation in different biotic regions (Dice, 1945).

The second study proceeds to compare the most frequent vocabulary types in each of the four text types using an equal-sized collection of each text type. Of special interest is the difference between spoken and written texts.

Though types are arguably the proper unit to investigate when comparing vocabulary variation, few learners would want to approach vocabulary learning one word type at a time (Nation & Meara, 2002; Bauer & Nation, 1993). The third study thus compares the effect reordering words as families (as opposed to types) has on core vocabulary lists. An analysis is made of the major differences resulting from grouping the members of each word family under a single headword and summing their individual frequencies.

Methods are then discussed for how core vocabulary lists of various sizes can be constructed based on the findings of these three studies. Recommendations are made regarding the size and composition of the source corpus and the core list extraction and construction methodology based on the learning objectives.

ACKNOWLEDGEMENTS

A simple reference in the bibliography would be insufficient recognition of the contribution made to my research by Prof. Stefan Gries. Without his text *Quantitative Corpus Linguistics with R* (2009e), I probably would not have finished this thesis. When the paperback edition came out in late 2010, I ordered it immediately and spent the next semester working page by page through that text and his *Statistics for Linguistics with R*. These texts are valuable resources, not only for learning how to use the R language but also for understanding corpus linguistic research.

Dr. Marc Hasenbank's one-day seminar on R sponsored by the University Teaching Development Centre at Victoria University in July of 2010 and subsequent explanations by Dr. Sasha Calhoun were especially helpful in pointing me toward R as a useful solution to the programming I needed to do for this thesis. As I would later learn, small influences can lead to significant changes.

I also wish to express special thanks to Prof. Estate Khmaladze, Dr. Yuichi Hirose, Dr. Giorgi Kvizhinadze and Dr. Haizhen Wu for their time and patience in discussing Zipf's law and word frequencies with someone whose math skills should have prevented him from even setting foot in the School of Mathematics.

Thanks go, as well, to the faculty and students of the School of Linguistics and Applied Language Studies at Victoria University for always welcoming me during my regular migrations between Taipei and Wellington, especially to the Vocabulary Discussion Group: Myq Larsen, Yosuke Sasao, Tatsuya Nakata, Michael Rogers, Betsy Quero, Tatsushiko Matsushita, Friederike Tegge and others.

One of the highlights of every stay in Wellington was the opportunity to spend each Shabbat at Temple Sinai. I want to especially thank Fred, Susan, Rick, Martin and many others for welcoming me into the community.

I owe an immense debt of gratitude to Prof. Paul Nation. His writings and the clarity with which he translates research findings into practical teaching applications were what first attracted me to Victoria University and the School of Linguistics and Applied Language Studies. At conferences in several countries, I have heard him encourage language teachers to engage in research, and to share their findings with the language learning community by publishing. To say that the writing topics he suggested to me were life-changing would not be an exaggeration. Pursuing a doctoral degree is not usually good for one's health, but the suggestion to write on Zipf's law and word frequencies lead to discoveries of how dynamic

systems work, including the human body. Prof. Nation continues to inspire my learning and teaching, and now, the teaching of many of my former students.

I also wish to thank many of my colleagues here in Taipei who regularly encouraged me and Christ's College for granting me leave on more than one occasion to focus more intently on research.

This thesis is dedicated to my wife, Shufeng, my son, Nathan, and my parents, Michael and Barbara. My parents were my first teachers and are still among the most dedicated I have ever met. During the time it took to write this thesis, Nathan has become a proficient user of two languages. Both of which, he uses to research the inner workings of computers. When my laptop began to black out from running week-long R-scripts, he single-handedly designed and assembled a more powerful computer that I have used to complete most of the important studies in this thesis. During the last few years, Shufeng has put up with a husband who spent more time in the office than at home. When I did make it home, she always made sure there were plenty of fresh fruits and vegetables to keep me going. Shufeng also made sure Nathan paid attention to things other than computers. I look forward to being able to spend more time with them now that this thesis is finished, and I have a better understanding of how research should be conducted.

TABLE OF CONTENTS

ABSTRACT.....	i
ACKNOWLEDGEMENTS.....	ii
TABLE OF CONTENTS.....	iv
LIST OF TABLES.....	x
LIST OF FIGURES.....	xiii
1.0 INTRODUCTION.....	1
1.1 Problems with vocabulary lists.....	5
1.1.1 Words are listed in an inappropriate sequence.....	5
1.1.2 Word lists are not consistent.....	7
1.1.3 Word lists are not representative.....	8
1.1.4 Which variety of English?.....	9
1.1.5 What is a word?.....	10
1.1.6 How big is the core?.....	10
1.1.7 Perceptions of word lists.....	12
1.1.8 Summary of problems with core vocabulary lists.....	12
1.2 Purpose and Significance.....	13
1.3 Research questions and hypotheses.....	14
1.4 Definition of terms.....	14
1.4.1 Core vocabulary.....	14
1.4.2 Corpus.....	14
1.4.3 General and comparable corpora.....	15
1.4.4 Word types and tokens.....	15
1.4.5 Hapax legomenon and dislegomenon.....	15
1.4.6 Lemmas and word families.....	16
1.4.7 Marginal words.....	16
1.4.8 Text type and genre.....	17
1.4.9 Word frequency, normalization, rank and range.....	17
1.4.10 Coverage.....	18
1.5 Organization of the Studies.....	18
2.0 BACKGROUND.....	20
2.1 Introduction to the background review.....	20
2.2 What is a word?.....	20

2.3	A Taxonomy of English Language Texts.....	22
2.3.1	Representiveness.....	22
2.3.2	Multidimensional analysis.....	23
2.4	Defining core vocabulary.....	27
2.5	Word lists and corpora.....	32
2.5.1	The origin of word lists.....	33
2.5.2	Early core vocabulary lists.....	34
2.5.3	Attempts at objectivity.....	37
2.5.4	Intuition versus objectivity.....	38
2.5.5	Early studies of spoken vocabulary.....	41
2.5.6	The Brown family of corpora.....	42
2.5.7	The International Corpus of English.....	42
2.5.8	The soup-pot method.....	43
2.5.9	Academic word lists.....	44
2.5.10	Mega-corpora.....	46
2.5.11	Word lists from mega-corpora.....	47
2.5.12	Ersatz conversation data.....	49
2.5.13	Modern corpora of spoken English.....	49
2.5.14	The Internet as corpus.....	50
2.5.15	Corpora and the core.....	51
2.6	Corpus similarity.....	51
2.6.1	Stylometry.....	52
2.6.2	Chi-squared.....	55
2.6.3	Standard Lexical Difference.....	58
2.7	Applications of core vocabulary lists.....	61
2.8	Conclusion to the background review.....	62
3.0	METHODOLOGY.....	64
3.1	Text Types and a Taxonomy of English of Language Texts.....	64
3.1.1	Out of order.....	64
3.1.2	A simplification of Biber's text type model.....	65
3.1.2.1	Interactive.....	66
3.1.2.2	General reported exposition.....	67
3.1.2.3	Imaginative narrative.....	68
3.1.2.4	Academic.....	68

3.1.2.5	Other text types.....	68
3.1.2.6	Key text types.....	69
3.2	Data preparation.....	69
3.2.1	Tags, annotations, etc.....	70
3.2.2	Non-ASCII characters.....	70
3.2.3	Hyphenated compounds.....	70
3.2.4	Letter-digit combinations.....	71
3.2.5	“Indigenous” vocabulary.....	71
3.2.6	Observer’s paradox.....	73
3.3	How much data is needed to give consistent results?.....	74
3.3.1	The Dice Coefficient with non-overlapping samples.....	75
3.3.2	Projections based on the Dice distance between non-overlapping samples.....	82
3.3.3	Dice distance study with replacement.....	83
3.4	How different are the text types?.....	84
3.4.1	One core vocabulary, or more?.....	84
3.4.2	Dice distance between text types.....	85
3.5	Which Unit Should Be Used For Counting Words?.....	86
3.5.1	Text samples.....	87
3.5.2	Family and type frequencies.....	87
3.6	Dispersion.....	89
3.7	Scope, limitations and assumptions.....	94
3.7.1	A tentative attempt at a core vocabulary for EIL.....	94
3.7.2	Orthographic words.....	94
3.7.3	Multiword Units.....	95
4.0	RESULTS.....	96
4.1	Dice distance between samples of the same text type.....	96
4.1.1	Dice distance without replacement.....	96
4.1.1.1	Conversation and general writing.....	100
4.1.1.2	Academic writing.....	109
4.1.1.3	Narrative writing.....	112
4.1.1.4	Conversation versus simulated dialogue.....	115
4.1.1.5	Graded writing.....	116
4.1.1.6	Projecting to larger corpus sizes.....	117

4.1.2	Dice distance with replacement.....	130
4.1.2.1	Conversation.....	136
4.1.2.2	Narrative writing.....	138
4.1.2.3	General writing.....	140
4.1.2.4	Academic writing.....	142
4.1.2.5	Projections of Dice distances with replacement.....	144
4.2	How different are the text types?.....	152
4.2.1	Dice distance of the key text types from each other.....	152
4.2.1.1	Conversation.....	153
4.2.1.2	Narrative writing.....	154
4.2.1.3	General writing.....	154
4.2.1.4	Academic writing.....	154
4.2.2	Analysis of word types shared by or unique to each text type.....	158
4.2.2.1	Distribution of marginal words.....	159
4.2.2.2	Words unique to the high-frequency word lists for each text type...	160
4.2.2.3	Vocabulary overlap among the high-frequency words of each text type.....	162
4.2.2.4	Quantifying the overlap of vocabulary between the key text types..	166
4.2.3	Where is the core?.....	169
4.2.3.1	Survival vocabulary.....	170
4.2.3.2	Kernel vocabulary.....	173
4.2.3.3	Lexical sets.....	178
4.2.3.4	Out of order words.....	180
4.3	Which Unit Should Be Used For Counting Words?.....	182
4.3.1	General writing.....	183
4.3.2	Academic writing.....	186
4.3.3	Narrative writing.....	188
4.3.4	Conversation.....	190
4.4	Constructing objective core vocabulary lists.....	193
4.4.1	Steps in constructing objective core vocabulary lists.....	193
4.4.1.1	Purpose.....	193
4.4.1.2	Data collection.....	194
4.4.1.3	Frequency lists.....	194
4.4.1.4	Lemmas or word families.....	194

4.4.1.5	Family or lemma frequency.....	194
4.4.1.6	Core list length.....	196
4.4.1.7	A joint table of the text types.....	196
4.4.1.8	Prioritizing text types.....	196
4.4.2	A trial core vocabulary list.....	196
4.4.2.1	Time of compilation.....	197
4.4.2.2	Definition of core vocabulary.....	198
4.4.2.3	Purpose of the lists.....	199
4.4.2.4	Varieties included.....	199
4.4.2.5	Tacit values.....	199
4.4.2.6	A work in progress.....	199
5.0	DISCUSSION AND CONCLUSION.....	201
5.1	A Model of the English Language.....	201
5.2	How much data is necessary to give consistent results for high-frequency vocabulary?.....	202
5.2.1	A lot!.....	202
5.2.2	The size and definition of the core vocabulary.....	205
5.3	How different are the text types?	206
5.3.1	A core or cores?.....	206
5.3.2	Narratives and proper nouns.....	209
5.3.3	Word types in spoken English, including EIL.....	209
5.3.4	Beyond the four key text types.....	210
5.4	How does the scope of the definition of a word influence the resulting list?....	212
5.4.1	Coverage and comprehension.....	212
5.4.2	Family size and other issues.....	216
5.5	Summary.....	217
5.5.1	What this study has shown.....	217
5.5.2	Applications.....	218
	BIBLIOGRAPHY.....	221
	APPENDICES.....	246
Appendix A.1	Corpus segments used to represent the <i>interactive</i> or <i>conversation</i> text type.....	246
Appendix A.2	Corpus segments used to represent the <i>narrative writing</i> text type	248
Appendix A.3	Corpus segments used to represent the <i>general writing</i> text type...	250

Appendix A.4	Corpus segments used to represent the <i>academic writing</i> text type.....	252
Appendix B.1	Sample R-scripts for “cleaning” corpus data.....	254
B.1.1	London Lund Corpus.....	254
B.1.2	Wellington Corpus of Spoken New Zealand English.....	262
B.1.3	International Corpus of English (written).....	264
Appendix B.2	R-script for combining corpus files.....	268
Appendix B.3	R-script for splitting corpus data into equal-sized blocks.....	269
Appendix C.1	Calculating Dice distance with replacement.....	274
Appendix C.2	Calculating Dice distance without replacement.....	286
Appendix D.1	Adding head words to the BNC/COCA lists.....	287
Appendix D.2	Merging text type frequency lists and the type-head word list.....	288
D.2.1	Merging the words common to both the family lists and the text type list.....	288
D.2.2	Listing words only found in the text type sample list.....	289
Appendix E.1	Multi-file frequency script.....	290
Appendix E.2	Multi-file dispersion and range script.....	293
Appendix F	Frequency lists from two general writing samples of 23.8 million tokens each.....	296

LIST OF TABLES

Table 1.1	A selection of core vocabulary lists.....	11
Table 2.1	Coverage of the Longman Defining Vocabulary in BNC genre subsections..	29
Table 2.2	A partial list of the frequency, standard deviation, and mean for the types in each of 10 samples of reportage from the International Corpus of English....	53
Table 2.3	A partial list of the z-scores (rounded) for the types in each sample from the International Corpus of English.....	53
Table 2.4	A partial frequency and z-score list for a 10,000-word sample of reportage from the BNC.....	54
Table 2.5	Delta scores for reportage samples from the BNC and ICE including each ICE sample's absolute distance from the BNC sample.....	55
Table 2.6	A matrix showing Standard Lexical Difference scores for a group of sample spoken and written corpora.	60
Table 3.1	Corpus sizes for the key text types compiled for these studies.....	69
Table 3.2	Excerpt from the ranked list of tokens in descending frequency order for two 500,000-token samples of face-to-face conversation.....	77
Table 3.3	Dice distance matrix for the top-ranked 100 types for ten 100,000-token samples composed of data from the BNC spoken demographic section.....	80
Table 3.4	Dice distance matrix for the top-ranked 200 types for ten 100,000-token samples composed of data from the BNC spoken demographic section.....	81
Table 3.5	Example calculation of Gries' DP (2010c) for a word with very uneven distribution.....	91
Table 3.6	Example calculation of Gries' DP (2010c) for a word with more even distribution.....	91
Table 3.7	Sample sizes used for a dispersion comparison of English varieties from the UK, New Zealand, North America, Hong Kong, the Philippines, Singapore, Jamaica, Ireland and a mixture from East Africa, Europe and elsewhere.....	92
Table 3.8	Examples of words with even and uneven dispersions in a 3-million-word collection of conversation from several varieties of English.....	93
Table 4.1	Example of a partial comparison of types used to calculate Dice distance....	99
Table 4.2	The most frequent n-grams of four or more words for conversation and general writing in samples from the International Corpus of English.....	106

Table 4.3	The top-ranked 25 word types on the complete frequency lists for the collections of the four key text types, a collection of OUP graded readers and the Brysbaert and New's (2009) SUBTLEXus corpus.....	156
Table 4.4	Samples of the vocabulary shared by more than one text type.....	164
Table 4.5	Matrix showing the quantity of shared or differing vocabulary for pair-wise comparisons of the 6,000 most frequent word types in each of the four key text types.....	167
Table 4.6	Matrix showing the quantity of shared or differing vocabulary for general and academic writing vis-à-vis the combined vocabulary of conversation and narrative writing.....	168
Table 4.7	Matrix showing the quantity of shared or differing vocabulary for conversation and narrative writing vis-à-vis the combined vocabulary of general and academic writing.	168
Table 4.8	Matrix showing the quantity of shared or differing vocabulary for academic writing vis-à-vis the combined vocabulary of conversation, narrative and general writing.....	169
Table 4.9	Ranks in each text type for a selection of words from Crabbe and Nation's (1991) survival vocabulary.....	171
Table 4.10	Ranks in each text type for a selection of word families from Crabbe and Nation's (1991) survival vocabulary.....	172
Table 4.11	Ranks in each text type for a selection of words from Swadesh's (1971: 283) kernel vocabulary.....	174
Table 4.12	Ranks in each text type for a selection of word families from Swadesh's (1971: 283) kernel vocabulary.....	176
Table 4.13	Ranks in each text type for the names of the days of the week.....	178
Table 4.14	Ranks in each text type for the names of the months.....	179
Table 4.15	Ranks in each text type for the names of the seasons.....	179
Table 4.16	Ranks in each text type for color terms.....	180
Table 4.17	Ranks in each text type for the words (word types) mentioned in Nation (1990: 21-24).....	181
Table 4.18	Ranks in each text type for the words (word families) mentioned in Nation (1990: 21-24).....	181
Table 4.19	Sample family headwords and members with their respective frequencies in a sample of 10 million tokens of general writing.....	185

Table 4.20	Analysis of the most frequent word 3,000 families from a 10-million-token sample of general writing.....	186
Table 4.21	Analysis of the most frequent 3,000 word types from a 10-million-token sample of general writing.....	186
Table 4.22	Analysis of the most frequent 3,000 word families from a 10-million-token sample of academic writing.....	188
Table 4.23	Analysis of the most frequent 3,000 word types from a 10-million-token sample of academic writing.....	188
Table 4.24	Analysis of the most frequent 3,000 word families from a 10-million-token sample of narrative writing.....	190
Table 4.25	Analysis of the most frequent 3,000 word types from a 10-million-token sample of narrative writing.....	190
Table 4.26	Analysis of the most frequent 3,000 word families from a 10-million-token sample of conversation.....	192
Table 4.27	Analysis of the most frequent 3,000 word types from a 10-million-token sample of conversation.....	193
Table 4.28	A portion of a pivot table showing the seven most frequent word families in a 10-million-token sample of general writing.....	195
Table 5.1	Approximate amount of variation that should be expected for various-sized frequency lists for each of the key text types using a corpus of 50 million tokens for each text type.....	203
Table 5.2	Approximate amount of variation that should be expected for various-sized frequency lists for each of the key text types using a corpus of 20 million tokens for each text type.....	204
Table 5.3	Approximate amount of variation that should be expected for various-sized frequency lists for each of the key text types using a corpus of 15 million tokens for each text type.....	204
Table 5.4	Token coverage for the most frequent 3,000 word families from a 10-million-token collection of the respective text types.....	214
Table 5.5	The percentage of proper nouns among the most frequent 3,000 word families from a 10-million-token collection of in each text type.....	214
Table 5.6	Token coverage for the most frequent 3,000 word types from a 10-million-token collection of the respective text types.....	214

Table 5.7	The percentage of proper nouns among the most frequent 3,000 word types from a 10-million-token collection of each text type.....	214
Table 5.8	The number of word families from each text type included in a combined list of the most frequent 3,000 word families from a 10-million-token collection of each text type.....	215
Table 5.9	Types covered by the most frequent 3,000 families for 10-million-token samples of each of the four key text types.....	216

LIST OF FIGURES

Figure 2.1	Biber's (1995) text types 1-4 showing the average score on each of five dimensions.....	26
Figure 2.2	Biber's (1995) text types 5-8 showing the average score on each of five dimensions.....	27
Figure 2.3	Delta scores for reportage samples from the BNC and ICE.....	55
Figure 3.1	Biber's (1995) text types 1-4 showing the average score on each of five dimensions. (Identical to Figure 2.1).....	66
Figure 3.2	Biber's (1995) text types 5-8 showing the average score on each of five dimensions. (Identical to Figure 2.2).....	66
Figure 3.3	Dr. Martha E. Apple laying out a grid to gather data for the GLORIA project – Global Observation Research Initiative in Alpine Environments monitoring network (EcoRover, 2010).	75
Figure 3.4	Illustration of the comparison of increasingly large portions of two frequency lists, calculating the Dice distance for the most frequent 100 word types, 200, 300, etc.....	79
Figure 3.5	Graph of an example comparison of the Dice distance for the most frequent 100, 200, 300 word types.....	79
Figure 3.6	Mean Dice distance scores between samples of conversation of 1 million, 2 million and 3 million tokens. Dice distance was calculated for the most frequent 100 word types and increasing in increments of 100 words until the comparisons included the top-ranked 10,000 word types.....	82

Figure 3.7	Projection from the Dice distance measures for the top-ranked 1,000 word types for samples of general writing. (Diamonds indicate data points on which the power law projection is made.).....	83
Figure 4.1	Dice distances for the four key text types for corpus samples of 3 million tokens in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.....	98
Figure 4.2	Dice distances for the four key text types for corpus samples of 3 million tokens in increments increasing by 100 words up to a rank of 10,000.....	100
Figure 4.3	Dice distances for conversation and general writing for samples of 1 million, 2 million and 3 million tokens in increments increasing by 100 words up to a rank of 10,000.....	101
Figure 4.4	Dice distances for conversation and general writing for samples of 1-3 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.....	102
Figure 4.5	Graph of the word frequencies for the Brown Corpus (Kučera & Francis, 1967).....	103
Figure 4.6	Double-logarithmic graph showing the Zipfian distribution of general writing texts (47.65 million tokens).....	104
Figure 4.7	Double-logarithmic graph showing the Zipfian distributions of conversation texts (7.05 million tokens).....	105
Figure 4.8	Dice distances for general writing samples of 1-5, 10, 15, 20 and 23.8 million tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	108
Figure 4.9	Dice distances for general writing samples of 1-5, 10, 15, 20 and 23.8 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded down to the last full 1,000 word types.....	109
Figure 4.10	Double-logarithmic graph showing the Zipfian distribution of academic writing.....	110
Figure 4.11	Dice distances for samples of academic writing from 1-8 million tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	111

Figure 4.12	Dice distances for samples of academic writing from 1-8 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.....	112
Figure 4.13	Double-logarithmic graph showing the Zipfian distribution of narrative writing.....	113
Figure 4.14	Dice distances for narrative writing for samples of 1-9 million tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	114
Figure 4.15	Dice distances for narrative writing for samples of 1-9 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.....	115
Figure 4.16	Double-logarithmic graphs showing the Zipfian distribution of the SUBTLEXus corpus.....	116
Figure 4.17	Double-logarithmic graphs showing the Zipfian distribution of a corpus of OUP graded readers (763,507 tokens).....	117
Figure 4.18	Projections from the Dice distance measures at various ranks for samples of general writing.....	119
Figure 4.19	Projections from the Dice distance measures at various ranks for samples of narrative writing.....	120
Figure 4.20	Projections from the Dice distance measures at various ranks for samples of academic writing.....	121
Figure 4.21	Projections from the Dice distance measures at various ranks for samples of conversation.....	122
Figure 4.22	Projections from the Dice distance measures the top-ranked 1,000, 2,000 and 3,000 word types for samples of all four key text types.....	123
Figure 4.23	Projected Dice distances for corpora of the four key text types for the top-ranked 1,000 words up to a sample size of 20 million tokens.....	124
Figure 4.24	Projected Dice distances for corpora of the four key text types for the top-ranked 3,000 word types up to a sample size of 20 million tokens.....	125
Figure 4.25	Projected Dice distances for corpora of the four key text types for the top-ranked 6,000 words up to a sample size of 20 million tokens.....	126

Figure 4.26	A comparison of Dice distance scores for narrative writing and conversation.....	127
Figure 4.27	A comparison of Dice distance scores for general and academic writing.....	128
Figure 4.28	A comparison of Dice distance scores for general and narrative writing.....	129
Figure 4.29	A comparison of Dice distance scores for academic and narrative writing...	130
Figure 4.30	Error estimates from sampling with replacement when measuring Dice distances for the top 10,000 words.....	132
Figure 4.31	Error estimates from using replacement when measuring Dice distances for the top 1,000 words.....	133
Figure 4.32	Error estimates from sampling with replacement when measuring Dice distances for the top 2,000 words.....	133
Figure 4.33	Error estimates from sampling with replacement when measuring Dice distances for the top 3,000 words.....	134
Figure 4.34	Error estimates from sampling with replacement when measuring Dice distances for the top 6,000 words.....	135
Figure 4.35	Error estimates from sampling with replacement when measuring Dice distances for the top 9,000 words.....	135
Figure 4.36	Mean Dice distances for 10 samples of 1-5 million tokens of conversation with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	136
Figure 4.37	Mean Dice distances for 10 samples of 1-5 million tokens of conversation with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types.....	137
Figure 4.38	A comparison of the Dice distance for conversation for samples of 1, 2 and 3 million tokens with and without replacement.....	138
Figure 4.39	Mean Dice distances for 10 samples of 1-10 million tokens of narrative writing with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	139
Figure 4.40	Mean Dice distances for 10 samples of 1-10 million tokens of narrative writing with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types.....	139

Figure 4.41	A comparison of the Dice distance samples of narrative writing for samples of 1, 2, 4 and 9 million tokens with and without replacement.....	140
Figure 4.42	Mean Dice distances for 10 samples of 1-10, 15, 20 and 25 million tokens of general writing with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	141
Figure 4.43	Mean Dice distances for 10 samples of 1-10, 15, 20 and 25 million tokens of general writing with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types.....	141
Figure 4.44	A comparison of the Dice distance samples of general writing for samples of 1, 5, 10 and 20 million tokens with and without replacement.....	142
Figure 4.45	Mean Dice distances for 10 samples of 1-10 million tokens of academic writing with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000.....	143
Figure 4.46	Mean Dice distances for 10 samples of 1-10 million tokens of academic writing with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types.....	143
Figure 4.47	A comparison of the Dice distance samples of academic writing for samples of 1, 2, 4 and 8 million tokens with and without replacement.....	144
Figure 4.48	Projected Dice distances for samples of conversation up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.....	145
Figure 4.49	Projected Dice distances for samples of narrative writing up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.....	146
Figure 4.50	Projected Dice distances for samples of general writing up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.....	147

Figure 4.51	Projected Dice distances for samples of academic writing up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.....	148
Figure 4.52	Projected Dice for top 1,000 words for the four key text types up to a size of 50 million tokens.....	149
Figure 4.53	Projected Dice for top 3,000 words for the four key text types up to a size of 50 million tokens.....	150
Figure 4.54	Projected Dice for top 6,000 words for the four key text types up to a size of 50 million tokens.....	151
Figure 4.55	Projected Dice for top 9,000 words for the four key text types up to a size of 50 million tokens.....	152
Figure 4.56	Radar diagram showing the Dice distances between the top 1,000 words of each of the key text types.....	153
Figure 4.57	Radar diagram showing the Dice distances between the top 6,000 words of each of the key text types.....	157
Figure 4.58	A dendogram showing the hierarchical clustering of the key text types using the complete-linkage or furthest-neighbor method based on the Dice distances between the top 1,000 words of each of text type.....	158
Figure 4.59	Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of general writing.....	183
Figure 4.60	Tokens covered by the most frequent 3,000 types and families for a 10-million-token sample of general writing.....	184
Figure 4.61	Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of academic writing.....	187
Figure 4.62	Tokens covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of academic writing.....	187
Figure 4.63	Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of narrative writing.	189
Figure 4.64	Tokens covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of narrative writing.....	189
Figure 4.65	Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of conversation.....	191

Figure 4.66	Tokens covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of conversation.....	192
Figure 5.1	Visualization of the relationship of the four key text types.....	207

CHAPTER 1

INTRODUCTION

The research presented in this thesis started as an applied linguistic project looking for practical solutions to prioritizing vocabulary learning in an EFL curriculum. What was unexpected was the way the research turned out to be intertwined with research into everything from physics and mathematics to biology and economics. Delving into the fundamental question of why there is a core vocabulary reshaped my view of language, language acquisition, and even the nature of life and cognition.

As an EFL instructor at a tertiary institution in Taiwan, one of the biggest struggles my students face is the large quantity of unknown vocabulary they encounter in their required reading and listening. One of the first steps I took in trying to improve that aspect of my teaching was to reorder the vocabulary learning objectives based on a version of West's *General Service List of English Words* (1953) that has been reorganized into word families by Paul Nation, who would later become my thesis advisor. This year marks the 60th anniversary of the publication of West's GSL. Much has changed in society, technology and in the research techniques available since West's list was first published in 1953. His list was based largely on the work of the Carnegie Conferences of 1934 and 1935, and it reflects a world that has changed in many ways.

In the early 20th century, the majority of people on earth lived in rural communities. Communication with people outside of one's town or village was either costly or time-consuming. Posting a letter could take days or weeks to arrive. A telegram was faster, but the price was calculated per character. In the industrialized world, telephone landlines had just recently become common, but were far from universal.

Today, we call or instant message a broad group of friends, acquaintances and colleagues from our pads or cell phones throughout the day. Even sitting down at a desktop computer to send an email can seem slow. A glance at a handheld screen tells us immediately how many of the people in our global network are accessible at that very moment. We even have tools to quantify how networked we are by looking at the number of friends we have on Facebook or contacts we have on LinkedIn or Academia.edu. This technology has raised our awareness of our social networks and, at least superficially, enlarged them. Language is the medium of communication in these social networks, and research into the structure of networks has also led to a better understanding of language itself.

In the past, social contacts were largely determined by geography, and interactions were almost exclusively face-to-face. Face-to-face conversation is rapid, multi-modal

communication that includes several nonverbal aspects. In many societies, written communication has been added as an adjunct to speech. Written communication is usually restricted to a single mode, orthographic text, and there are numerous formal conventions on how it is produced, e.g. spelling, punctuation, rhetorical structures. Unlike speech, which is produced in real time, writing is often composed “offline” where it can be edited and revised to varying degrees before being released. All of today’s communication technologies are simply extensions of these two modes. Typical social interactions today are still face-to-face, but technology has created variations of both the spoken and written communication. How different written modes of communication are from spoken conversation is an important question to consider when determining the core vocabulary of a language.

The evolving structures of our telephone and electronic communication systems have helped us see the network structure of our social contacts. At the same time, it was language research which led the way to the discovery of these network patterns – patterns that go much deeper than the electronic web that now encircles the globe.

In the early 20th century, a French stenographer, Jean-Baptiste Estoup (1916), noticed an unexpected pattern in the frequencies of words. Word frequencies are an important consideration in stenography, since slight improvements in the speed of transcribing high-frequency words can lead to major gains in efficiency. One might expect a gradual decrease in frequency from the most frequent word to the second most frequent, to the third most frequent and so on. Instead, Estoup – and later George Kingsley Zipf (1935) – discovered that in texts and collections of texts, there is an extremely sharp decrease in the frequencies of words. The pattern is such that the second most frequent word should occur roughly half as many times as the first. The third most frequent word should occur only about one-third as often as the first. This inverse proportional relationship between a word’s rank and its frequency is fairly consistently through the full range of word frequencies of any naturally produced text (Sorell, 2012).

Proportional patterns of this sort are known in mathematics as power law distributions. The power law distribution that characterizes word frequencies is usually referred to as Zipf’s law, since it was Zipf who popularized it. Shortly after Zipf’s *Human Behavior and the Principle of Least Effort* was published, a young mathematician, Benoit Mandelbrot, retrieved a review of the book from the wastepaper bin in his uncle’s office so he would have something to read on the subway (Mandelbrot, 1982: 346). He found this distribution of word frequencies so intriguing that it led to his first published academic article, and it set him on a course toward the discovery of fractal geometry.

A fractal is an object in which the small scale structures share the same pattern as those on larger scales. The branching pattern of a tree, for example, is the same if one is looking at the large branches attached to the trunk, or smaller branches further out, or even the twigs at the extremities of the tree. The relationship is the same at every scale.

The proportional frequency distribution described by Estoup and Zipf is related to an underlying hierarchical and fractal pattern (Chen, 2012a, 2012b). Interestingly, the shapes of the antennas in cell phones we use so often are based on Mandelbrot's work. Their fractal shapes allow a single antenna to resonate at a variety of frequencies so that they can transmit and receive a variety of signals, such as Bluetooth, Wi-Fi, 3G and GPS (Werner, Haupt, & Werner, 1999). Language is similarly scalable to fit different communicative contexts (Sorell, to appear). A writer can adapt the wording of a message to fit a technical journal, a popular magazine, or a children's book depending on the audience.

Euclidean geometry is able to describe cubes, cylinders, pyramids, and other regular shapes, but fails when it comes to the organic forms found in nature. Mandelbrot's discovery of fractal geometry provided the mathematical vocabulary to describe the structure of the trees, mountains, clouds, and other dynamically evolving structures like the pattern of connections on the Internet. Even though the physical connections of the Internet and the software network of World Wide Web that it hosts are man-made structures, no one was certain of their architecture. Both had evolved organically without any overall design. Working independently, teams of researchers discovered that both the Internet and the World Wide Web have evolved a very particular type network structure known as a small world configuration (Albert, Jeong & Barabási, 1999; Faloutsos, Faloutsos & Faloutsos, 1999; Kumar, Raghavan, Rajagopalan & Tomkins, 1999). In a small world network "just a few nodes have so many links that 80 to 90 percent of the network's total number of links feed into just a small fraction of the nodes" (Buchanan 2002: 84). A small world network pattern has also been found in the biological processes of living things in everything from the interactions of proteins in yeast (Barabasi & Bonabeau, 2003: 54; Wagner & Fell, 2001) to the connections and firings in the nervous system (Buchanan, 2002: 61-72; Scannell, 1997; Klaus, Shan & Dietmar, 2011; Shew, Yang, Yu, Roy & Plenz, 2011).

Most importantly for this study is that a small world configuration has also been found in the patterning of words in texts (Ferrer-i-Cancho & Solé, 2001a). Ferrer-i-Cancho and Solé's study defined links between words based on proximity. If a word occurred within one or two words of another word in texts, it was considered to have a link to that word. They found a core of 5,000-6,000 highly connected words that seem to act as hubs linking the rest of the words in the lexicon. Imagining words as nodes on a network, one could navigate from

one word to another in the lexicon with “less than three intermediate words on average” (Ferrer-i-Cancho & Solé, 2001b: 4). This conception fits very well with a constructionist or usage-based model of language acquisition in which speakers acquire language by observing frequent word patterns (Barlow & Kemmer, 2000; Bybee, 1985, 1995, 2006, 2010; Bybee & Hopper 2001; Croft, 2001; Fillmore & Kay, 1993; Fillmore, Kay, & O’Connor, 1988; Goldberg, 1995; Langacker, 1982, 1987, 1988, 1990, 1991, 2000, 2008; Taylor, 2002).

Probably the dominant field in applied linguistics is the endeavour to network with speakers of languages one does not yet understand. This is a daunting task since a native speaker of a language typically knows many thousands of words and has a nuanced knowledge of how to combine them in utterances in a wide variety of contexts (Nation, 2001: 8-9). Coming full-circle, Zipf’s law and the network structure of the lexicon itself provide a strategic starting point on the journey towards full participation in the language.

Even though this mathematical pattern in vocabulary was apparently not noticed until the 20th century, language learners and teachers long before intuited that lists of core vocabulary would be an excellent tool for initial language learning. Early successes should motivate learners in a virtuous cycle that can lead to proficient use of the new language. Learning a core of 1,000 word families covers 70-80% of the words in almost any type of text (Nation, 2001: 17). Knowing a core of 2,000 word families plus the 570 families on the Academic Word List (Coxhead, 2000) would mean that the learner would know almost 90% of the running words of an academic text (Nation, 2001: 17-18). Coverage of 90% is impressive, but that still leaves one in 10 words that would be unknown. Reading would be feasible, but very intense.

Hu and Nation (2000: 423) as well as Schmitt, Jiang and Grabe (2011) found that 90-98% coverage is necessary for comprehension of spoken or written texts depending on how well one needs to understand the texts. A total of 6,000 word families would allow a learner “to cope with unsimplified spoken language” (Nation & Newton, 2009: 133; Nation, 2006). Raising the number to roughly 8,000-9,000 word families would allow the learner to read ungraded narrative or general written texts with ease (Nation, 2006). These are substantial, but achievable long-term goals that could be accomplished over several years. That learners see substantial gains early on is crucial, and reliable core vocabulary lists can be a significant help.

Researchers have demonstrated and practitioners often intuitively realize that core vocabulary lists are valuable assets in the acquisition of a new language. Such lists should guide pedagogical decisions, such as syllabus design, testing and the creation of learning materials. However, certain problems undermine teachers’ and learners’ confidence in

vocabulary lists, so they are often not given the recognition or priority they deserve in language instruction and teacher training. During my own training as a language teacher, I do not ever remember a discussion of the value and use of core vocabulary lists. As a foreign language learner, there was only one instance where vocabulary learning objectives were based on frequency. In learning Biblical Hebrew, we received a list of the most common words in the Hebrew Bible. Since it is a closed text, language instructors know the exact frequency of each word. In a living language, a once-and-for-all count is impossible. How much has vocabulary use in English changed in just the last sixty years? How much will it change in the next sixty? It seems highly unlikely that any list produced today would still be in use sixty years from now. What is needed, instead, is an objective methodology for creating core vocabulary lists that can be replicated often and for varying situations.

1.1 Problems with vocabulary lists

Despite their potential usefulness, several factors have hindered the widespread use of core vocabulary lists in foreign language learning, especially those that are objectively compiled. Some are valid criticisms that need to be corrected, while others stem from a lack of understanding of what can be expected of core vocabulary lists or from negative perceptions of how vocabulary lists have been used.

As mentioned above, the present study stems from a review of my own teaching and the frustrating lack of progress seen in students' vocabulary learning. In hindsight I now realize that many of the students could well have quoted Job 19:2, "How long will you torment me and crush me with words?" After an investigation with Nation's Range program (Nation & Heatley, 2002), I discovered that nearly 90% of the words I had originally selected for explicit instruction in one segment of a course were too infrequent to be of any significant future benefit to the students.

Published frequency-based lists are usually far better than individual intuition, but the skepticism they sometimes face is not without reason. Nation (1990: 20-21) identifies five key problems which are presented below in a slightly modified order along with a few other issues.

1.1.1 Words are listed in an inappropriate sequence.

- 1) "The most serious problem with word-frequency lists is that certain useful and important words do not occur in the first or second 1000 words."

Nation gives the examples of *soap*, *bath*, *chalk*, *damage*, and *stomach* that are not in the top 2,000 of Thorndike and Lorge's *The Teacher's Word Book of 30,000 Words* (1944). The words *tidy*, *stupid*, and *behavior* are not in first 3,000. Fries and Traver (1960) also found that common words covering food, clothing and the classroom typically do not appear on frequency lists. In one very early corpus study of spoken French, the names of days were present except for *mercredi* "Wednesday" and *vendredi* "Friday" (Gougenheim, Michéa, Rivenc, & Sauvageot, 1964: 102, 112; Halliday, McIntosh, & Strevens, 1964: 194).

- 2) "[S]ome words that are not suitable for a beginner's vocabulary come in the first 1000 words of most frequency lists."

The business-oriented vocabulary items *bank* (v), *bill*, *company*, *deal*, *issue*, *labor*, *stock*, *supply*, the proper noun *Chicago*, and the archaic pronouns *thee* and *thou* are in the first thousand of Thorndike and Lorge's list.

- 4) "Usually the order of words in a frequency list is not the best order in which to teach the words."

Nation cites *his* which is the 74th word on Thorndike and Lorge's list and its feminine parallel *hers* which is the 4,151st. One useful application of word lists is the ordering of vocabulary instruction. On the other hand, strictly following most lists would lead to a bizarre arrangement in which related and similarly useful items would be introduced at very different times in the curriculum.

The Thorndike and Lorge list that Nation criticizes was compiled before computer-based corpora were available, but these criticisms hold for contemporary lists as well. For example, in compiling a core vocabulary list based on the BNC, Nation (2004: 3) discovered that lists based on frequency and distribution in the whole BNC were not suitable core vocabulary lists since "the BNC is predominately a corpus of British, adult, formal, informative language."

Mark Davies' 5,000-word list, just released in 2012, would still be subject to some of the same criticisms cited for Thorndike and Lorge's list almost 70 years earlier. The COCA list is based on Davies' 450-million-word Corpus of Contemporary American English that is comprised of equal portions of text from 1990-2012. Of the words, Nation argues should be in top 2,000, only one is within that range on the COCA list (A word's rank is given in parentheses): *soap* (4,389), *bath* (3,923), *chalk* (not found), *damage* (1,531), *stomach* (2,525). Of those that should be in top 3,000, only *tidy* is not found, but *tidy* is less common in American English: *tidy* (not found), *stupid* (2,725), and *behavior* (674). The most persistent problem is with words which are in top 1,000, but are not appropriate for most beginners: *bank* v. (not found), *bill* (809), *company* (189), *deal* (675), *issue* (1,875), *labor* (1,158), *stock*

(740), *supply* (1,503). Scanning the first few hundred items on the COCA list, one sees numerous terms from the world of politics and business. Nation's criticism that *his* and *hers* are not found close to each other may not be entirely fair since *his* (25) parallels both the pronominal usage *hers* (3895) and the more frequent determiner *her* (42). The COCA list does not include proper or quasi-proper nouns, such as *Chicago* or the names of the days or months. The list does, however, contain nouns like *American n.* (545), *Catholic n.* (3127) *Russian n.* (4033) and their more frequent adjective counterparts.

1.1.2 Word lists are not consistent.

3) "Word-frequency lists often disagree."

Nation (2001: 15-16) wrote that "if the research is based on a well-designed corpus there is generally about 80% agreement about what particular words should be included." This is still a substantial amount of variation, which leads some learners and instructors to ignore the benefits of using frequency-based lists.

5) "Word frequency lists are not reliable above a certain level."

As one progresses down the frequency list reliability decreases. Going back to Zipf's law, one remembers that word frequencies drop sharply. Since there are fewer and fewer instances of these words, the probability of their occurring becomes more and more uncertain. To some degree, this is inevitable, but this variation has not been quantified.

Unfortunately, when word lists are seen as unreliable, many language learners and instructors abandon them in favor of their own intuitions. Ironically, the most common words are so frequent that they often slip by without being noticed (especially by native speakers) or they do not lend themselves to quick instruction, and so they are skipped in favor of more concrete and content-laden words.

Is too much not enough? This study began with a tentative hypothesis that the quantity of data would reach a point beyond which no significant gains would be seen for a given length list for a particular text type. For example, when looking for the most frequent 1,000 words in narrative writing, would one find that a corpus 20 million tokens produced essentially the same list of 1,000 words that a corpus of 10 million would produce? If so, there would be no cost benefit to compiling the additional 10 million words. Whatever "best case scenario" was discovered could then serve as a minimum corpus size for studies wishing to find a particular of length frequency list.

Previously, the upper limit of a cost-effective corpus was assumed to be a few million tokens at most and would only give reliable results for the most frequent 1,000 to 1,500 words

(Fries & Traver, 1960: 26). Gathering more data was considered “disproportionate to the effort required” (Howatt, 2004: 290).

1.1.3 Word lists are not representative.

It has long been understood that frequency lists simply reflect the types of texts used as source material (Fries & Traver, 1960: 26). This has led to two different reactions. One was to gather a diverse amount of material and hope that it was sufficiently representative. The other was to distrust frequency lists and rely instead on intuition (Howatt, 2004: 288-293). After all, how was one to know what sampling frame would adequately represent all the diverse oral and written texts produced in a language? If the selection of source texts is subjective, how is that different from subjectively choosing the words themselves?

This issue has not gone away with the advent of very large corpora either. The composition of many large corpora is based more on what texts are readily available than on achieving a balanced representation of the language. Over half of the Bank of English, which currently contains 650-million-tokens, is written reportage. Speech makes up 11% of the total corpus and is almost entirely represented by transcripts of radio and TV. British spoken English makes up around two-thirds of the Bank of English spoken data about half of which is composed of transcripts from the BBC World Service. The other half is the Cobuild British Spoken Corpus, the contents of which are not described. About one-third of the total spoken portion of the corpus is American English, which is taken entirely from the Voice of America (Collins WordBanks Online, 2009).

Since the 1990s, corpus linguists have begun compiling multi-billion-token corpora by collecting texts online, such as the above-mentioned COCA, the 1.9 billion-token Corpus of Global Web-Based English (Davies, 2013), the EnTenTen12 corpus of 12 billion words (Jakubíček, Kilgarriff, Kovář, Rychlý, & Suchomel, 2013) and the Big Web Corpus of English which has a target size of 20 billion (Pomikálek, Rychlý, & Kilgarriff, 2009). Responding to criticism that corpora gleaned from the Internet are not representative, Kilgarriff and Grefenstette (2003) argue that “[t]he Web is not representative of anything else. But neither are other corpora, in any well-understood sense” (2003: 342). In fact, “[o]utside very narrow, specialized domains, we do not know with any precision what existing corpora might be representative of” (Kilgarriff & Grefenstette, 2003: 340).

Small talk is not small. Section 3.1 will describe how Biber (1995) provides an objective framework for dividing the major text types produced in English. Two of these text types, “intimate interpersonal interaction” and “informational interaction” are approximately

equivalent to face-to-face, especially casual, conversation. An important hypothesis of this study is that such conversation is the primary source of what is intuitively thought of as core vocabulary.

Some of the problems with core vocabulary lists stem from the choice of exclusively written texts as source material – something that would have been hard to avoid before the late 20th century. Fries and Traver (1960: 40-41) point out this as a flaw in all of the lists in their survey.

Peyawary (1999) attempted to find the core vocabulary of EIL using the Brown, LOB and Kolhapur corpora. With one million words from American, British and Indian English, he expected to find the core vocabulary occurring frequently in all three varieties. Unfortunately, 1 million words is still too few, and the corpora that were used contain no spoken language. Peyawary's resulting list includes *latter*, but not *laugh*; *finance*, but not *fish*; *military*, but not *milk*. *Nevertheless* and *nuclear* are considered core, but not *nose*!

Even today, including comparably-sized samples of spoken language is a challenge. The fact that almost all frequency lists have been based exclusively or predominately on written texts is likely why the need has always been felt to subjectively fix lists, even when the goal was to produce a supposedly objective standard. Even though the Oxford 3000 list (See section 1.1.6) is based on a corpus of several hundred million words, the creators had to add “some very important words which happen not to be used frequently, even though they are very familiar to most users of English” (Turnbull, 2010: R43).

1.1.4 Which variety of English?

West's *General Service List of English Words* (1953) was a product of the Carnegie Conferences that were collaborations between British and North American scholars. They represent the two dominant varieties in what Kachru (1985: 12) termed the “inner circle” of English. However, even agreeing on a core list compatible with both of these varieties is often difficult, as Kilgarriff discovered when creating frequency bands for the *Longman Dictionary of Contemporary English*. After using objective, frequency-based methods, Kilgarriff manually added about a hundred American words on the advice of an American lexicographer (1997: 147).

English as an International Language (EIL). In a world where “the majority of English speakers are now multilingual people who have learned English and who communicate with fellow multilinguals” there is reason to wonder if core lists based exclusively on Anglo-American norms are still the most desirable (Kirkpatrick, 2007: 1).

Larry E. Smith (1976) coined the term English as an International Language (EIL) to highlight the role of English as “an international language...used by people of different nations to communicate with one another.” The term EIL shifts the focus toward the role of English in many different multilingual contexts, not just contexts where it is the dominant public and private language. The concept of EIL is relevant to this study since the goal of this study should be to produce core vocabulary lists that are drawn from and applicable to English in a variety of global contexts.

The original proposal for this study envisioned a set of parallel collections covering several varieties of English. Beyond the International Corpus of English (Greenbaum, 1996), very little data is available from the outer or expanding circles of English. Nonetheless, the study will demonstrate how to find the common ground shared by different varieties of English.

1.1.5 What is a word?

One additional inconsistency which causes differences between lists is what unit is counted as a word. In Thorndike’s first *Teacher’s Wordbook* (1921), word types are counted individually, so *walk*, *walks* and *walking* are counted as separate words. All of the inflected forms (types) of a word make one lemma, so the above words would all be members of the lemma *walk*. The lemma is used as the unit of counting in some of the word counts that are used for the subsequent *Teacher’s Wordbook of 20,000 Words* (1932). Since this was before the era of computer processing, these lemma counts were simply spliced together with the type counts ignoring the fact that they were based on a different definition of what a word is.

In their study, Carroll, Davies and Richman (1971) count capitalized and lowercase word forms as separate words. This is useful in distinguishing the lexical noun *bill* from the proper noun *Bill*, but capitalization is usually “not used as a distinguishing feature unless there is a very strong reason to do so” (Nation, to appear).

1.1.6 How big is the core?

A final problem is the actual length of a core vocabulary list. Lists range in size from the 850 words of Ogden’s BASIC to the 5,000 lemmas of Davies’ downloadable COCA list (Table 1.1). Even taking into account the differences in what is counted as a word, this is still a considerable range. The four shorter lists in the table were designed to be a restricted vocabulary that would guide the composition of vocabulary-controlled materials, such as

dictionary definitions, simplified news broadcasts or graded readers. They needed to be small enough to be a reasonably achievable goal for beginning-level learners.

Table 1.1

A selection of core vocabulary lists.

List	Length
Ogden's BASIC (1930)	850 words
Voice of America's Special English Word List (1959, 2009)	1,510 words
General Service List (West, 1953; Nation & Heatley, 2002)	2,000 word families
Longman Defining Vocabulary (Proctor, 1987)	2,197 words, 10 prefixes, 39 suffixes
Longman Communication 3000 (n.d.)	3,000 words
Oxford 3000 (Turnbull, 2010)	3,000 words (includes some two-word compounds, such as <i>trash can</i>)
Hindmarsh's Cambridge English Lexicon (1980)	5,000 words
COCA (Davies & Gardner, 2010)	5,000 lemmas

Each of the lists in Table 1.1 was designed to be adequate for general communication and, except for Ogden's BASIC, to be a stepping stone to more advanced language learning. Hindmarsh's list was specifically designed for language testing. Davies' list may not have been designed for English language teaching, but that use was certainly foreseen.

Based on Nation & Newton (2009: 133) and Nation (2006), the two longest lists are only approaching the length necessary for comprehending ungraded speech and are still far short of the vocabulary needed for ungraded reading. Would 6,000-9,000 words exceed the limits of a core vocabulary? Ferrer-i-Cancho and Solé's (2001a) study implies the core does contain around 6,000 words.

Creators of core vocabulary lists are caught between two contradictory goals: ease of learning and adequate coverage. Learners do better with learning objectives that do not intimidate them into inaction. At the same time, learners often aspire to become proficient users of the language in which case they will need many more words. A list which is short enough for rapid learning will certainly lack many very useful words, but when one includes those words, the list quickly expands.

1.1.7 Perceptions of word lists

One key problem hindering the more widespread use of core vocabulary lists is not intrinsic to word lists, but in how they are perceived. In the minds of both learners and teachers word lists are often associated with currently unpopular teaching methodologies, such as grammar-translation. The grammar-translation method was developed by educational reformers interested in making language learning less difficult for young learners. The grammar-translation method did improve the lot of many pupils learning Latin in the late 18th and 19th centuries in Germany, where it originated. However, modern perceptions of the method are based on later developments in the UK and elsewhere. The Cambridge and Oxford Local Examinations were implemented in 1858 to maintain academic standards across the wide variety of British secondary schools. The exams tested classical and modern languages, but had an unfortunate wash back effect on language teaching that resulted in an obsessive focus on accuracy, completeness and fine distinctions between often archaic vocabulary items (Howatt, 2004: 150-165). Modern language instruction was also expected to emulate the teaching of the classical languages, and so the spoken language was undervalued in favor of written literature. This sort of language teaching did not die out in the 19th century (Coady & Huckin, 1997:6). For many language instructors today, the use of word lists brings back unpleasant memories of teaching methodologies they endured as students. This is regrettable since words lists have much to offer.

1.1.8 Summary of problems with core vocabulary lists

When a core vocabulary list lacks words learners or instructors feel are essential, and instead includes archaic or other irrelevant words, this undermines confidence in the whole list and may even discourage the use of core vocabulary lists. In other words, core vocabulary lists are not valued if there is a discrepancy between the words people feel the core list should contain and the words that are actually found to be high-frequency in the study that produced them. This goes back to the reliability of the list which rests largely on what texts are in the corpus that produced the list and the quantity of those texts that were available to the creator of the list. If the corpus is not representative, the word list will not be either. If the sample is too small, the list will be unreliable.

Word lists also suffer from negative associations when learners and instructors do not separate useful techniques from methodologies which may be unpopular. The reality is that

language learning requires a considerable amount of time and effort even with effective teaching techniques.

1.2 Purpose and Significance

Since, according to Nation (2006) and Nation and Newton (2009: 133), learners need 6,000-9,000 words to comprehend spoken and written texts, Swan and Walter (1984 [cited in Thornbury, 2002: 14]) are not exaggerating when they write that “vocabulary acquisition is the largest and most important task facing the language learner.” A principled guide to the presentation and study of those words would be a very valuable asset to students, instructors, textbook writers and test creators.

Wilkins’ aphorism, “Without grammar very little can be conveyed, without vocabulary nothing can be conveyed” (1972: 111), takes on new meaning when considers how the network of words may be at the root of syntax. The advocates of a usage-based or connectionist grammar mentioned earlier in this chapter would argue that grammar and vocabulary learning are at least intertwined, if not the same endeavor. Researchers that view language as a complex adaptive system would agree (Larsen-Freeman & Cameron, 2008: 129-130; Schmitt, 2000: 14; van Geert, 2009: 63). Even one of the pioneers of EFL teaching, Harold Palmer, hinted at this long ago in the title of his (1938) *Grammar of English Words*. Rather than simply emphasizing the importance of vocabulary in language learning, this view of language implies that vocabulary learning *is* language learning (with a wavy equal sign $\approx$). Therefore, helping language learners and teachers improve vocabulary learning should be a very high priority. This study proposes to do that by addressing each of the problems outlined in Section 1.1 and offering practical solutions to each.

The goal of this study is to outline methodological principles for constructing core vocabulary lists for English as an International Language in an objective, replicable manner. The impetus for this study arose from a tertiary-level, EFL teaching context in the “expanding circle” of English where most of the students are learners of English as a foreign language, but such lists would be useful to learners, teachers, writers and designers of course materials in a variety of contexts (Nation, 1990:18ff). Many of the findings should also be applicable to other situations where objective measures of vocabulary use would be of value.

1.3 Research questions and hypotheses

This thesis seeks to answer the overarching question, “How can valid core vocabulary lists for English as an International Language be created?” by addressing the following research questions in three linked studies:

1. How much data is necessary to give consistent results for high-frequency vocabulary?
That is, how large a corpus is necessary to achieve nearly identical results for a given length of frequency list? For example, does one need a corpus of 20 million tokens before one ceases to see variation in the most frequent 1,000 words for a certain type of text?
2. How much difference is there between the core of oral and written language? Should there be distinct oral and written core lists, or should there be individual lists for each type of text? More precisely, how different are the high-frequency words of each of the text types?
3. How does the scope of the definition of a word as type or family influence the resulting list?

1.4 Definition of terms

1.4.1 Core vocabulary

A core vocabulary is “those lexical items which are accepted as being central and indispensable to language use” (Bell, 2012). However, one of the main problems in finding a core vocabulary is the difficulty in arriving at a common definition of what is central and indispensable. This thesis will broadly define a core vocabulary as a set of words that will enable learners to adequately function in the contexts where they use English. This is a very open-ended definition of core vocabulary due to the multiplicity of contexts where English is used, especially as an international language. Various definitions that have been proposed for core vocabulary will be discussed in Section 2.4.

1.4.2 Corpus

A **corpus** (plural: **corpora**) refers to “a large and principled collection of natural texts” (Biber, Conrad, & Reppen, 1998: 12). Today it is assumed that these are “machine readable

texts,” i.e. they are stored in an electronic form (Crystal, 2003: 95). Most of the data for the studies in this thesis are drawn from general corpora.

1.4.3 General and comparable corpora

A general corpus is a collection of “texts of many types” and “may include written or spoken language” from “as wide a spread of texts as possible” (Hunston, 2002: 15). The corpora used in these studies were collected according to a sampling frame, but for these studies they have been disassembled and certain types of texts have been reassembled according to a framework that will be described in the next chapter.

Many of these corpora, such as the Brown family of corpora or those that are part of the International Corpus of English are comparable corpora. They are designed to be parallel so that (in this case) different varieties of English can be compared.

1.4.4 Word types and tokens

A word type in this thesis refers to a particular surface form surrounded by punctuation or white space in a text. Therefore a contraction, such as *they’re* is counted as two types *they* plus *re*. This orthographic definition of a word is a practical solution related to how the data will be processed. The text file is the most basic format used for computer processing. It serves as a common denominator for the variety of formats encountered in collecting texts for this study. As a requirement of the methodology for one study, the data collections also had to be divided into equal length sections of precisely 10,000 words (tokens) ignoring utterance, sentence and text boundaries. Recognizing compounds written as two words, like *trash can* mentioned above, was not possible, and compounds written without spaces, e.g. *nonetheless*, are counted as a single word. Homographs, are also treated as a single word since they are by definition orthographically indistinguishable. This includes word forms that can be more than one part of speech, such as the noun and the verb *bank*. In these studies, case is also ignored, so one could say the type *a* occurs three times in the first sentence of this paragraph. Each occurrence of this word is called a token.

1.4.5 Hapax legomenon and dislegomenon

A type which occurs only once, i.e. it represents only one token in a text or corpus, is called a hapax legomenon (pl. hapax legomena) “said once,” a term coined by ancient Greek

grammarians (Greenspahn, 1986: 217). A type that occurs only two times is sometimes called a dislegomenon “said twice.”

1.4.6 Lemmas and word families

A lemma includes all of the inflectional forms of a word. For example, the types *meet*, *meets*, *met* and *meeting* (*v*) are members of the lemma MEET (Hunston, 2002: 17-18). Since only inflectional forms are included, all the members of a lemma are the same part of speech. The noun *meeting* would, therefore, part of a different lemma. Knowles and Don (2004) argue that this English-based definition of lemma is too narrow.

Nagy, Anderson, Schommer, Scott and Stallman (1989) found that words were organized in families in speakers’ mental lexicons. In addition to inflectional forms, a word family includes frequent, productive and regular derivational word forms that are semantically related, so *meeting* (*n*) would be included in the word family MEET (Bauer and Nation: 1993). The operational definition of a word family in this thesis will be the word family lists designed for the BNC/COCA or GSL/AWL described in chapter 8 of Nation and Webb (2011). Interestingly, the idea of a word “family” (משפחה *mishpacha*) was already being used by Ibn Ezra, a Jewish grammarian born in the eleventh century CE (Greenspahn, 1986: 221, note 17).

1.4.7 Marginal words

The practical definition of a word that is being adopted in this thesis – that a word is an orthographic word (sections 1.4.1 and 2.2) – is to some degree in conflict with the premise that speech is the primary form of language. Written language has long dominated the study of syntax and the categorization of words. Structuralist labels of *noun*, *verb*, *adjective*, *adverb*, *conjunction*, *preposition*, *determiner*, etc. have a long history that dates back to the study of Latin and Greek. The study of syntax was predominately the study of complete sentences, often in isolation from discourse. The role of words in the messy, interactive world of conversation was ignored. It is only relatively recently that linguists have begun giving the structure of speech the same level of analysis (Labov, 1972a, 1972b; Labov & Waletzky, 1967; Maas & Wunderlich, 1972; Sacks, Schegloff, & Jefferson, 1974; Sadock, 1974). Terms like *vocalized pause*, *back channeling*, *discourse marker* and *insert* have been coined, but are not universally understood even among linguists. Terms like *noun*, *verb*, *adjective*, etc., on the other hand, are familiar to school children around the world.

Some recent ESL grammars (Biber, Conrad, & Leech, 2002; Carter & McCarthy, 2006) have begun to devote key chapters to spoken conversation and the functions of these often amorphous words, since they facilitate the flow of interaction and understanding in face-to-face conversation. Many of these forms are problematic, though, for corpus analysis because they sometimes lack standardized spellings or are indistinguishable in written transcription when they are separated from facial expression, intonation, voice quality, as well as their physical, epistemic and social contexts (Biber, Conrad, & Leech, 2002: 430; Carter & McCarthy, 2006: 164). What is transcribed as *hmm* could be a comprehension check, but could just as easily signal agreement, surprise, or disappointment. Should all of these meanings be considered a single word? Should *hm*, *hmm* and *hmmm* be considered members of the same lemma or word family?

Including such forms in the finished version of a core vocabulary list is probably unhelpful. The inclusion of taboo words would also be inappropriate in most learning contexts. Nation and Webb (2011) created a set of word lists from the BNC and COCA, one of which is a list of what they termed *marginal words*. Their list of marginal words includes interjections, taboo words and the letters of the alphabet that are not words in their own right. In the final study of this thesis that compares word types and word families, marginal words will be filtered out prior to the comparison. The operational definition of *marginal word* will be the word forms included on Nation and Webb's (2011) list, except for the individual letters which have not been removed.

1.4.8 Text type and genre

In this thesis, text type will always refer to the text types identified by Biber (1995) based on his multidimensional analysis of spoken and written, English-language texts. Genre, such as the novel or lecture, on the other hand, are culturally defined. The text types may include multiple genre or may cut across genre boundaries. Which texts are used to represent each text type will be discussed in chapter three.

1.4.9 Word frequency, normalization, rank and range

A word's frequency refers to how often a word is used in a source or group of sources. It is often useful to report a normalized frequency, for example, how many times a word occurs per million tokens. This makes the frequencies from different-sized texts or corpora easier to compare (McEnery, Xiao, & Tono, 2006: 52-3).

A word's rank refers to its order in a list arranged in descending order of frequency. The ranking function on a spreadsheet, such as MS-Excel, assigns the same rank to all words with the same frequency. The rank that is assigned is equal to the rank of the first word in that group. In other words, if 100 words have already been listed (i.e. 100 words have a higher frequency), the next word will be given the rank 101, as will all other words that have that same frequency.

A word's range refers to the number of individual texts or samples in which a word occurs in a particular study. This term was apparently coined by Thorndike in his pioneering word frequency study (Fries & Traver, 1960: 21).

1.4.10 Coverage

Coverage has two different senses. One is semantic coverage in the sense that *walk* can cover hyponyms like *pace* "to walk back and forth", *shuffle* "walk slowly dragging one's feet" and *stride* "walk with long steps." The second sense refers to how many times a word or the words on a wordlist are found in a text or collection of texts. For example, the word *walk* covers five of the 75 words in this paragraph.

1.5 Organization of the Studies

Chapter two begins by examining definitions of what is considered a word and several criteria for determining which words are part of a core vocabulary. Defining a core vocabulary is dependent on having a framework for analyzing the texts produced in English. A taxonomy of English language texts is proposed based on Biber's (1995) text type categories. The bulk of the chapter is a discussion of how the conceptions and uses of word lists have evolved in foreign language education through history. Their methodologies are evaluated and related to the current studies. Corpora are also reviewed since they are foundational to any objective method for creating a core vocabulary list. Some measures of corpus similarity are then evaluated for possible use in the methodologies that will be described in chapter 3. The chapter then concludes with a summary of some important applications for core vocabulary lists.

Chapter three outlines the methodology of three studies. The Dice coefficient is explained and how it is used in the first study to quantify the difference between samples of the same type of text in order to determine the correlation between sample size and variance among the high-frequency words of each text type. In the second study the Dice coefficient is

used to discover how different the text types are by comparing samples of different text types. A final study compares the coverage of lists of the most frequent word types and word families in different text types. A usage coefficient is also introduced since it is needed to adjust for poorly dispersed words. Chapter three concludes with some notes on scope and limitations of these studies.

Chapter four shares the results of the three studies and some initial analysis. The results are compared with several word lists and lexical sets to try to determine where the core vocabulary is found. This chapter also outlines practical steps for constructing core vocabulary lists, and analyzes a preliminary trial list.

Chapter five discusses the findings of the studies and their implications for the English teaching profession and further research. The chapter concludes by exploring possible applications of the findings of these studies.

CHAPTER 2

BACKGROUND

2.1 Introduction to the background review

Before discussing a vocabulary one must first come to terms with what constitutes a word. This chapter will, therefore, begin with a review of the issues involved in delimiting what turns out to be a fairly fuzzy concept (2.2). The term *core*, in *core vocabulary*, implies that it is the center of some structure. To describe a collection of words as the core vocabulary of a language, one must first have some notion as to the structure of the language. Section 2.3 will review some of the difficulties involved in categorizing the texts of a language and describe a set of text types suggested by Biber's (1995) research. Some proposed definitions of core vocabulary are then discussed in 2.4. Section 2.5 will then review the history of word lists and the methodologies used to create them, especially corpora upon which modern objective methodologies are based. To date, corpus linguistics lacks a recognized measure of corpus similarity, a tool which is essential in creating a taxonomy of text types. Section 2.6 will describe some suggested measures including one attempt by the author to fill this methodological gap. Section 2.7 reviews some of the more significant applications of core vocabulary lists.

2.2 What is a word?

This seemingly obvious concept turns out to be very slippery. Julien (2006) surmises that our intuitions about what counts as a word are probably based in criteria described by structuralist linguists early in the last century. Bloomfield defined a word as a "minimum free form" (1933: 178). By this definition, a word is the smallest independent form that can normally be used as an utterance. Bound morphemes, such as the prefix *un-* or the plural suffix *-s*, are not commonly used as utterances on their own and would, therefore, not be considered words.

Boas (1911/1969) had already "suggested that (relative) freedom of position, or independent distribution, is a necessary part of the definition of the word" (Julien, 2006). For example, the sentences "*Yesterday we had a meeting at three.*" and "*We had a meeting at three yesterday.*" are both grammatical, despite moving the word *yesterday*. *Yesterday*, therefore, qualifies as a word. Moving the morpheme *-ing* in *meeting*, however, would

produce an ungrammatical utterance no matter where else it is placed, e.g. **“We had a ing meet at three yesterday.”*

With this definition, words are composed of morphemes, while words themselves are combined to form utterances or phrases. A problem occurs, however, with expressions like *of course* or *by and large*. These expressions are thought of as multiple words, since each of its members occurs independently in different contexts. However, moving one part of the expression would produce an ungrammatical utterance, e.g. **“Course of, I do know what a word is.”* Such lexicalized phrases (Moon, 1998; Wray, 2002) must be moved as a unit, *“I do, of course, know what a word is.”* The term *lexeme* or *lexical item* covers both multiword units and what we traditionally think of as single words (Schmitt, 2000: 2). Lexemes are very problematic, though, when it comes to creating a methodology for identifying and counting them in large corpora.

Further problems arise when one considers different modes of communication. In speech, various phonological or segmental features, such as stress, help to mark word boundaries. The definition of a phonological word runs into trouble with the category of words called *clitics*. A clitic might be a recognized member of a lemma, which is by definition a free form, but at the same time the clitic may not be phonologically independent. For example, *she’s* [ʃɪz] and *they’ll* [ðeɪl] can each be said to contain two words. The *’s* in *she’s* is an inflectional form of the BE verb, and *’ll* in *they’ll* is a member of the lemma WILL. However, the [z] and [l] are both phonologically bound to the pronoun they follow. Through frequent use these forms have fused into a single phonological unit. Over time, this process, called grammaticalization (Meillet, 1912), produces new word forms and affixes. The indefinite articles, *a* and *an* are from *one*. The morpheme *-ly* in *only* and *really* is from a now extinct word for “body.” The three-letter word *not* was once the Old English phrase *ná wiht* “not a thing” (Aitchison, 2000: 116). It can be seen today in new oral word forms *gonna*, *gotta* and *hafta*.

The second mode of communication, writing, introduces even more idiosyncrasies in how a word is delimited. The orthographic word is based in the writing system that has been developed, in this case, for English. Despite the fact that this thesis stresses the primacy of speech over writing, this study will be restricted to the orthographic definition of a word. For practical reasons, a word is a string of characters bounded on each side by white space or punctuation. This results in clitics, like those mentioned above, being counted as full words, while multi-word lexical phrases are treated as individual items, except for cases like *inasmuch*, *heretofore* or *nonetheless* where they happen to be written as a single word. Compounds count as words a single word when they are written without an intervening space,

e.g. *farmhouse*, but as two words when there is a space or hyphen between them, as in *tennis shoe* and *action-packed*. This is admittedly inconsistent, but nevertheless unavoidable. This definition also, regrettably, ignores gesture as an element or mode of communication.

2.3 A Taxonomy of English Language Texts

2.3.1 Representativeness. The English language is used in a myriad of oral and written contexts every day. How many of these genres or text categories does one need to include in a corpus for it to be representatively general? For published written genres, an objective sampling frame can be based on listings of published texts, such as *The British National Bibliography Cumulated Subject Index, 1960-1964* and the *Willing's Press Guide (1961)* that were used for the London-Oslo-Bergen Corpus (Johansson, Leech & Goodluck, 1978). These categorized lists served as an objective operational definition of the available genre for published written British English for that time period. No such catalogues exist, however, for speech or unpublished writing. Biber (1993) suggests using the setting of the interaction as a basis for categorizing speech. He divides the possible settings into three domains: “institutions (offices, factories, businesses, schools, churches, hospitals, etc.), private settings (homes), and other public settings (shopping areas, recreation centres, etc.)” (Biber, 1993: 246).

However, even assuming one knew what categories of texts to include, how does one balance the quantity of those texts in the corpus? Does one count the production or reception of language or both? Newspaper articles may be written by one individual, edited by others and then read by hundreds of thousands. Most casual conversations are cooperatively composed and shared by only a few individuals, but this type of language use is by far the most common. Hoffmann, Evert, Smith, Lee, and Berglund Prytz point out that “[i]f language production is the decisive factor in the choice of texts to be included, a representative corpus would consist mostly of spontaneous spoken interaction” (2008: 16). “A corpus with this design might contain roughly 90% conversation and 3% letters and notes, with the remaining 7% divided among registers such as press reportage, popular magazines, academic prose, fiction, lectures, news broadcasts, and unpublished writing” (Biber, 1993: 247).

Kilgariff and Grefenstette (2003: 341) also raise the issues of repetition and attention. If one rereads a chapter, does each reading constitute an instance of language? If a song is playing in the background, is that the same as singing the song oneself? For the song played in background, the text constructed in the mind of the listener may be minimal. In any communicative situation there are likely to be significant differences between the text that the

speaker or writer produces and the one the listener or reader constructs mentally (Rosenblatt, 1978). A corpus could not possibly account for these differences.

Biber (1993: 245) sees three possible design strategies for constructing a corpus. The corpus could attempt to represent text production, text reception or the texts as products. The first two strategies would result in collections which omitted or included a vanishingly small quantity of many types of texts. Most people simply do not frequently read academic journal articles or legal briefs, for instance. These types of corpora would not be useful for most types of linguistic research. For the final design strategy, the corpus tries to capture the range of texts produced within certain temporal, social or geographic parameters without attempting to represent the quantity of those texts that were produced or received. The balance is instead based on a subjective judgement of the relative importance of each type of text in the society. This is the basic strategy compilers of general corpora are forced to adopt. An attempt is made to represent a wide variety of texts and contexts, but “there are simply no absolute or objective criteria that can be used to gauge representativeness” (Hoffmann et al., 2008: 17).

Genres and speaking situations are extrinsic constructs. There are linguistic differences between them, but they are not linguistically defined. Methods have been suggested for measuring the intrinsic differences between texts. Manning and Schütze (1999: 575-608) discuss methods for categorizing texts, as in how one would automatically sort a stream of news reports into topic categories. However, these methods require that one first determine the categories into which the texts will be sorted before sorting begins. The problem corpus designers face is that there is no comprehensive set of categories that are based on objective linguistic criteria. Genre classifications of texts and speaking situations are based in cultural conventions which may or may not correspond to a linguistic taxonomy.

2.3.2 Multidimensional analysis. Rather than using a predetermined set of categories, Biber (1988) designed a tagger that marked a wide variety of texts for every linguistic feature he could find reference to in linguistics research. He then looked for strong positive or negative correlations among the texts. He discovered five significant dimensions along which texts vary depending on their purpose and context. Each of these dimensions is made up of certain features which are either positively or negatively associated with one end of the dimension.

1. *Involved versus Informational*: Involved texts are marked by a higher frequency of second-person reference and private verbs, like *think* and *feel*. Informational texts, on the other hand, contain more frequent nouns and a more precise and varied vocabulary.

2. *Narrative*: Narrative texts naturally have a large number of past tense forms, perfect aspect verbs, and third-person personal pronouns, whereas non-narrative texts lack these features.

3. *Situated versus Elaborated*: Texts which are highly situated are characterized by frequent place and time adverbs. Elaborated texts, on the contrary, do not assume a shared context and are, instead, intended to be self-explanatory. Elaborated texts are noticeable by their use of wh-relative clauses, pied-piping constructions (*for which...*), and nominalizations.

4. *Persuasive*: Overtly persuasive texts contain a greater than average number of “prediction modals (*would, shall*), necessity modals (*should, must*), possibility modals (*might*), suasive verbs (*agree*) and conditional subordination” (Biber, 1995: 343).

5. *Abstractness*: Abstract texts are primarily marked by a high frequency of passive constructions.

This multi-dimensional analysis of texts is a tremendous insight into how a diverse set of lexical and syntactic features are related to the purpose of a text and the context in which, or for which, it was composed. This study shows the shape of the discourse space, but still does not provide a clear scheme for dividing the wide variety of texts in a language. Biber’s tagger program could analyze a particular text and assign it a score for each of the dimensions. These scores would give a text’s position within this five-dimensional space. However, that is somewhat like having the GPS coordinates of your car’s current position without having the map overlay to guide you as you drive. The map came later when Biber (1989, 1995) examined how particular scores on each of these five dimensions correlated to form several constellations of texts. For English, Biber found eight relatively distinct text types:

1. Intimate interpersonal interaction
2. Informational interaction
3. “Scientific” exposition
4. “Learned” exposition
5. Imaginative narrative
6. General reported exposition
7. Situated reportage
8. Involved persuasion

These “text types are identified on the basis of shared linguistic co-occurrence patterns, so that the texts within each type are maximally similar in their linguistic characteristics, while the different types are maximally distinct from one another” (Biber, 1993: 245).

Similar multidimensional analyses were done for Somali, Korean and Nukulaelae Tuvuluan, Taiwanese and Spanish (Biber, 1995; Jang, 1998). These languages were selected since each presented a contrasting case study in the role of literacy in the culture. Each study was independent and used linguistic features specific to that language. A distinct set of text types was discovered in each case. Surprisingly, though, “the analyses of all languages have uncovered dimensions relating to interactiveness/involvement versus informational focus, the expression of personal stance, and narrative versus non-narrative discourse” (Biber, 1995: chapter 7; 2004: 16). Folch, Heiden, Habert, Fleury, Illouz, Lafon, Nioche and Prévost (2000) described a methodology for this type of analysis for French, but did not suggest a taxonomy of text types.

Figures 2.1-2 show the typical scores for each text type on each of the five dimensions. The dimensional profiles of text types 1 and 2 are almost identical except that texts of type 1 have a somewhat higher score on dimension one. Text type 1 consists of interactions that are “situated in reference, non-abstract in style, and extremely interpersonal and involved in focus” (Biber, 1995: 333). Texts of type 2 are still focused on maintaining social relationships, but are slightly more informational in focus (Biber, 1995: 333-334). All of the texts in text type 1 are face-to-face conversations or telephone calls (Biber, 1995: 328). No written texts are found in the first text type. The vast majority of texts in type 2 are still face-to-face conversations, though some personal letters, one piece of business correspondence, and one text of general fiction fall into text type 2 as well.

Biber found a significant difference between the academic writing of scholars in the natural sciences and those in the humanities. A far more abstract style is preferred in the natural sciences, whereas scholars in the humanities apparently do not usually feel the need to distance themselves quite so markedly from their research (Biber, 1995: 334).

Figure 2.1

Biber's (1995) text types 1-4 showing the average score on each of five dimensions.

Texts of type 5 *imaginative narrative* have slightly higher scores on all dimensions with the highest – not surprisingly – being the narrative dimension. More than three-quarters of the texts that fit into this category are from some genre of fiction (Biber, 1995: 330). Seven of the 14 prepared speeches in Biber's study were also members of this text type.

The most common text type in Biber's study – the category into which most texts fell – is text type 6 *general reported exposition*. Texts of type 6 have nearly neutral scores on all five dimensions (Biber, 1995: 339). They include a wide range of written texts, as well as a several prepared spoken texts. The largest percentage are from press reportage and press editorials followed by academic prose, general fiction, religion, humor, biographies, press reviews, hobbies, non-sport broadcasts, prepared speeches, adventure fiction, science fiction, mystery fiction, popular lore, professional letters, sports broadcasts, official documents and one sample of romance fiction.

Type 7 *situated on-line reportage* is “the smallest and most distinct text type identified in [Biber's] study” (Biber, 1995: 340). It is primarily the highly situated speech of live sports commentary (Biber, 1995: 331).

According to Biber, text type 8 *involved persuasion* is also “a relatively specialized text type” (Biber, 1995: 341). This text type is spread across 14 different spoken and written genres, but is not the center of gravity for any of them (Biber, 1995: 331). One could think of

it as a discourse space that a speaker or writer can resort to when needed, whether it be in a phone call to a colleague or in an academic journal article.

Figure 2.2

Biber's (1995) text types 5-8 showing the average score on each of five dimensions.

Biber's text types provide a linguistically-based taxonomy of texts that represent groupings of texts that are intrinsically different and not simply defined by cultural conventions. Nonetheless, there are some strong correlations between some genres and some text types.

2.4 Defining core vocabulary

Bell's (2012) definition of core vocabulary as "those lexical items which are accepted as being central and indispensable to language use" is deliberately open-ended. Several authors have suggested more explicit criteria for determining which words they might consider central and indispensable. Stubbs (1986) and Carter (1982, 1987b, 2012: chapter 2) list several traits they believe are characteristic of core words. Lee (2001), however, lists seven – not necessarily compatible – conceptions of what a core vocabulary might be. In his study, he opts for a very practical definition. This section will review the array of definitions given in Lee (2001) followed by the characteristics suggested by Stubbs (1986) and Carter

(1987b, 2012: chapter 2). (The order in which Lee’s definitions are discussed below has been altered.)

1. “words useful for dictionary definitions”: One could object that since a vocabulary of this type is designed for a very specific function, it cannot be accepted *a priori* as a core vocabulary for general communication. However, defining vocabularies are often used as de facto core vocabulary lists. They do at least provide access to the definitions used in that particular learner’s dictionary.

2. “the most frequent words in the language as a whole”: One would assume that core vocabulary items would be the most frequent, but as was seen in Section 2.3, designing a corpus that is a representative sample is no simple task. Sutarsyah, Nation and Kennedy (1994: 41) found marked differences in a comparison of just the fifty most frequent words in an economics text and a corpus of general academic English. The content words *price, cost, demand, curve, firm, supply, quantity, margin, economy, produce, income, market, labour, increase, consume, total, change* and *rate* were among the fifty most frequent words in the economics text, while the fifty most frequent items in the general academic corpus contained only three content words: *use, make* and *year*. This is a small, but dramatic example of the effect corpus sampling can have on a frequency count.

3. “the most frequent in terms of a particular medium”: Lee’s (2001) study sought to discover if there is one core vocabulary, or separate cores for spoken and written English. He sought to answer that question by checking the distribution of the Longman Defining Vocabulary across the different genres that compose the BNC. For the purposes of his study, Lee assumes that the LDV closely approximates a core vocabulary and so should be distributed in a similar manner across each genre. The table below shows the percentage of the LDV that occurred in each genre subsection.

Table 2.1

Coverage of the Longman Defining Vocabulary in BNC genre subsections (adapted from Lee, 2001: 267).

Spoken	
Demographic B	92.76
Demographic C2	92.69
Demographic C1	92.53
Demographic D	91.94
Demographic E	91.69
Demographic unclassified	91.56
Demographic A	91.04
Context-governed educational	90.34
Context-governed leisure	88.86
Context-governed public	86.65
Context-governed business	85.60
Average across spoken domains	90.52
Written	
Fiction	83.70
Poetry	81.36
Drama	80.89
Community and social science	77.67
Arts	75.38
Belief and thought	73.35
Commerce and finance	72.61
Leisure	71.72
World affairs	71.10
Applied science	70.53
Pure science	65.62
Average across written domains	74.90

Lee's study shows a smooth gradient from middle class speakers in the B section of the spoken demographic collection (92.76%) to scientific writing (65.62%). This would seem to indicate spoken and written genres form a continuum rather than two distinct categories. In the border region, the context-governed speech of business meetings closely approximates the coverage found in some forms of written language, and fictional writing naturally includes a

large quantity of simulated dialogue. Oral and written language, nonetheless, represent two discernable sides of the continuum of language usage. Worth noting is the strong association between this definition of core vocabulary and face-to-face conversation as represented by the BNC demographic sections. One could argue that face-to-face conversation is the core of the language (c.f. Hoffmann et al., 2008: 16; Biber 1993: 247).

4. “the most frequent words for a particular demographic grouping”: One could restrict sampling based on some combination of the criteria “age, socioeconomic status, educational level, region, and so on” (Lee, 2001: 254). Lee, however, argues that too narrow a focus contradicts the idea of a core vocabulary. “[S]lang and markedly colloquial items are generally not considered part of any ‘core’ vocabulary” even though they may be more frequently used by some demographic groups (Lee, 2001: 254). The same argument could be made for not including overly formal or specialized terms. West’s *General Service List* (see Section 2.5) is sometimes criticized for the inclusion of words like *ornament*, *mere* and *vessel* (Carter, 1987a: 165).

5. “words that, in their most general sense, have the most widespread usage across a wide range of genres”: Core words, according to this definition, should be words used in as many different situations as possible. These are highly dispersed terms, not those restricted to a small range of contexts.

6. “words that are most general, or unmarked, or central to the language”: General words by this definition are those that “can substitute for other words” (Lee, 2001: 254). Superordinates would fall into this category. *Bird* would qualify, but not *robin*. The words in Ogden’s BASIC English (1930) would fit this definition (See section 2.5).

7. “words that are cognitively basic or most salient”: Words that are cognitively basic are those that generally come to mind, when speakers imagine something. For example, *water*, *tea* and *coffee* are more salient than *liquid*. Americans typically think of a *robin* when they hear the word *bird* (Rosch, 1976). In both of these examples, the hyponyms are arguably more salient than the superordinates which creates a certain amount of tension between this and the previous definition. This conception of core vocabulary was first introduced by Gougenheim Michéa, Rivenc, and Sauvageot (1964: 13, 135ff). They used to the term *disponibilité* to describe a word’s mentally “availability” despite the fact it may not show up frequently in utterances (See section 1.1.1).

The list of characteristics below follows the order given in Stubbs (1986). He also organized these characteristics into three categories. Stubbs used the term *nuclear* rather than *core*, but his list of characteristics is nearly identical to Carter’s (1987b). The number in

parentheses after each defining characteristic is the number that characteristic is given in Carter (1987b).

A. Core words are “pragmatically neutral and occur in a wide range of contexts.” These characteristics correspond roughly to definitions 4-6 in Lee (2001).

1. Neutral register: core words have no “attitudinal, emotional, evaluative connotations.” (8)
2. Culture-free: core words are less likely to have culture-specific uses. One could rephrase this to say that core words are less-culturally marked, since all utterances are cultural acts. (6)
3. Pragmatically neutral: core words are not restricted to a particular semantic context. (9)
4. Neutral tenor: core words are not marked as formal or informal. (10)
5. Summarizing: informants are more likely to use core words when summarizing. (7)

B. Syntactic and Semantic relations between nuclear words:

6. Superordinates: core words tend to be superordinates rather than hyponyms, but not exclusively. (5)
7. Generic: core words have evolved a wider range of syntactic patterns than non-core words. Carter (1987b: 180) cites the example of *give* which could be followed by an indirect object and a prepositional phrase, as in *I gave it to him*. The direct object could follow the indirect object without a preposition: *I gave it him*. Or, the order of the objects could be reversed: *I gave him it*. A non-core word like *donate* is only used in the first pattern: *I donated it to him*. (1)
8. Core words are used to define other non-core words. This characteristic is identical to the first definition from Lee (2001) above. (1)
9. Collocability: core words have a wider collocational range. That is, they are not strongly tied to particular collocational patterns. *Blonde*, for example is strongly associated with its collocate *hair*, whereas *yellow* is much more flexible in its collocations. (3)
10. Extension: core words have a more developed range of meanings, are more often a part of multi-word units, idioms and phrasal verbs. Ironically, Stubbs is arguing that polysemy is evidence of membership in the core, whereas others might consider it necessary to count markedly divergent senses separately. (4)
11. Core words are more likely to be incorporated into compound lexical items. (4)

C. The core is “a unified whole” (Stubbs 1986: 109). Core words have well-developed semantic links to each other. Meara (1980: 12ff) reviews studies of semantic word associations in native speakers, learners and bilinguals of many languages.

12. Antonymy: core words have well known antonyms. (2)

A well-worn list like the Longman Defining Vocabulary satisfies Lee’s first definition of a core vocabulary. Lee’s second through fifth definitions are amenable to corpus studies. The remaining definitions and characteristics would require subjective judgements or experimental studies of each word one was considering for inclusion in the core vocabulary.

An intuitively crafted list, like West’s GSL, is quite similar to a defining vocabulary. West’s GSL was honed over many years as he composed graded reading materials. The LDV was developed for lexicography, whereas the GSL grew out of narrative and informational writing. One could also use a survival vocabulary, such as Crabbe and Nation (1991), as an intuitive core vocabulary list. Such lists are designed to be an essential vocabulary for oral communication. In this sense, a survival vocabulary is to speech what the LDV and GSL are to writing.

One additional test of a core vocabulary could be a kernel vocabulary used by historical linguists in looking for historical links between languages, e.g. Swadesh (1971: 283). In related languages, these kernel words tend to be cognates. These words are so central to human languages, that they are less likely to be replaced by borrowing. At 100 words, the Swadesh word list is hardly a complete core vocabulary, but a core vocabulary should include the majority of these words that have stood the test of millennia. A kernel vocabulary list is still an intuitive creation, in this case, crafted through years of research reconstructing proto-forms of ancient languages. The biggest problem with these intuitively crafted lists is not their content, but the methodologies used to create them. An intuitive methodology that takes decades to compile is not easily replicated. These existing lists could, however, serve as tests of objectively compiled lists.

2.5 Word lists and corpora

The following review of the history of word lists and corpora is roughly chronological rather than thematic. The stories behind word lists and corpora are intertwined and involve multiple overlapping themes. Teasing the thematic strands apart into isolated discussions would create a misleading impression of linear progress. One might fail to see the meandering and sometimes tumultuous journey which brings us to the present day. This review also

delves further into the past than some might consider necessary. However, seeing researchers and practitioners centuries in the past confronting the same issues that modern applied linguists face is both fascinating and humbling. Solomon's ancient observation still holds true: "There is no new thing under the sun."

2.5.1 The origin of word lists dates from at least the time of the first known writing system around 3400-3300 BCE in the Sumerian city of Uruk in modern-day Iraq. Fifteen percent of extant proto-cuneiform tablets are word lists (Gnanadesikan, 2009: 15). The other 85% are related to administration or accounting. Written texts are prototypically informational. This tendency, which Biber (1988, 1989, 1995) discovered in his multidimensional studies, has its origin in the very first written texts. Writing systems were almost universally created for the exchange of information. One very rare exception is the script invented by the Hanunóo people of Mindoro in the Philippines which they have used "primarily for exchanging love poetry" carved on pieces of bamboo (Gnanadesikan, 2009: 180). Unlike the vast majority of early written texts, these poems would certainly have high positive scores for involvement.

Corpus linguistics begins with the study of fixed religious or classical texts, such as the Bible, the Qur'an or Homer's *Iliad*. Some of the techniques that were used, such as concordancing, are still essential tools for modern research. Modern corpus linguists are indebted to those who expended a lifetime of effort pioneering techniques that we can now execute in milliseconds. As was seen in section 1.4, many of the terms these ancient scholars coined while pouring over scrolls or sheets of papyrus are still in circulation. Their focus on written discourse is also a lasting legacy that influences the search for core vocabulary to this day.

Not long after Greek grammarians had codified the text of Homer's *Iliad* in Alexandria, Roman children were learning Greek as a second language. Some of the texts they read contained alphabetical or topical wordlists. Such lists focused on rare rather than core vocabulary. The Roman rhetorician Quintilian writes that, "following the traditional practice . . . it will prove advantageous to ensure that he* does not waste his time on pedestrian, everyday words. He can from the start memorize the meanings of more recondite vocabulary" (cited in Joyal, McDougall, & Yardley, 2009:172). Quintilian assumes the pupils are boys since girls rarely received formal education. "[T]he educated Roman elite culture was essentially bilingual" so the learners apparently were expected to have already acquired the core vocabulary of Greek naturalistically before formal instruction began (Saenger, 2000: 91; Bowen, Masden & Hilfery: 1985: 6ff). Unfortunately, no data exists on the initial vocabulary knowledge of those Roman school children. As mentioned in chapter one, I

mistakenly repeated this focus on rare vocabulary in a teaching context where the learners were not yet familiar with the core vocabulary.

2.5.2 Early core vocabulary lists. Some of the first known word lists that better approximate a core vocabulary list can be seen in the 16th century. Two of these early foreign language textbooks are *The Schoolmaster* and *The Littleton* by Claudius Holyband, a protestant refugee from France, who taught French in London. “Both books contain lengthy vocabulary lists arranged in topic areas” (Howatt, 2004: 27). The vocabulary is presumably related to the scene-based dialogues that are the focus of the lessons. Each lesson is a scene from an extended narrative – almost like a drama. Not unlike many modern textbooks, the dialogues dealt with school, home, travel and business among other topics. Business might seem out of place as a topic for school children. In fact, Nation (1990: 20-21) cited the inclusion of this sort of vocabulary as one of the key problems with objective frequency lists. Holyband’s pupils, however, came from merchant families who would have wanted such values instilled in their children at a young age. In these early texts, one already one sees a focus on speech and an exploration of different speech domains (c.f. Biber, 1993: 246).

In 1588, Timothe Bright published a text on shorthand in English, but it did not contain symbols for phonetic transcription (cited in Fries & Traver, 1960: 2). Instead it contained a list of 559 words and a set of symbols to represent them. These 559 words were intended to cover the meanings of the majority of words, so the symbol for *agree* would also cover *consent*. West, Palmer and Ogden (Richards, 1974: 73-75) would later suggest this sort of broad semantic coverage as a criterion for selecting core vocabulary (c.f. Carter, 1987b; Lee, 2001; Stubbs 1986 in section 2.4 above).

Richard Mulcaster, a schoolmaster from Cumbria and a contemporary of Bright’s, was eager to see English spelling standardized, and so compiled a very long wordlist with more systematic spellings for his students. Variant spellings are still an issue, especially in the transcription of speech (as will be seen in the chapter three), but without standardized spellings most modern corpus techniques would be impossible.

In the early 17th century, Jan Amos Comenius, an innovative Latin teacher from Moravia, encouraged the use of realia (often live animals) in the classroom to teach vocabulary. Comenius actually advocated mother-tongue instruction, but he designed a four-stage curriculum for teaching Latin as a foreign language that exemplifies some of the same goals seen in the current study.

The first stage of Comenius’ curriculum, “the *Vestibulum* should contain a few hundred words, sufficient for simple conversation on everyday things with an accompanying word list” (Howatt, 2004: 47). The second stage, *Janua*, “should aim at teaching about 8,000

words in a series of graded texts of intrinsic interest and educational value.” These textbooks sold very successfully, but Comenius “was particularly angry with teachers who tried to ‘improve’ the *Janua* by stuffing it with ‘uncommon words and with matter quite unsuited to a boy’s* comprehension’” (Howatt, 2004: 49). Girls were not typically sent to school in 17th century Europe either. Comenius’ recommendations closely resemble Lee’s definition of core vocabulary as those “that are cognitively basic or most salient” (2001: 254). Words that are cognitively basic are those that generally come to mind, when speakers imagine something.

The textbooks for the third and fourth stages were to cover style, usage, translation and linguistic comparisons, but these were never completed. He did, however, create a bilingual picture dictionary with woodcut illustrations entitled *Orbis Sensualium Pictus*. Unlike, modern EFL publications, the use of pictures in a textbook would have been considered very unusual at the time. He designed the text to be at the center of engaging classroom discussions about students’ thoughts and feelings, so cognitively basic terms would have been a high priority. Comenius’ focus on face-to-face conversation related to the “here-and-now,” his conscious grading of vocabulary, avoidance of unnecessary low-frequency words, and building from a small core of common words toward a larger literate vocabulary of several thousand words are all ideas which would resurface in more modern language teaching. In fact, while ignoring the difference in language and the lack of a common definition of what will count as a word, the size of the graded vocabulary Comenius suggested is roughly the same as this thesis will propose – only he did so 400 years ago! Unfortunately, the wisdom of many of his methods was not recognized for another two hundred years or more (Howatt, 2004: 53).

Holyband’s and Comenius’ texts and vocabulary choices were designed to fit the learning needs of school-age foreign language learners. Another émigré from France, Paul Festau, was teaching French in London to adult English speakers who wanted to be fashionable. When persecution of the Huguenots resumed in France in 1685, a new wave of refugees arrived, and Festau quickly adapted his French as a foreign language textbook to serve as a text for teaching English as a second language to the newly arrived French speakers. The “frivolous salon chit-chat” in the dialogues and the accompanying vocabulary lists had been appealing to his middle-class English learners, but did not fit the practical needs of the new immigrants (Howatt, 2004: 58). These EFL learners desperately needed basic communication skills, including a survival vocabulary (c.f. Crabbe & Nation, 1991).

A few years later, Guy Miège, a Swiss immigrant to England who had worked as a lexicographer, published *Nouvelle Méthode*, which “is written in a much more appropriate manner and is a landmark in the development of English teaching” (Howatt, 2004: 58).

Miège's texts contain much more linguistic information about English grammar, pronunciation, orthography and word forms. Miège's background as a lexicographer enabled him to begin to address the complexity of word knowledge and what constitutes a word.

Samuel Johnson's landmark *Dictionary of the English Language* (1755) is also a word list, in a sense, since he sought to capture current (written) word usage in the 1750s. Speech is recognized only in so far as he documented the current pronunciation of words. Though not corpus-based, the dictionary was at least based in quotations from authentic texts (Howatt, 2004: 110-116). Johnson had to struggle with some of the same methodological choices that surface in the studies of this thesis. For example, what does one do with marginal words (Johnson called them "barbarous.") and borrowings from other languages? Johnson also sought to include everyday vocabulary as well as specialized vocabulary. Johnson's dictionary contained 45,000 headwords. In the next century, Noah Webster (1828) expanded on Johnson's work bringing the total to 70,000. One could now begin to contemplate the question, "How many words are there in a language?"

By the end of the 18th century, English was being taught as a foreign language across Europe. (The map on Howatt, 2004: 66 shows the spread of publications.) Miller (1797) published the first known non-European EFL textbook in Bengal. The text included a vocabulary list that is "remarkable for the period in that it avoids all the 'worthy' and 'over-literate' words which authors of the time would almost certainly have included as evidence of their own erudition and to 'improve' the natives" (Howatt, 2004: 72). In giving the forms of the verb *speak*, however, Miller does include *speaketh*, which was considered archaic even at that time. Since writing preserves as well as transmits language, its very existence creates tension between past and present speech. Overcoming this pull toward written language is one of the greatest challenges faced when creating core vocabulary lists.

The grammar-translation method was developed in the late 18th century in Germany and became the dominant methodology in the early 19th century. Grammar-translation made extensive use of wordlists, but also ignored spoken language. Vocabulary items were often chosen to illustrate grammatical structures rather than for their usefulness in communication (Zimmerman, 1997: 6). Except for Pestalozzi (1836), who used a natural approach to language teaching and also produced primers that contained graded vocabulary, conversation was not taken seriously again in language instruction until Berlitz opened his language schools in the United States in the 1870s.

In the second half of the 19th century, a British civil servant and avid language learner, Thomas Prendergast, devised an elaborate system for combining a limited vocabulary to cover a wide range of communicative functions. Of the 214 words in his core vocabulary, "82 per

cent are among the most frequent words on the Thorndike-Lorge (1944) list and another 14 percent in the second 500” (Howatt, 2004: 176). He replicated this limited productive vocabulary for French, German, Spanish, Latin and Hebrew (Howatt, 2004: 175). This idea of core words having broad functionality fits Lee’s (2001) fifth definition of coreness.

2.5.3 Attempts at objectivity. Until the 19th century, word lists were compiled on a purely intuitive basis. It was arduous enough to apply objective methods like frequency counting to a single, static text like the Bible. Trying to objectively explore the vocabulary of a living language must have been unthinkable to most educators. Nonetheless, several researchers embarked on such projects. These counts are usually related to writing or literacy training. Not surprisingly, quantifying even written language proved extremely difficult.

Except for two outstanding examples, these early studies were small by today’s standards. Pitman (1837) collected 10,000 words in the development of his stenographic system. Freeman (1820) compiled a 20,000 word corpus for his work in adult education (McArthur 1998: 51-52). The exception was Käding’s enormous *Häufigkeitwörterbuch der Deutschen Sprache* (1898). Käding compiled his corpus to inform the teaching of German stenography. Five thousand stenographers worked on the project collecting almost 11 million words – larger than many modern computer-based corpora (Fries & Traver, 1950: 4).

In the United Kingdom, Knowles (1904) collected 100,000 words as he was developing a reading system for the blind. He extracted 353 words which he claimed gave a text coverage of 75% (McArthur, 1998: 52). *Coverage* here does not refer to coverage of words’ meanings, but to how many of the tokens (running words of a text) are represented by the words on the list. From a receptive perspective, a core list should be composed of words that give the greatest coverage. On the other hand, words that may be the most syntactically flexible, semantically general or culturally neutral may not be the most frequent. As Bell (2012) points out, “the relationship between frequency and coreness is not simple.”

In these early studies, one can see the beginnings of corpus design as researchers decide what type of texts to collect. Eldridge, for example, searched four newspapers published in Buffalo, New York to discover “what words occur the most frequently in ordinary use” in order to develop a “limited universal vocabulary” (1911: 3). His study in some ways foreshadows Ogden’s goals in creating BASIC English (see section 2.5.4).

In 1914, Jones “analyzed a count of 15,000,000 words from children’s compositions” (Nisbet, 1960: 53). Cook and O’Shea (1914) created a corpus from the personal correspondence of thirteen adults in the United States as part of research related to spelling. They also noticed that a small number of very frequent words make up a large percentage of

any text. This is the frequency distribution Estoup (1916) and Zipf (1935, 1949) would soon describe mathematically.

In preparing another frequency list for stenography, Dewey (1923) decided to compile his own corpus of 100,000 words since he felt previous lists had not drawn upon a sufficiently broad group of texts (McArthur, 1998: 52).

In these early attempts at objective word lists, one already sees many of the themes related to the research questions of this thesis: designing a corpus that is representative of everyday language, objectively determining which items occur most frequently, and the aim of facilitating language learning or international communication through a limited core vocabulary and graded learning materials. During the 20th century, researchers began to explicitly debate whether intuitive or objective methodologies are more appropriate for creating core vocabulary lists. This a debate continues today.

2.5.4 Intuition versus objectivity. With the Reform Movement in the late 19th century an emphasis was again placed on spoken language in teaching. These “new use-based ideas . . . coalesced into what became known as the *Direct Method*” (Schmitt, 2000: 12).

In 1902, Harold Palmer started his teaching career at a Berlitz-style school in Belgium where he became familiar with the Direct Method and eventually opened his own language school. Even then, Palmer “already had the idea that an economical approach to English might be made by first learning the words most frequent in occurrence . . .” (Palmer, 1936: 371). However, after seeing Eldridge’s (1911) list, “he concluded that either there was something radically wrong about the Eldridge principles of listing, or that such statistics were of little or no value to students of foreign languages” (Bongers, 1947: 74).

Unfortunately, since the existing frequency lists were idiosyncratic in many respects, Palmer had reason to doubt the usefulness of frequency-based methods. Palmer argued that, if necessary, one had to “subordinate frequency to expediency” (Palmer, 1917: 130). One tested a word list, e.g. Palmer and Hornby’s *Thousand Word English* (1937), by actually using that range of words to write ELT materials. “Just as the proof of the pudding is in its eating, the proof of the vocabulary is in the degree of its smooth and natural functioning when put to the test” (Palmer, 1934: 51).

When Palmer’s word choices were criticized as subjective, he countered that the objective lists were compiled from sources that were selected subjectively (Bongers, 1947: 12-13; Tickoo, 2008: 297). This was, of course, true at the time. For example, in creating his first *Teacher’s Wordbook* (1921), E. L. Thorndike, compiled a corpus of 4 million words, 3 million of which “came from the Bible and the English classics” (McArthur, 1998: 53). This selection likely reflects his own tastes and values rather than an attempt to represent English

in the 20th century. Even in Thorndike and Lorge's final refinement, *The Teacher's Wordbook of 30,000 Words* (1944), "doorknob [is] as rare as *dulcet*" and "damsel [is] as frequent as *error*" (Bright & MacGregor, 1970: 17). As seen in Section 2.3, the choice of texts to include is a key methodological question when compiling a corpus.

Besides the subjective choice of source texts, Palmer also pointed out that the frequency lists lacked clarity on what counted as a word. Most lists relied on a simple orthographic definition. In Thorndike's first *Teacher's Wordbook* (1921), words were defined as types, i.e. *am*, *is*, *are*, *was*, *were*, and *be* counted as separate words. However, some of the counts used in the subsequent *Teacher's Wordbook of 20,000 Words* (1932) grouped types together, counting each lemma as one word. As was mentioned earlier, these differing word counts were spliced together without redoing the first counts. "List-makers were content to tinker with counts already made, probably because of the sheer difficulty of making the counts in the days before computers could absorb the tedium" (McArthur, 1998: 54-55).

A further problem Palmer criticized was that frequency lists ignored homographs and polysemes. A fairly useful word may not have occurred in the selected material with sufficient frequency to be considered, whereas a rare sense of an otherwise frequent word could slip through unchallenged. "The text simplifier must not exaggerate the claims of quantitative objectivity, for if he does, he will be straining at vocabulary gnats while swallowing semantic camels" (Palmer, 1934: 49). Thorndike and Lorge tackled the problem of polysemy in the *Teacher's Wordbook of 30,000 Words* (1944), but they based their sense divisions on the *Oxford English Dictionary*, itself an amalgam compiled layer upon layer (McArthur, 1998: 56). As noted in the previous section, Stubbs (1986) argues that polysemy and a word's inclusion in many collocations or idiomatic expressions indicate it is a core word since it has been central to the language for long enough to have evolved a variety of uses.

Thorndike and Palmer represent these two different philosophical approaches – objective and subjective – to capturing the core of the English language in a wordlist. Thorndike's objective approach sought to document word usage through real-world observation. Palmer, was unconvinced since he believed it failed to match the intuition of experienced teachers.

McArthur, in his 1998 review, places Michael West, who was teaching at the Teachers' Training College and the University of Dacca, and Lawrence Faucett, at St. John's University in Suzhou, in the subjective camp along with Palmer, but Palmer was "much more reluctant than West or Faucett to acknowledge the relevance of existing word counts" (Smith, 2003: xiii). Faucett, for example, suggested using a combination of Thorndike's (1921) and Horn's (1926) frequency lists to divide the radii (levels) of Palmer's IRET 3000-word list. Faucett

and Maki published their combination of the Thorndike and Horn lists in 1932. This list and Palmer's IRET word-lists were foundational to the Carnegie Report in 1936 (Smith, 2003: xviii).

The Carnegie Conferences in New York in 1934 and London in 1935 were pivotal in the history of word list construction. How these researchers came together has as much to do with personalities as it does with methodologies. The conferences were initiated by West and were also attended by Palmer, Faucet and Thorndike. The catalyst that brought them together was in some ways another researcher, C. K. Ogden, who was invited but declined to attend.

In 1930, C. K. Ogden had published a short list of 850 words he called *British American Scientific International Commercial* (BASIC) English. He argued BASIC could function as a medium for international communication in its own right. Unlike Palmer, West, Faucett, etc., Ogden was not concerned with whether his list was an adequate introduction to the language. Ogden wanted his list to be a language – similar to Esperanto (Howatt, 2004: 283-288).

West (West et al., 1934) criticized BASIC, which apparently deeply offended Ogden who responded with a book entitled, *Counter-Offensive* (1935), in which he claimed BASIC could be learned in “a week or at worst a month.” BASIC does limit the number of vocabulary forms a learner needs to remember, but this limited vocabulary is still required to cover a large number of meanings. Fries and Traver (1960: 81) estimated that Ogden's list of 850 words includes 12,425 different meanings. It is possible to learn 850 word forms in a month, but that is not the same as learning a language, even an artificial one like, BASIC.

Faced with the possibility of this, at times, pidgin-like form of English being taught in place of standard English, “the two ELT reformers [Palmer and West] [saw] the need to come together under the same umbrella in their shared agenda of serving situated EFL with tools more responsive to its perceived needs and demands” (Tickoo, 2008: 325).

If one looks at Ogden's BASIC as a core vocabulary list, rather than a language in its own right, it is not unlike the subjective lists produced by Palmer. In fact, there is a 93% overlap between the words in Ogden's BASIC and the list produced by the Carnegie Conference (Fries & Traver, 1960: 74). “Ogden may not have attended the conference, but his words did” (McArthur, 1998: 65). The acrimonious way in which the debate took place is unfortunate, but Ogden's search for a core English vocabulary did influence the development of EFL vocabulary teaching, if only from the sidelines.

West eventually published the *General Service List of English Words* (1953) based on the results of the Carnegie Conferences. The GSL in some ways represents a merging of corpus-based and intuitive methodologies (Nation, 1990: 22-23). Cowie (1999: 22-25) and

Howatt (2004: 289-290) argue that the frequency figures were added after the list had already been agreed upon subjectively which would make the GSL more subjective than objective. Either way, the GSL marks an important milestone in the creation of core vocabulary lists. It does not, however, represent a replicable methodology.

Bongers, a long-time friend and admirer of Palmer, continued with a purely subjective approach, however. Bongers divided his final revision of Palmer's 3000-word list into three sublists, which he claimed gave a coverage of 89.5%, 5.6% and 2.4% respectively (97.5% in total) for a group of mostly narrative texts (Bongers, 1947: 165-66).

2.5.5 Early studies of spoken vocabulary. Probably the single greatest flaw in the objective methodologies was the neglect of conversational data. Josselson (1953) had attempted to balance "conversational" and "nonconversational" language in his Russian corpus though all of the material taken was from written texts (Ornstein, 1954: 428). Vakar (1966) attempted to improve on this by collecting 10,000 words "from 200 acts in 93 Russian plays" (Tolpin, 1967: 54). He believed this method would give "a close approximation of actual spoken Russian" (cited in Tolpin, 1967: 55).

Schonell, Meddleton and Shaw (1956) compiled a 500,000-words corpus for their Oral Vocabulary of the Australian Worker in order to create a basic vocabulary list for non-English-speaking immigrants to Australia. They reported that 2,000 word families covered 99% of the words in their corpus.

As part of an effort to support the role of French as a world language – a role it was losing to English – Gougenheim, Rivenc, Michéa and Sauvageot (1964) compiled a corpus of 312,135 tokens of authentic spoken French by recording and transcribing conversations and interviews with 275 informants from a variety of socio-economic and geographic backgrounds (Gougenheim et al., 1964: 64-66; Stern, 1983: 55). Their corpus contained 8,774 word types, of which they selected 1,063 (those with a frequency of 20 or more) to be the first stage in a vocabulary for the learning of French as a foreign language.

Howes (1966) collected 250,000 tokens of spoken English for use in psychometric research from interviews with university students and patients. He felt the Thorndike-Lorge list (1944) was inappropriate since it was based on writing. "For most experimental applications . . . the frequency of occurrence of a word in the spoken discourse . . . is the quantity one wants to estimate" (Howes, 1966: 572).

Beier, Starkweather and Miller's (1967) corpus of around 80,000 words is intriguing, not for its size, but due to the research question they sought to answer. The authors wanted to discover if Zipf's law also held true for the spoken vocabulary of children and teenagers. They concluded that the spoken vocabulary of their subjects did fit Zipf's law and that spoken

data of this sort could “help in building reading material which can be easily understood” (Beier et al., 1967: 225).

The London-Lund Corpus of Spoken English eventually provided a 500,000-word parallel for spoken British English (Greenbaum & Svartvik, 1990; Svartvik, 1990; Svartvik & Quirk, 1980). The corpus combines two projects: one by Quirk and Greenbaum at University College London begun in 1959 and another by Svartvik at Lund University in 1975.

Unfortunately, these studies did not set a precedent for the first computer-based corpora in English.

2.5.6 The Brown family of corpora. In the meantime, modern corpus linguistics and objective methods began in earnest with Kučera and Francis (1967). At Brown University, they compiled a million-word corpus of American English texts printed in 1961. The Brown corpus is one of the first attempts to create a balanced corpus that could represent a language variety. Kučera and Francis used stratified random sampling to select 500 passages of 2,000 words each from a variety of published materials. Their sampling frame, however, consisted of the books and periodicals held in the Brown University Library and the Providence Athenaeum. Spoken English was completely ignored.

The Brown corpus became a model for several similar corpora. The London Oslo Bergen corpus was compiled as a parallel corpus of British English texts (Johansson, Leech, & Goodluck, 1978). Shastri, who had done research under Leech at the University of Lancaster, went to work on a similar corpus of Indian English at Shivaji University in Kohlapur (Shastri, 1986). At Macquarie University in Sydney, researchers began compiling the Australian Corpus of English (Peters, 1987). At Victoria University of Wellington, two corpora were compiled, one of which finally recognized the importance of speech: the Wellington Corpus of Written New Zealand English (Bauer, 1993) and the Wellington Corpus of Spoken New Zealand English (Holmes, Vine, & Johnson, 1998).

In the 1990s, Mair compiled the Freiburg Brown and Freiburg LOB corpora as parallels to the original Brown and LOB corpora (Hundt, Sand, & Siemund, 1998; Hundt, Sand, & Skandera, 1999). Xu and Liang (2012) from the Beijing Foreign Studies University have just recently published another pair of corpora, the CROWN and CLOB corpora, that parallel to the Brown and LOB but are compiled from texts published during 2009.

2.5.7 The International Corpus of English. In 1988, Greenbaum proposed the International Corpus of English which should eventually include million-word corpora from over twenty countries where English is “a majority first language (for example, Canada and Australia) or an official additional language (for example, India or Nigeria)” (Greenbaum, 1996: 3). The ICE family of corpora are “informally based on the design of Brown and LOB,”

but they mark a significant advance in corpus design in that they contain “spoken English, and in a greater proportion than writing” (Nelson, 2006b: 737). Members of the ICE family of corpora contain 300 samples of spoken English and 200 samples of writing. Like the Brown corpus, each sample is 2,000 words in length. To date, ICE corpora are available from Canada, East Africa, Hong Kong, India, Ireland, Jamaica, Great Britain, New Zealand, the Philippines and Singapore. Parallel corpora, like ICE and the Brown family of corpora, make it possible to begin investigating English as an International Language.

With the WSC and ICE family of corpora, speech was at least put on equal footing with writing, but this discussion has deviated from the chronological narrative. To appreciate this advance in methodology, one needs to first return to the early 1970s before the London-Lund corpus was published.

2.5.8 The soup-pot method. In 1971, Carroll, Davies and Richman published a study of the vocabulary being used in schools in United States. The study was not designed to sample the language as a whole, but is noteworthy for its size and detailed statistical analysis. The researchers compiled a 5-million-word corpus comprised of extracts from elementary school textbooks used in various grades and subjects. *The American Heritage Word Frequency Book* contains the resulting frequency distributions for the corpus as a whole, as well as each grade level and subject (Carroll, Davies & Richman, 1971: 753ff). Though their list was designed for U.S. school children, the authors saw it as useful to EFL professionals as well:

One obvious major use of word frequency counts is to assist in the compilation of lists of words which language users are most likely or least likely to encounter, and hence to know or not know. The new word frequency book will supplement previous counts in yielding this type of information. The resulting lists would presumably have value in teaching either native speakers of English or persons learning English as a foreign language (Carroll, 1972: 1073).

Carroll’s suggestion that the American Heritage list could “supplement previous counts” seems to imply a sort of soup-pot methodology that remains appealing. The *Cambridge English Lexicon* is a good example.

In the introduction to the *Cambridge English Lexicon* (1980), Hindmarsh describes how he first created a list in 1970-71. After checking West’s GSL against the Cambridge First Certificate in English and other examinations, Hindmarsh deleted a portion of the GSL. He then added words from a series of graded readers, “a number of basic English vocabulary

lists,” and three frequency counts. Principles for adding or deleting items are not given. This produced the *Intermediate English Word List* which contained 4,200 words.

Hindmarsh then launched a new attempt using Thorndike and Lorge’s frequency lists. “I therefore took frequency 13 as my focal point, but broadened the focus up and down the frequency scale so that I ranged in fact from 19 down to 8, using *a teacher’s discretion* [emphasis added] in so doing”. To this, he “added a limited number drawn from the New Oxford English Readers for East Africa (OUP) Books 1-5” and “all the words from the Longman Structural Readers Handbook not already included.” The produced list was about 4,700 words. He then added words not already included from a frequency list from the Brown corpus and a South African word count by Wright in 1965. The list now contained 5,400 words of which he “now cut away some 200 items whose frequency did not meet certain criteria of statistical constancy through the three frequency lists, giving 5200 items.” He then added in any of the words from his first word list that were not already on the list. Eight hundred words from the first list were not already included giving a total of 6,000 words.

Six thousand words was beyond the scope of the examinations for which he was hoping to prepare students, so he pared the list down to 4,000 by cutting out “slightly archaic,” “academic” and “specialized” terminology. “At this stage the cutting out process was therefore radical and based principally on the *intuitions* [emphasis added] of teachers of EFL. At the same time lexical items reflecting the world today were drawn together from a number of sources and added to the list.” The final published list contains 4,500 items with 8,000 semantic senses.

The methodology Hindmarsh describes reflects years of teaching experience, but it would be impossible to replicate or even to consistently describe the principles used in creating this list. Nonetheless, it is perceived as authoritative. In a recent ESL blog post, Hindmarsh’s *CEL* was described as “a rare gem” (gramarye, 2012).

2.5.9 Academic word lists. The two lists discussed below are of academic rather than core vocabulary, but they demonstrate developments in objective methodologies for creating vocabulary lists.

The University Word List by Xue and Nation (1984) was designed to set vocabulary learning goals for EFL university students. The UWL sought to fill a gap between a core vocabulary and the specialized vocabulary that would be the subject of courses in each student’s discipline. Xue and Nation’s list is not directly corpus-based, but involved splicing four existing lists of academic vocabulary. The first list is Champion and Elley’s (1971) list of academic vocabulary that was created for an EFL entrance examination for New Zealand universities. The Champion and Elley list was based on a corpus of 301,800 words from

lectures, journals and examination papers from the 19 disciplines with the highest enrollment in New Zealand universities. Their list assumed knowledge of the most frequent 5,000 words on Thorndike and Lorge's *The Teacher's Wordbook of 30,000 Words* (1944). Xue and Nation first merged the Campion and Elley list with the American University Word List by Praninskas (1972). The AUWL was based on an analysis of 272,466 words from ten first-year university textbooks from ten different disciplines. Praninskas excluded the words on West's GSL from his list.

After merging these two lists, Xue and Nation compared this combined list with two additional lists by Lynn (1973) and Ghadessy (1979). Lynn, at Nanyang University in Singapore, and Ghadessy, at Pahlavi University in Iran, compiled their lists by recording words which students had glossed or annotated in their English-language textbooks. Xue and Nation found an approximately 70% overlap between each of these lists and their combination of the two corpus-based lists. They then added high-frequency items from the Lynn and Ghadessy lists to create their final University Word List. The threshold for high-frequency is not given.

On the surface, Xue and Nation's methodology is objective. However, by utilizing existing lists they have inherited all of the subjective decisions and inconsistencies in the previous studies. For example, Praninskas excluded words she considered to be technical terms specific to particular disciplines, but kept some technical terms since they also had more general senses. She also excluded some function words, such as *nevertheless*, that she thought could be better dealt with in a grammar course. Each list was based on a different range of academic disciplines and also excluded a different list of basic vocabulary. Thorndike and Lorge's list and West's GSL each comes with its own mix of subjectivity and inconsistency. The procedure for creating the UWL has been described objectively, but it has a thoroughly subjective ancestry.

The Academic Word List (Coxhead, 2000) marks a major step toward a principled objective methodology. Coxhead constructed an academic corpus of 3.5 million words drawn from "28 subject areas organised into 7 general areas within each of four disciplines: arts, commerce, law, and science" (2000: 216). The corpus is made up of numerous journal articles, textbooks and manuals which Coxhead collected herself or were part of existing corpora: the Brown, LOB, Wellington Written and MicroConcord (Murison-Bowie, 1993) corpora. "The resulting corpus contained 414 academic texts by more than 400 authors" (Coxhead, 2000: 219). Coxhead actually fell short of her original goal of collecting 4 million tokens of academic writing. The difficulty in collecting an amount close to that meant that she had to include 2,000-word samples from existing corpora as well as many complete textbooks.

Coxhead did, however, balance the number of long, medium and short texts within each subject area (Coxhead, 2000: 221).

Coxhead chose the word family as the unit of counting so as not to discount closely related word forms. Nagy, Anderson, Schommer, Scott and Stallman (1989) found evidence that both the inflectional and derivational relationships between words, i.e. word families are represented in a speaker's mental lexicon.

Since Coxhead was not seeking to develop a general vocabulary list, she excluded words from West's GSL, which she considered adequate for her purposes. In selecting words for the AWL, Coxhead sorted by range (the number of texts in which a word occurred) and then by frequency (total number of occurrences in the corpus) so as not to bias the list in favor of words from the longer texts in the corpus. To be included "[a] member of a word family had to occur at least 10 times in each of the four main sections of the corpus and in 15 or more of the 28 subject areas" (2000: 221). She also set a frequency threshold of 100 tokens to exclude widely dispersed, low-frequency words. In all, Coxhead isolated 570 word families that met her criteria.

Hyland and Tse (2007: 239) criticize the use of different-sized texts in the corpus and the fact that there are a differing number of texts in each of the subject subcorpora. They also suggest that since words often have significantly different senses in different disciplines, it may not always be a valid conclusion that a shared orthographic form is part of a common academic vocabulary.

Despite the perceived weaknesses pointed out by Hyland and Tse, the AWL represents a major step forward in the methodology for creating lists. A conscious effort was made to balance the types of texts that represent the type of language being studied. In this case, a corpus composed of purely written texts is appropriate since the purpose is to create a list to aid the development of academic reading skills. The use of both range and frequency to objectively isolate the words included on the list is also informative. The fact that an objective methodology is clearly explained and adhered to is in itself a major advance.

2.5.10 Mega-corpora.

Collins COBUILD and the Bank of English. Given the Zipfian distribution of words, even a million tokens is too small a sample for examining lower frequency words and features. In 1980, a joint venture between Collins Publishing and a team of researchers from the University of Birmingham led by John Sinclair established the *Collins Birmingham University International Language Database* (COBUILD) that went far beyond the million-word mark that had become the norm. In 1982, their initial corpus contained around 7 million words of spoken (25%) and written (75%), British (70%), American (20%) and other varieties

of English (Kennedy, 1998: 46ff). A *Reserve Corpus* was also compiled to which texts were continuously added, “not according to a pre-conceived framework giving a balance of genres, but especially according to variety of topic” (Kennedy, 1998: 47). In the 1990s, COBUILD became part of an expanded corpus called *The Bank of English* that contained 320 million words by 1997. *The Bank of English* is now part of the Collins Word Web which contains 2.5 billion words (HarperCollins Publishers, 2004a, 2004b). With *The Bank of English* and the Collins Word Web, the quantity of data has been radically increased, but how that relates to representativeness is very unclear since the collections seem to be opportunistic rather than principled in their gathering of texts.

The British National Corpus is a more easily accessible mega-corpus. The BNC was developed “by an industrial/academic consortium led by Oxford University Press” (Daulton, 2003: 33). The British government also supported the project providing half of the funding. When it was completed in 1994, the corpus contained 90 million written and 10 million spoken words of British English. The spoken portion of the corpus is made up of 5 million words that were collected by asking a group of demographically selected informants to carry *Walkman*™ tape recorders with them recording their conversations throughout the day and another 5 million words taken from speaking situations where the context strongly determined what was said, such as a business meeting or public speech. Though only 10% of the total corpus, this “was at the time the largest collection of spoken English ever assembled” (Kennedy, 1998: 50). A corpus of American English comparable to the BNC was proposed in 1998 (Fillmore, Ide, Jurafsky, & Macleod, 1998). To date, 22 million words of the corpus are available (Ide & Suderman, 2006).

2.5.11 Word lists from mega-corpora.

Frequency bands for the LDOCE. Using the British National Corpus, Longman Lancaster and the Longman Corpus of Spoken American English, Kilgarriff (1997) created frequency lists for the first 3,000 words in English for spoken and written usage. These were used to mark frequent headwords in the *Longman Dictionary of Contemporary English 3*. Kilgarriff’s study, is especially interesting since it concludes that there should be two core vocabulary lists – one for spoken and the other for written English.

The range of texts in which a word occurred was apparently not used as a criterion when sorting, since Kilgarriff discovered “[w]ords such as *peptide* and *endoscopy* needed weeding out of the frequency list” since one source, *GUT: Journal of Gastroenterology and Hepatology*, “accounted for 713,000 words (0.7%) of the [written] corpus” (Kilgarriff, 1997: 139). Kilgarriff reports that this problem was more common in the American corpora and the BNC spoken corpus since they were much smaller. In other words, the corpora were not large

enough to give stable results for these words. Words that were high frequency, but were restricted to very few texts, had to be manually deleted.

Kilgarriff's team felt the context-governed portion of the spoken BNC also unduly raised the frequencies of formal words. "The lexis of meetings was playing a dominant role. So we subjected the list to a further edit, on the basis of frequencies of the demographic half alone" (Kilgarriff, 1997: 138). The "demographic half" refers to the collection of largely casual conversation that makes up around half of the BNC spoken corpus. This implies that casual conversation actually gives a more reliable indication of the frequency of core vocabulary, and that mixing even casual and formal speech leads to muddled results.

In the end, Kilgarriff still considered meetings to be overrepresented in the BNC spoken corpus, so two hundred items of "meetingese" were deleted. One hundred American words were added on the advice of an American lexicographer, since the American spoken corpora were much smaller than the British (Kilgarriff, 1997: 147).

Kilgarriff acknowledged that these decisions were subjective, and he speculated that researchers working from another corpus would find different results for both the third written and spoken frequency bands. There would be "very little change to the W1 and S1 items" (Kilgarriff, 1997: 151). These estimates are similar to those of Fries and Traver (1960: 17). Though Kilgarriff began the project of assigning frequency bands to the LDOCE using an objective methodology, he was eventually forced to compromise and merge both subjective and objective techniques.

The introduction to Hindmarsh's (1980) *Cambridge English Lexicon* contains an ominous warning:

At first sight it may appear that making a lexicon ought to be easy enough; all one should need is a set of clear working principles, and the will to apply them. But language is too subtle a matter to submit to such regimentation; exceptions begin to occur almost at once, and principles lose their clarity of contour (vii).

Nation (2004) created a 3,000-word list base on the BNC as a possible replacement for the GSL. He began with a list of around 6,500 of the most frequent lemmas from the BNC as a whole. He then removed those words which had a range of less than 98 – that is, they occurred in less than 98 of the 100 sections of the BNC. This list was then sorted by dispersion, and those with a dispersion of 80 or greater were subsequently sorted by frequency. From this list, Nation selected the most frequent thousand lemmas and expanded them to word families. To complete the first 1,000-word list, he had to manually add *goodbye*, *OK*, *Oh* and "[a] full list of days of the week, months, numbers, and letters of the alphabet . . . even

though several of these did not meet the frequency, range, or dispersion criteria” (Nation, 2004: 5). The lists of the second and third thousand were created in the same way, except that five word families were added to the 3,000 list because they were very frequent in the spoken corpus, but not in the written.

The use of dispersion and range in addition to frequency is instructive, but the over-representation of written data was still a hindrance to creating a list on a purely objective basis. This problem resurfaced when Nation and Beglar (2007) created their Vocabulary Size Test based on the BNC.

2.5.12 Ersatz conversation data. Given the prohibitive costs of compiling large quantities of casual spoken language, Burgess and Livesay (1998) compiled the Hyperspace Analog to Language corpus of approximately 131 million words from Internet newsgroups since “the text is very conversational and noisy, much like spoken language” (1998: 273). Burgess and Livesay found that a sample of nouns with a frequency of greater than or equal to 100 in the Brown corpus correlated well with the frequencies in their HAL corpus. Those nouns in the sample with lower frequencies correlated much more weakly. However, when compared to reaction times in word recognition tests, the frequencies from HAL correlated much better than those from Brown.

The reliability of frequency lists is important in such psychology studies since “[w]ord frequency is the most important variable in research on word processing and memory” (Brysbaert & New, 2009: 977). This led Brysbaert and New to wonder how large a cost effective corpus needs to be. They collected another analogue to speech in the form of subtitles from TV and films to build their SUBTLEXus corpus. In all, they compiled “51.0 million words (16.1 million from television series, 14.3 million from films before 1990, and 20.6 million from films after 1990)” (Brysbaert & New, 2009: 980). When they correlated these with reaction times, they found a similar degree of correlation for words with a frequency of 20 or more per million at a corpus size of 1 million. Lower frequency words (<10 per million) “seem to require a corpus size of at least 16 million words for reliable estimates” and “little gain seems to be made beyond” (Brysbaert & New, 2009: 980). Reaction times are not a precise enough measure for distinguishing words for a core vocabulary list, but these corpus sizes are encouragingly small.

2.5.13 Modern corpora of spoken English. Excellent efforts to collect authentic face-to-face conversation continue. Du Bois, Chafe, Meyer, Thompson, Englebreton and Martey (2000-2005) compiled a quarter of a million words from 1988-1996 that make up the Santa Barbara Corpus of Spoken American English. The team did an exemplary job of collecting data from a very wide variety of authentic conversational contexts.

The 5-million-word Cambridge and Nottingham Corpus of Discourse in English (CANCODE) was specifically collected for research into spoken discourse (McCarthy, 1998). Unfortunately, access to this corpus is restricted to researchers connected with Cambridge University Press.

In 2011, the XML version of the Vienna-Oxford International Corpus of English became available. Compiled by Seidlhofer, Breiteneder, Klimpfinger, Majewski, Osimk and Pitzl at the University of Vienna with the support of Oxford University Press, VOICE contains over 1 million words of English as a lingua franca in a variety of speaking situations involving 1,250 speakers from around 50 different language backgrounds.

2.5.14 The Internet as corpus. The Internet is itself an unprecedented resource of digitized texts, the potential of which researchers are just beginning to utilize. An example is The Corpus of Contemporary American English (COCA) developed by Davies (2008-, 2010) at Brigham Young University. Like *The Bank of English*, COCA is a monitor corpus to which texts are added continuously. COCA currently contains approximately 450 million words of written American English made up of roughly equal portions of fiction, popular magazines, newspapers and academic journals. In addition there are approximately 95 million words of spoken American English collected from reasonably accurate transcripts of oral interviews and conversations from TV and radio. An intriguing question is how closely the vocabulary of these public conversations matches that of everyday face-to-face speech. Since they are public conversations, they are intrinsically different from the private conversations recorded in corpora like the London-Lund, BNC, Santa Barbara Corpus and VOICE. Unfortunately, for copyright reasons, COCA can only be searched online. The text of the corpus cannot be downloaded.

COCA was designed to be a general corpus that just happens to be compiled from texts that are available online. As mentioned in section 1.1.1, the list of 5,000 words created from COCA still seems to list words in an order that would not fit most beginning EFL syllabi or graded reader schema.

Davies' Corpus of Global Web-Based English (GloWbE) is composed of 1.9 billion words from 1.8 million web pages in 20 different English-speaking countries. The corpus was created by Mark Davies of Brigham Young University, and it was released in April 2013. In this case, the corpus could be said to be representative of the English speaking portion of the World Wide Web.

Jakubíček, Kilgariff, Kovář, Rychlý, and Suchomel (2013) describe a family of TenTen corpora they have created, such as enTenTen12, by amassing 10 billion words (10^{10} , hence the name) from web pages in a particular language in a particular year. The

enTenTen12 corpus is thus 10 billion words from English-language web pages compiled during 2012. Again, such a corpus could be said to represent the Web, but how it relates to the language as a whole is uncertain.

2.5.15 Corpora and the core. Today researchers have access to a great deal of data, but is it the right kind, or the right mix? The common sense of many early language educators was that casual conversation, i.e. face-to-face speech that is not overly constrained by its context, is the core of the language. The status of writing and the relative ease with which it can be collected is an irresistible combination. Collecting published written texts is also less threatening. After all, they were written and edited for publication. Most people would feel very uncomfortable with having large amounts of their speech recorded and transcribed for researchers to pore over and analyze. Even the possibility that our online communication may have been intercepted by unknown security agents has recently been cause for alarm. For these reasons, many corpora have been compiled solely of writing, or speech is given a token representation. Occasionally, speech is given a more equal footing as it has in the ICE family of corpora or in COCA, but even then casual speech is often overwhelmed by public speech. Such corpus designs may be appropriate for comparative studies of certain features of speech and writing, but such a corpus as a whole does not give an accurate picture of the core vocabulary of the language.

2.6 Corpus similarity

Kilgarriff (2001: 232) argued that “[t]here is a void at the heart of corpus linguistics” since the discipline lacks real tools for measuring the very thing it claims to study.

The terms [corpus linguistics] has used for taxonomising corpora have been unrelated to any measurement: a corpus is described as “Wall Street Journal” or “transcripts of business meetings” or “foreign learners’ essays (intermediate grade)”, but if a new corpus is to be compared with existing ones, there are no methods for quantifying how it stands in relation to them (Kilgarriff, 2001: 232).

Comparing texts to determine how similar they are is central to stylometry, a field related to corpus linguistics. One currently popular and easily deployed method, Burrows’ Delta, will be demonstrated in 2.6.1. In his review of methods used in corpus linguistics for determining corpus similarity, Kilgarriff (2001) found chi-squared to be the most accurate. Chi-squared will be evaluated in 2.6.2. This is followed by an attempt by the author to improve on these

methods in 2.6.3. In the end, none of these methods was found to be appropriate for the current study.

2.6.1 Stylometry. There is a history of measuring the similarity of texts in statistical stylistics (or stylometry) that dates back at least to the nineteenth century, if not earlier (Lutosławski, 1890; Mosteller & Wallace, 1964). Stylometry is usually concerned with identifying authors of disputed texts or analysing differences between the known works of an author. Most techniques examine high frequency function words since they are less likely to be influenced by topic and are not typically under the conscious control of the author.

One currently popular method, Burrows' (2002, 2007) Delta, typically analyses the most frequent 150 words, most of which are function words. In an author attribution study using Delta, a group of texts is first collected to serve as a comparison corpus, e.g. a group of similar texts from the same time period as the disputed text. For example if one were trying to determine if Shakespeare might be the author of a recently discovered poem, one would compile a collection of poems from Shakespeare and several of his contemporaries. As a demonstration of this method that is closer to the concerns of this thesis, a 10,000-word sample of reportage has been selected from each national corpus in the International Corpus of English. An additional excerpt of 10,000 words of reportage has been selected from the BNC to be the test sample. In this test, Delta is being used in an attempt to identify the national origin of the BNC sample. In the case of the anonymous poem, Delta should show whether the poem is most similar to other poems by Shakespeare or possibly some other author.

The frequency for each of the individual word types in each of the national samples and the sample from the BNC has been calculated using Nation's *Range* program. The frequency for each of the individual word types is written as $f_i(D)$ in the formula for Burrows' Delta. The mean μ_i and standard deviation σ_i were then calculated for each of the most frequent 150 types (see Table 2.2). The types used to calculate Delta did not include any that have variant spellings, so the results are not based on simple orthographic differences. This collection of ten samples forms the comparison corpus. Finding the mean for this collection will allow us to calculate how far each word in each sample deviates from the mean for that word.

Table 2.2

A partial list of the frequency, standard deviation, and mean for the types in each of 10 samples of reportage from the International Corpus of English.

type	IRL	EA	GB	IND	CA	NZ	HK	JA	PHIL	SING	standard deviation	mean
the	584	865	670	793	663	596	787	784	889	611	112.95	724.2
to	286	353	321	297	303	286	332	247	295	278	29.817	299.8
of	237	305	327	284	271	222	327	282	291	248	35.744	279.4
and	224	220	166	221	235	244	211	264	223	213	25.309	222.1
a	277	140	238	173	223	222	173	213	158	191	41.435	200.8
in	204	199	188	220	209	184	184	205	184	167	15.672	194.4

How far each word deviates from the mean must be considered relative to its standard deviation. If two words have the same-sized deviation from the mean, but one word has a much smaller standard deviation, then the deviation will be more significant for the word with the smaller standard deviation. The z-score calculates the relative size of this deviation. In the formula below, the z-score is written as $z(f_i(D))$. One finds the z-score for each type in each corpus by subtracting a type's frequency in a corpus from the mean for all 10 corpora, divided by the standard deviation.

$$z(f_i(D)) = \frac{f_i(D) - \mu_i}{\sigma_i}$$

Table 2.3

A partial list of the z-scores (rounded) for the types in each of 10 samples of reportage from the International Corpus of English.

type	IRL	EA	GB	IND	CA	NZ	HK	JA	PHIL	SING
the	-1.24	1.25	-0.48	0.61	-0.54	-1.13	0.56	0.53	1.46	-1.00
to	-0.46	1.78	0.71	-0.09	0.11	-0.46	1.08	-1.77	-0.16	-0.73
of	-1.19	0.72	1.33	0.13	-0.24	-1.61	1.33	0.07	0.32	-0.88
and	0.08	-0.08	-2.22	-0.04	0.51	0.87	-0.44	1.66	0.04	-0.36
a	1.84	-1.47	0.90	-0.67	0.54	0.51	-0.67	0.29	-1.03	-0.24
in	0.61	0.29	-0.41	1.63	0.93	-0.66	-0.66	0.68	-0.66	-1.75

To find the Delta score for a sample, the z-scores for all the word types in that sample need to be combined into a single score. To do this one calculates the mean of the absolute z-

scores. The term *absolute* means all the z-scores are treated as positive, disregarding whether they are actually positive or negative. If one were to average the z-scores for the all 150 of the words as they are, the positive and negative deviations from the mean would balance each other to some degree. This would give a false impression that the text is more similar to the mean than it really is, since it often deviates in both directions from the mean. How far each word deviates from the mean is more important than whether it was more or less frequent than the mean. The mean is then calculated for the absolute differences for the most frequent 150 word types in each sample. This is the Delta value for that sample.

The same calculation is then made for the test sample. In this case, the frequency of each of the same 150 word types in the BNC sample is subtracted from the mean frequency for the same word type in the ICE collection and divided by the standard deviation for that word type. The average of the absolute z-scores for the BNC sample is then calculated to find its Delta score.

Table 2.4

A partial frequency and z-score list for a 10,000-word sample of reportage from the BNC.

type	frequency	z-score
the	785	0.54
to	261	-1.30
of	341	1.72
and	213	-0.36
a	213	0.29
in	243	3.10
.....		

In this example, Delta correctly indicates that the BNC sample most closely resembles the sample from ICE-Great Britain. The next closest sample in absolute terms is Hong Kong. (It would be interesting to investigate how many of the news writers and editors of the English-language newspapers at the time in Hong Kong were British or attended universities in the UK.) Most of the Commonwealth countries are somewhat more distant, followed closely by Ireland. Delta for the samples from the Philippines, Jamaica, and Canada are the most distant.

Table 2.5

Delta scores for reportage samples from the BNC and ICE including each ICE sample's absolute distance from the BNC sample.

	BNC	GB	HK	EA	IND	NZ	SING	IRL	PHIL	JA	CA
Delta	0.807	0.798	0.833	0.87	0.743	0.741	0.874	0.878	0.703	0.667	0.609
Absolute Distance from BNC		0.009	0.026	0.062	0.064	0.066	0.067	0.071	0.104	0.141	0.198

Figure 2.3

Delta scores for reportage samples from the BNC and ICE.

Hoover (2004a) shows that Burrows' Delta is effective for the top 600 word types, and even up to 4,000 with longer texts. Delta essentially treats all of the words it compares as equally important, and so, does not take into account the huge differences in frequencies caused by Zipf's law (Zipf, 1949). Though the results from Delta are intriguing, it may not be sensitive enough for this study.

2.6.2 Chi-squared. Kilgariff (2001) compared chi-squared, the Spearman rank correlation coefficient and three variations of perplexity, an information theoretic measure. He found that chi-squared performed significantly better than Spearman, and that these two methods easily outperformed the perplexity measures, despite that fact that "the perplexity measures required far more computation: ca. 12 hours on a Sun, as opposed to around a minute" (Kilgariff, 2001: 257).

The chi-squared statistic has been used in corpus linguistics to find English words that are distinctly American or British by comparing word frequencies in the Brown and LOB corpora (Hofland & Johansson, 1982; Leech & Fallon, 1992). Rayson, Leech and Hodges (1997) used chi-squared to analyze vocabulary variation in the spoken demographic component of the BNC based on speakers' gender, age and socio-economic background.

The formula for computing chi-squared χ^2 is quite simple. In essence, chi-squared compares how often each word occurs in one corpus with the average frequency of that word for both corpora. In comparing two corpora, one looks at a frequency list for each corpus and subtracts the observed frequency of each word, written o in the formula, from the expected frequency of that word, written e . The result is then squared before it is divided by the expected frequency. Before going any further, we need to understand the expected frequency.

$$\chi^2 = \sum \frac{(o - e)^2}{e}$$

Calculating the expected frequency for chi-squared is also very straightforward. The formula for calculating the expected frequency is shown below. The expected frequency for one of the words in corpus 1 is written $e_{w,1}$. N_1 is the total number of running words in corpus 1. N_2 is the total number of running words in a corpus 2. In corpus linguistics, it is best to keep sample sizes the same. Otherwise one cannot be sure whether the differences are intrinsic to the samples or simply the result of the size difference. Let us assume, for example, that each corpus contains 1 million running words. The observed frequency for one of the words in corpus one $o_{w,1}$ is added to the observed frequency for that word in corpus 2 $o_{w,2}$ and the sum is multiplied by the total number of words in corpus 1. This sum is then divided by the sum of the number of words in both corpus 1 and corpus 2. The formula is set up in this way to allow for the possibility of different-sized corpora. The expected frequency is thus a weighted ratio.

$$e_{w,1} = \frac{N_1 \times (o_{w,1} + o_{w,2})}{N_1 + N_2}$$

Three example words will provide an insight into how chi-squared can be used to compare corpora. For a word that occurs only once in one of the corpora, the expected frequency is 1 million times the sum of 1 plus 0, divided by 2 million.

$$e_{w,1} = \frac{1,000,000 \times (1 + 0)}{1,000,000 + 1,000,000}$$

The expected frequency for this word is, therefore, 0.5. Now that the expected frequency is known, we return to the first formula and subtract the observed frequency 1 from the expected frequency 0.5 and square the result to get 0.25. This is then divided by that same expected frequency, which yields a chi-squared figure for that word of 0.5.

$$\chi^2 = \frac{(1 - 0.5)^2}{0.5}$$

Imagine now, a medium frequency word that occurs 110 times in the first corpus and only 100 times in the second. The expected frequency for this word would be 105.

$$e_{w,1} = \frac{1,000,000 \times (110 + 100)}{1,000,000 + 1,000,000}$$

The difference between the observed and expected frequency is 5, which is then squared to get 25. Twenty-five is then divided by the expected frequency of 105 to yield a chi-squared score of 0.2381 for this word.

$$\chi^2 = \frac{(110 - 105)^2}{105}$$

A third example is a higher frequency word which occurs 11,000 times in the first corpus and 10,000 times in the second. Though this difference is proportional to the previous example, the resulting chi-squared value is very different. The expected frequency is 10,500.

$$e_{w,1} = \frac{1,000,000 \times (110,000 + 10,000)}{1,000,000 + 1,000,000}$$

The difference of 500 squared is 2,500. Dividing by the expected frequency of 10,500 yields a much larger chi-squared value of 23.8095.

$$\chi^2 = \frac{(11,000 - 10,500)^2}{10,500}$$

The sum symbol $\sum$ indicates one should sum up the chi-squared values for all of the words in the two corpora to find the overall distance between the two corpora. Imagining – very unrealistically – that these are the only words in these two corpora, one would find that they have a chi-squared distance of 24.55.

	Chi-squared
1 vs. 0	0.50
110 vs. 105	0.24
11,000 vs. 10,500	23.81
Sum	24.55

As one can see, the higher frequency word contributes far more to the overall chi-squared value even though its difference is not proportionally larger. In this way, chi-squared overemphasizes common words.

2.6.3 Standard Lexical Difference. Given Zipf's law, there will always be a few very frequent words which will dominate the results of chi-squared. Burrows' Delta, on the other hand, completely ignores all but the very most frequent words. Over the course of a year, and after discussions with Prof. Estate Khmaladze, Dr. Yuichi Hirose, Dr. Giorgi Kvizhinadze and Dr. Haizhen Wu of the School of Mathematics at Victoria University, the author sought to find a method for comparing corpora that treated the whole range of frequencies more even-handedly. Prof. Khmaladze suggested graphically comparing differences in the logarithms of the frequencies of the words in two corpora. However, Prof. Khmaladze, his colleagues and students did not have input into the construction of the formula shown here. Any mathematical naiveté is the responsibility of the author.

The method is called Standard Lexical Difference since it is based on the formula for the standard deviation. It also seeks to compensate for the skewed distribution of frequencies caused by Zipf's law. The SLD methodology yielded promising results, but is rather complicated to implement, besides being unproven statistically. The standard deviation assumes a normal distribution. However, the natural distribution of word frequencies is anything but normal. Using the logarithm of the frequencies perhaps makes this somewhat defensible. When multiple texts are combined in corpora, one begins to see a variety of power law distribution known as a double Pareto distribution (Ferrer-i-Cancho & Solé, 2001b; Montemurro, 2001). The body of a log-normal distribution and such a double power-law distribution are very similar (Giesen, Zimmerman, & Suedekum, 2010; Mitzenmacher, 2003a). In fact, Carroll (1972: xxi) was adamant that word frequency distributions are log-normal rather than Zipfian.

Standard Lexical Difference begins with the familiar formula for standard deviation. The innovation is in how the value for each word type is calculated prior to finding the standard deviation. The second formula, which is used to calculate the value difference for each word type has two parts. These two parts function much like the bicameral legislature in the United States. The US congress has two houses. In the Senate, each state is represented equally with each state having two Senators. In the House of Representatives each state is allotted a number of congressmen based on the population of the state. This system was designed to protect the rights of small states while guarding against tyranny by a few. Corpus linguists face a similar dilemma since corpora always contain a few very frequent words and many that are infrequent.

Value of difference for each individual type

Average value of differences for all types

Sum of calculations for all types from 1 to however many there are

$$\sqrt{\sum_{i=1}^n \frac{(v_i - \bar{v})^2}{n}}$$

Number of values in comparison

Frequency of each individual type in corpus a and corpus b

$$v_i = (\ln a_i - \ln b_i) * \frac{(a_i + b_i)/2}{N}$$

Natural logarithm of frequency of each individual type in corpus a and b

Total number of tokens in the text.

The first part of the formula subtracts the logarithm of the frequency of a word in one corpus $\ln a_i$ from the logarithm of the frequency of that word in the second corpus $\ln b_i$. Logarithms are great equalizers. The natural logarithm of 100 is 4.61, while the natural logarithm for 1,000 is 6.91, a difference of only 2.3. The logarithm reduces differences of hundreds of thousands to only a dozen or so.

The second half of the formula determines how much influence this difference in frequency should have by calculating the average proportion of the corpus this word type represents. For each word type, the actual frequency in the first corpus a_i is added to its actual frequency in the second corpus b_i . The average is then divided by the total number of tokens

in the corpora. Here again, the corpora are assumed to be the same size. The value v_i is the proportional logarithmic difference between the corpora for each word type.

The value $\bar{v}$ is the average of these proportional logarithmic differences for all the word types that occur in these two corpora. The standard deviation formula then calculates how broadly the majority of these differences are scattered in the corpus. This standard deviation of the proportional logarithmic differences is the Standard Lexical Difference between the two corpora.

A group of six corpora were compared using the SLD statistic. Three of the corpora are of face-to-face conversation, one of TV and movie subtitles, another of writing for children and adolescents, and a final one of novels. Each corpus contained approximately 1 million words. The results are intuitively valid with the lower middle-class (C1) and upper working class (C2) conversation corpora being most similar followed by a mixed corpus of lower working class (DE) and New Zealand conversation. Next was the simulated dialogue of TV and film, children's and adolescent literature, and finally novels.

Table 2.6

A matrix showing the Standard Lexical Difference scores for a group of sample spoken and written corpora.

Cross comparison: <i>Standard Lexical Difference</i> > SLD value = > dissimilarity Each corpus contains approx. 1 million words	BNC C1 conversation	BNC C2 conversation	WLG Spoken & BNC D-E conversation	TV and movies	NZ school journal & children's books	BNC Novels
BNC C1 conversation		0.068	0.163	0.236	0.453	0.545
BNC C2 conversation	0.068		0.160	0.254	0.572	0.567
Wellington Spoken & BNC D-E conversation	0.163	0.160		0.247	0.476	0.482
TV and movies	0.236	0.254	0.247		0.420	0.350
NZ school journal & children's books	0.453	0.572	0.476	0.420		0.197
BNC Novels	0.545	0.567	0.482	0.350	0.197	

Though these results are encouraging, it was realized that none of these three measures, Burrows' Delta, chi-squared or Standard Lexical Difference actually fits the needs of these studies. Further investigation of the SLD statistic was, therefore, put off till a later date. Since

the goal of these studies is a methodology for finding core vocabulary lists, what is needed is a measure which quantifies exactly how many words from a particular word list are shared by another word list. This is actually a very simple question related to set theory that will be explored in the next chapter.

2.7 Applications of core vocabulary lists

Some important applications for vocabulary lists have been seen repeatedly in this survey.

1) Prioritizing vocabulary teaching and learning in a foreign language

Probably the foremost application of core vocabulary lists is the prioritizing of vocabulary teaching and learning in foreign languages. The vocabulary of native speakers of a language increases by around 1,000 words per year (Nation, 1990: 11). All, or nearly all, these new words are learned incidentally through listening or reading. However, learning vocabulary incidentally requires extensive amounts of reading and / or listening (Nagy, Herman & Andersen, 1985; Nagy, 1997). Liu and Nation (1985) also found that learners needed to know at least 3,000 word families before learners they could successfully engage in extensive reading. It makes sense then to emphasize the direct learning of the most frequent words in order to get beginning FL learners over this threshold and on their way to a steady acquisition of the language. An accurate core vocabulary list would, therefore, be of considerable benefit.

2) Calculating the vocabulary load of a text

With accurate estimates of the frequencies of words, one can calculate how much of a text will likely be covered at a given vocabulary size (Nation & Beglar, 2007). Depending on what one deems to be an adequate level of comprehension, one can determine the approximate vocabulary size that will be necessary to achieve that level of comprehension. Different types of texts can have very different vocabulary loads. As was mentioned earlier, 6,000 words are required for comprehension of unsimplified conversation, whereas 8,000-9,000 words are needed to read ungraded narrative or general written texts with ease (Nation, 2006).

3) Determining a learner's vocabulary size

Knowing a learner's vocabulary size allows one to then match learning materials to a learner's level. The accuracy of a vocabulary size test depends largely on the quality of the list

it is based on. The Eurocentres Vocabulary Size Test 10ka (Meara & Jones, 1987, 1990) was based on Thorndike and Lorge's (1944) list. Schmitt, Schmitt and Clapham's Vocabulary Level's Test (2001: 58) was created using "stratified sampling from the Thorndike and Lorge list, with reference to frequency data from Kučera and Francis and the GSL." Their Vocabulary Levels Test includes subtests for the most frequent 1, 2, 3, 5 and 10,000 words plus a subtest for the Academic Word List (Coxhead, 2000).

Nation and Beglar's Vocabulary Size Test (2007) is based on an analysis of the much larger British National Corpus. The larger corpus allowed them to create an extended and more finely grained test. The Vocabulary Size Test gives estimates for the most frequent 14,000 word families for each 1,000-word level. There is also had a telling difference between Nation's (2004) BNC vocabulary lists and the underlying list used to create the Vocabulary Size Test. The formal written character of most of the texts in the BNC had already been shown to adversely affect the results of the first 3,000 words in Nation's (2004) BNC lists. For that reason, Nation and Beglar elected to order "the first twelve 1000 word lists were revised using word family range and frequency figures from only the 10 million token spoken section of the British National Corpus" (2007: 10).

4) Providing accurate estimates of word familiarity for psychology research

Research into human memory and cognition often relies on estimates of word familiarity from vocabulary lists. Unreliable lists weaken the results of this research. This is the specific motivation behind the corpora and lists created by Howes (1966), Burgess and Livesay (1998) and Brysbaert and New (2009).

2.8 Conclusion to the background review

In summary, this chapter has discussed the difficulties involved in defining the concept of word and the inevitable inconsistencies that have to be accepted when using the orthographic word as an operational definition. Corpus linguists have striven to construct representative corpora for more than a hundred years, but have either lacked sufficient data or an adequate framework for categorizing the texts of the language. A taxonomy for the texts of the English language that is based in an objective analysis of linguistic features can be found in Biber's (1995) text types. Chapter three will propose a practical application of Biber's text types to the problem of creating core vocabulary lists. There is also a gap in the current research in regards to how different the vocabulary of these text types are.

Immense quantities of corpus data and processing power are available to corpus linguists today, but the amount of variability that is associated with a given amount of data has not been quantified. A practical measure for quantifying that variability and the distance between text types will be described in the next chapter.

Many educators have suspected that spoken language, especially casual face-to-face conversation is the key to finding the core vocabulary of a language. However, conversation has typically been downplayed or ignored. This is likely the reason so many researchers have resorted to intuitive methods to create or “fix” their core vocabulary lists. As Nation and Beglar (2007) discovered, reordering results based on spoken data gives intuitively more acceptable results. The hypothesis that conversation is the foundation of a *prima facie* core vocabulary list will be checked by comparing an objectively compiled list of high-frequency words from conversation with some of the intuitive lists that have been described in this chapter. That the primary form of language – face-to-face conversation – has been neglected in the creation of core vocabulary lists is a significant problem since these core vocabulary lists have far reaching impacts on language research, teaching, materials production and testing.

Various-sized units for counting words have been recommended to learners or used to create core vocabulary lists, but researchers have not determined how significantly this impacts the text coverage or the quantity of word forms that would need to be learned. The final study in this thesis will explore that question.

CHAPTER 3

METHODOLOGY

3.1 Text Types and a Taxonomy of English Language Texts

3.1.1 Out of order.

Of the five problems with objectively derived frequency lists identified by Nation (1990: 21-24), it was seen in section 1.1.1 that problems one, two and four are really the same problem. Certain words that should be among the most frequent items, like the common words for food, clothing and the classroom that Fries and Traver (1960: 40-41) mention, often do not appear among the highest frequency words (problem 1), while other words that are not suitable, do appear (problem 2). In other words, the words are in the wrong order for instruction (problem 4). This problem with the ordering of words is the single biggest failing of objective vocabulary studies.

This research project originally started as a prelude to investigating other research questions related to vocabulary learning and reading. Do marginal or in-line glosses better facilitate fluent reading and comprehension? Do glosses in the reader's L1 or L2 better facilitate engagement and vocabulary learning? A core vocabulary list was needed for this research project since high and low frequency words should be handled differently (Nation, 1983). Not being satisfied with any of the available lists, I made a rather naïve attempt to find a new core vocabulary list by cobbling together a very large corpus from many existing corpora of spoken and written English. This was based on the hope that if one included large quantities of data from multiple varieties and balanced the corpus so that it contained approximately 50% spoken and 50% written English, a useful core vocabulary would rise to the surface as the most frequent and most widely dispersed items .

In retrospect, it is now clear that this method is similar to taking a serving of each dish on tonight's menu at *Logan Brown* and tossing them into a blender. No one would expect a few spoonfuls off the top of this mixture to give an accurate introduction to the food served at the restaurant. A mass spectrometer could identify many of the ingredients, but the results would be far from palatable. Similarly, the list that was derived using this blender methodology was not pedagogically useful. The resulting list resembled many of the lists seen in chapter two. Many obviously necessary words were far down the list or very poorly dispersed. Casual and academic items were jumbled together without any way to distinguish them in a principled manner.

Frequency lists can only reflect the source corpora they are derived from, so mixing the types of texts in the corpus inevitably produces mixed results. The question then becomes, how does one divide a language in a principled and consistent manner? This is why a taxonomy of texts in the language is needed to guide the analysis.

3.1.2 A simplification of Biber's text type categories

Biber's multi-dimensional analysis produced a set of text types for English. These text types provide a framework that is based in the linguistic characteristics of texts rather than conventional labels applied to the contexts in which they are found. The traditional taxonomy of genre and registers is not based on any measurement of similarities or differences. In the Brown corpus, mystery novels and adventure novels are distinguished as separate genre, but how different are they, and in what ways? Are science fiction novels more similar to these other two genre of novels or to some other type of text, popular scientific writing, for example?

Figures 3.1 and 3.2 show Biber's eight text types that will serve as a map for classifying the texts used in this study to represent each text type. The figures show the typical scores for each text type on each of the five dimensions. These figures also indicate some possible simplifications to this text type taxonomy. This simplified model of the language consists of four key text types. Data has been collected from existing corpora to match the texts Biber (1995) found to be characteristic of each of these key text types. Rather than forming a single corpus, each of these four collections represents one of the four key text types. The studies in this thesis investigate each of these text types individually and compare the results. It is an important point that the four collections are not combined to form one mega-corpus. The mixing of different text types in a single corpus is at the root of many of the problems with previous core vocabulary lists.

Figure 3.1

Biber's (1995) text types 1-4 showing the average score on each of five dimensions. (Identical to Figure 2.1).

Figure 3.2

Biber (1995) text types 5-8 showing the average score on each of five dimensions. (Identical to Figure 2.2).

3.1.2.1 Interactive. The dimensional profiles of text types 1 and 2 are almost identical except that texts of type 1 have a somewhat higher score on dimension one. Text type 1

consists of interactions that are “situated in reference, non-abstract in style, and extremely interpersonal and involved in focus” (Biber, 1995: 333). Texts of type 2 are still focused on maintaining social relationships, but are slightly more informational in focus (Biber, 1995: 333-334).

All of the texts in text type 1 are face-to-face conversations or telephone calls (Biber, 1995: 328). No written texts are found in the first text type. The vast majority of texts in type 2 are still face-to-face conversations, though some personal letters, one piece of business correspondence, and one text of general fiction fall into text type 2 as well. Given the similarity of these two text types and the fact that face-to-face conversations make up the majority of both text types, it makes sense to combine these two text types into a single “interactive” type. This type of discourse typically serves both to maintain and make use of social networks.

Spoken conversation texts have been collected from a variety of corpora (Appendix A.1) to represent this text type. A determined effort was made not to allow informational interactions to overwhelm interpersonal ones by limiting the inclusion of formal contexts as much as was practical. Despite its linguistic importance, data for face-to-face conversation are the hardest to come by and the most diverse in format. This text type will be referred to as “conversation” in the remainder of the thesis.

3.1.2.2 General reported exposition. The most common text type in Biber’s study – the category into which most texts fell – is text type 6 “general reported exposition.” Texts of type 6 have nearly neutral scores on all five dimensions (Biber, 1995: 339). They include a wide range of written texts, as well as a several prepared spoken texts. To represent text type 6 a collection of non-academic writing and reportage was compiled that matches the text categories best represented in this text type in Biber’s study. These include texts that were classified as biographies, editorials, essays, humor, instructional writing, persuasive writing, popular lore, popular writing and reportage (Appendix A.3).

The above collection does not include any spoken texts, despite the fact that, in Biber’s study, this type also included some prepared speeches. It was felt that these written-to-be-spoken texts were classified into the general text type primarily since they were prepared in advance. The category “prepared speeches” really represents a middle ground between speech and writing. This study seeks to investigate the four most distinct text types seen in Figures 3.1 and 3.2. Studies of more finely nuanced categories have been left for a future study. In this thesis, the text type “general reported exposition” will be referred to simply as “general writing.”

3.1.2.3 Imaginative narrative. Texts of type 5 “imaginative narrative” have slightly higher scores on all dimensions with the highest – not surprisingly – being the narrative dimension. More than three-quarters of the texts that fit into this category are from some genre of fiction (Biber, 1995: 330). Seven of the 14 prepared speeches in Biber’s study were also members of this text type, but in keeping with same principle mentioned above, spoken and written texts were not mixed in this study. A collection made up of texts in the categories “novels and short stories,” “fiction and drama,” “general fiction,” “mystery and detective fiction,” “science fiction,” “western and adventure fiction,” and “romantic fiction” will represent text type 5 in this study. The imaginative narrative text type will be referred to as “narrative writing” (Appendix A.2).

3.1.2.4 Academic. Biber found a significant difference between the academic writing of scholars in the natural sciences and those in the humanities. His study divided these into text type 3 “scientific exposition” and text type 4 “learned exposition.” A far more abstract style is preferred in the natural sciences, whereas scholars in the humanities apparently do not usually feel the need to distance themselves quite so markedly from their research (Biber, 1995: 334). Though there is undoubtedly a difference in the vocabulary of the humanities and that of the natural sciences, that distinction will be saved for a future study. This thesis will treat texts from both the humanities and natural sciences as a single “academic” category. Academic prose texts were by far the most common texts in these two text types. No spoken texts occurred in either. A collection of texts labeled “humanities,” “social sciences,” “natural sciences,” and “technology,” “written academic” and “science / academic prose” will be used to represent the academic text type (Appendix A.4).

3.1.2.5 Other text types. For this study, it is practical to ignore types 7 and 8. Type 7 “situated on-line reportage” is “the smallest and most distinct text type identified in [Biber’s] study” (Biber, 1995: 340). It is primarily the highly situated speech of live sports commentary (Biber, 1995: 331). Though interesting, this text type is both rare and narrow in focus.

According to Biber, text type 8 “involved persuasion” is also “a relatively specialized text type” (Biber, 1995: 341). This text type is spread across 14 different spoken and written genres, but is not the center of gravity for any of them (Biber, 1995: 331). One could think of it as a discourse space that a speaker or writer can resort to when needed, whether it be in a phone call to a colleague or in an academic journal article. Since it is so widely dispersed, it becomes impractical to isolate. The two most common genres in this type are interviews and spontaneous speeches. However, more interviews and spontaneous speeches are classified as “informational interaction” than as “involved persuasion” (Biber, 1995: 228-331).

3.1.2.6 Key text types. This leaves four key text types to investigate: conversation, as well as general, narrative and academic writing. An almost identical reduction to these four text types is seen in the *Longman Grammar of Spoken and Written English* (Biber, Johansson, Leech, Conrad, & Finegan, 1999: 16) and the *Longman Student Grammar of Spoken and Written English* (Biber, Conrad, & Leech, 2002: 23). The total quantity of data available for each text type varied considerably (Table 3.1). Since the majority of texts are members of the text type “general writing,” it was the easiest to collect. There were nearly seven times as many tokens of general writing as there were for conversation. The collections of narrative and academic writing were mid-way between the sizes of the other two text types.

Table 3.1

Corpus sizes for the key text types compiled for these studies

Text type	N = * million
Conversation	7.05
Narrative Writing	18.42
General Writing	47.65
Academic Writing	17.91

3.2 Data preparation

As described above, data has been collected for these studies from many different corpora. The different corpora from which the data was collected have differing transcription and mark up conventions that would affect the comparison of the vocabulary they contain. Before the comparisons could begin, all of the data needed to be in a uniform format that would not influence the results. R-scripts had to be written to clean each of the corpora individually so that the resulting texts were consistent in form (c.f. Appendix B.1.1-2). Some corpora included glosses of non-English vocabulary or non-standard usages. For example, in one of the samples from VOICE, the word *U-Bahn* is glossed as *underground*. The word *underground* was not actually uttered in the conversation.

```
how do you get home from the city center do you take a <LNger> u
bahn? {underground} </LNger> or a tram . . .
```

These glosses and other comments or annotations were eliminated, so as not to inflate the frequencies for those words. Some of the less-straightforward issues and the rationale behind each solution are described below. Most of the data manipulation and analysis presented in

this thesis was performed using scripts written in the statistical programming language R (R Core Team, 2012) and executed using either the 32-bit or the 64-bit R packages in a Windows 7.0 operating environment.

3.2.1 Tags, annotations, etc. The corpus data for each of the four key text types was stripped of tags, annotations, line numbers etc. using a series of R-scripts written to fit each of the various corpora (Appendix B.1). The scripts in Appendix B.1 are only a sample, as a different script had to be written for almost every corpus and often for different sections of the same corpus. The corpus files then had to be manually checked for unexpected tags or typographic errors that allowed annotations to remain in the text. Multiple attempts were usually required to clean each text type of each corpus. This meticulous work was warranted since the corpus data forms the foundation of all of the studies in the thesis. It was vitally important to remove the varying tags from the differing corpora. Any remaining tags or fragments would have been counted as words and corrupted the resulting frequency lists.

With another script, the resulting text files were then cut into 10,000-token bricks that could be randomly grouped into corpora of various sizes (Appendix B.2). Texts were cut at exactly 10,000 tokens, even if that meant dividing a sentence or published text. Corpus compilers often collect samples of 2,000 tokens, but cut off a sample at the end of the next sentence. The first study seeks to measure the variability in word types between samples. A precise number of words was used to insure that any variability was due to variation in which words were included not how many.

3.2.2 Non-ASCII characters. The corpora used in this study differ considerably in how they represent non-ASCII characters, such as letters with a grave or acute accent, umlaut or tilde. Each of the varying conventions was converted to the appropriate Unicode character, otherwise each of the encoding schemes would have created unique word forms. For example, the word *fiancé* would be encoded as *fiancé* in many of the ICE corpora, but *fianc&/e* in ICE-Ireland. If the ampersand and slash were simply deleted this would create two differing forms *fiance* and *fianceacute* even though they are actually the same word. As one of the last steps in cleaning a text, the various scripts replaced punctuation marks with a space so as not to accidentally join two nearby words. If the codes were not first replaced with Unicode characters, the resulting fragments would have created a number of new and irregular word types.

3.2.3 Hyphenated compounds. All hyphenated compounds are treated as separate words, since it would have been impractical to link hyphenated compounds with their non-hyphenated variants. For example, the compound *aircrew* can also be written as *air-crew* or as *air crew*, whereas other hyphenated forms are not combined as a single word, e.g. *father-*

in-law. Hyphens were, therefore, globally replaced with spaces. The line-break hyphen is marked in corpora by a tag – usually some variation of a bracketed lowercase letter *L*, e.g. `con<l>sistently`. Line-break hyphens were removed reuniting the word segments.

3.2.4 Letter-digit combinations. In the corpora of spoken English, numbers are always written out as words. Corpora of written texts contain digits that could be considered words or are parts of words. Strings, such as *A4*, *37th*, *3D*, *MP3*, *M25*, that contain both digits and alphabetic characters have been counted as types. One could argue that cardinal endings like *-st*, *-th*, *-nd* are really instances of one type, but the same could be said of all the inflectional morphemes. This could also be seen as compensating for the loss of certain digit-only items that could also be considered word types, e.g. *747*. Hyphenated items like *I-75*, *F-16* or *C-130* are counted simply as an instance of the letter since the digits are left stranded (and are deleted) after the hyphen is removed. The decision to keep letter-digit combinations was an effort to err on the conservative side so as not to underestimate the number of word types. In the end analysis, this was probably unnecessary and slowed the processing of the corpora of general writing considerably. General writing, it turns out, contains a large number of measurements, e.g. *3mm* or *15in* where there is no intervening space between the number and the unit of measure. In the analysis of these texts, each of these combinations was counted as a new word type.

3.2.5 “Indigenous” vocabulary. Since as Kirkpatrick (2007: 1) points out, most English speakers today are multilingual, the line between English and other languages is often fuzzy. Languages and the communities that speak them are not discrete entities. Rather, they are complex, overlapping networks. Most corpora of spoken English contain some quantity of words or utterances that are arguably not English. In some situations, these are clearly cases in which the speaker is consciously code-switching, e.g. this example from ICE Hong Kong:

<code><ICE-HK:S1A-009#147:1:A></code>	Uh huh
<code><ICE-HK:S1A-009#148:1:C></code>	<code><indig> Sihk daan gou </indig> <&></code> translation: eat cake <code></&></code>
<code><ICE-HK:S1A-009#149:1:A></code>	Thank you

The Santa Barbara Corpus contains conversations in which speakers switch between Spanish and English, and VOICE is a polyglot of European languages. However, in many instances, supposedly non-English words are individual lexical items that are naturally part of English usage, at least in that variety. For example, in ICE East Africa, the word *chapattis* is tagged as East African, despite the fact that the term is not African in origin and that English

speakers around the world use it. If one were to remove items like this, *cheddar* would also have to be removed from the UK corpus because it should be considered a British word. Removing local vocabulary would often mean taking away essential casual vocabulary. Words connected to food are very frequent in casual conversation. (Their absence was one of the problems identified with many frequency lists.) Since one would typically refer to the cuisine that is in one's environment, food-related vocabulary would logically be local as well.

Working with the New Zealand, Jamaican and East African corpora, it seemed reasonable to argue that indigenous words are an essential part of these English varieties and that all words tagged as indigenous should be left in the data. The Philippine corpus challenges that principle repeatedly, though. ICE Philippines contains extensive use of Tagalog – much more than one finds examples of indigenous languages in the other ICE corpora. Lexical items are very often thoroughly integrated into the English usage. These are usually discourse markers like *talaga* “really” or *oo* “yes.” Sometimes, however, the code switches are three or four turns in length. There are also numerous examples of a variety referred to as Taglish. In Taglish, the speaker creatively fuses both Tagalog and English in such a way that it is sometimes difficult to say which language is dominant.

<ICE-PHI:S1A-002#39:1:B> Yeah nice we went to Harvard it was so
cold and ka-cheap-an namin we (laughter)
we were trying to find out where Sharon
Cuneta was living (laughter)

In the above example, *ka-* and *-an* are a pair of Tagalog morphemes that together indicate superlative. Here they are creatively placed on an English adjective. This seems to trigger a switch to Tagalog, since the speaker continues with *namin* “we,” but then quickly reverts to English “we...”

In the next example, the number of Tagalog lexical items is greater. It is difficult to say whether the utterance is Tagalog with English borrowings or English with Tagalog borrowings. The preposition *sa* marks the point where the utterance switches to Taglish. From that point, one could argue that the code is basically Tagalog, but with English proper nouns. This dual language identity is confirmed by Speaker A who responds in English and follows up with an equivalent discourse marker in Tagalog.

<ICE-PHI:S1A-036#178:1:C> I had winter clothes because of April
sa [in] Paris *hindi pala* [not actually]
November sa [in] Paris
<ICE-PHI:S1A-036#179:1:A> That's right oo [yes]

Similarly, the utterance below really switches to Tagalog after *okay*, except for the English borrowing *communicate*.

<ICE-PHI:S1A-036#194:1:C> And they know Chinese
<ICE-PHI:S1A-036#195:1:C> That's why okay *na okay kasi nakakapag-*
communicate sila

Socio-linguistically, this is a brilliant way to maintain both one's identity as a Pilipino and as an educated English speaker. A similarly sophisticated mixing of Malay and English is also common, but is not represented in this study (Kirkpatrick, 2007: 127-129). Code-mixing leaves the corpus linguist to face a difficult dilemma, however. If all items are admitted as English, the study would overestimate the diversity of vocabulary used in any individual variety of spoken English. But, to go to the other extreme, and filter out all supposedly non-English words, would create an overly homogenous collection that underestimates the diversity of vocabulary in English as an International Language, not to mention missing important new word types.

In the end, a practical compromise was used. Whenever more than two indigenous items are consecutive, those items are excluded. This allows most of the naturally embedded indigenous vocabulary to remain while excluding most true switches to another code. In the New Zealand corpora, the occasional longer phrase will unfortunately be omitted, such as *Hei konei ra* below.

<ICE-NZ:S1A-004#217:1:G> and we we um put *hei konei ra* on
<ICE-NZ:S1A-004#218:1:G> which is a farewell song see you later
we're going home

3.2.6 Observer's paradox

The data that is available for conversation does sometimes suffer from the observer's paradox. Since corpora are usually collected by academic researchers, it is not surprising that the subjects being recorded are often associated with academia and, therefore, their discussion

topics tend to be academic. Even in the BNC demographic recordings, where great pains were taken to gather the everyday conversation of a broad spectrum of society, the recording devices themselves, the Sony *Walkman*TM, often became the subject of conversation. The act of recording the conversation has often obviously altered the interactions, e.g. this excerpt from the Bergen Corpus of London Teenage Language (COLT) which was incorporated into the BNC spoken demographic section:

Do you get to keep the walkman
No
Did you ask
Swap it around
Go on swear swear in front of all those Norwegians
Piss off

In the International Corpus of English, especially in the corpora for Hong Kong and Jamaica, language seems to be a frequent the subject of conversation. Both Hong Kong and Jamaica are very interesting linguistic contexts, but this may also be because linguists were involved in the conversations being recorded.

3.3 How much data is needed to give consistent results?

Once the text types have been identified and representative collections of texts have been assembled, the first research question can be addressed: “How much data is necessary to give consistent results for high-frequency vocabulary?”

Among the articles reviewed while searching for a corpus similarity/difference measure to use in the first study was Popescu and Dinu's (2008) “Rank Distance as a Stylistic Similarity.” In this article, they mentioned that “the process of ranking makes the distance measure more robust acting as a filter, eliminating the noise contained in the values of the frequencies.” This led to the realization that a different approach was needed to answer the first research question. Though interesting, trying to directly compare corpora or even word frequencies would not really answer this question. The real question is how many word types differ when one compares the ranked frequency lists from different samples. How many words do they share, and how many are different? Or to be more specific, “What amount of variation can be expected for each of the four key text types for a certain rank level if one has a particular amount of data?” This would address two problems that Nation (2001: 15-16) identified: that word lists often disagree or “are not reliable above a certain level.”

3.3.1 The Dice Coefficient with non-overlapping samples. Inspiration for a practical method of indirectly comparing corpora by comparing the ranked frequency lists derived from them came from ecological research. Ecologists regularly need to compare populations of living things in different environments. This is often done by plotting a grid over a particular environment and comparing the quantities of certain species in each section of the grid. For example, does a particular plant species grow better in shaded areas or in direct sunlight? Population changes can also be compared from season to season or from year to year to look for trends or dynamic patterns.

Figure 3.3

Dr. Martha E. Apple laying out a grid to gather data for the GLORIA project – Global Observation Research Initiative in Alpine Environments monitoring network (EcoRover, 2010).

One useful metric for this sort of study was developed by Lee Raymond Dice as he compared biotic regions in North America (Dice, 1945).

$$\frac{2|\alpha \cap \beta|}{|\alpha| \cup |\beta|}$$

The Dice coefficient compares two quantities, which are represented in the formula as α and β . The upper part of the formula is the amount of overlap between the two quantities. The amount of overlap is multiplied by two, since this amount is present in both samples. The lower part of the formula sums the total amount in both samples. For the sake of simplicity, let us assume both samples are of equal size, for example 100, and 75 is the quantity common to both samples. The Dice coefficient would be calculated as $\frac{2|75|}{|100|+|100|}$ which would be $\frac{150}{200}$ yielding a Dice similarity of 0.75. This is a simple and intuitive quantification of the similarity between the two samples.

The Dice coefficient has also been used to quantify genetic variation and map the relatedness of species (e.g. Chou & Wang, 2006; Sanchez et. al., 2009). This application is similar to how the Dice coefficient has been used in this study to compare the vocabulary of texts or corpora. To compare two strands of DNA one aligns the sequences of molecules in each strand of DNA. One then counts how many of the nucleotides in the two sequences are identical out the total number being compared. The example below shows two versions of the beta chain in the gene for the haemoglobin protein (Clancey, 2008). The second sample contains a single mutation among the sequence of 39 nucleotides. These two sequences have a Dice similarity of 0.97. However, the small mutation in the second sample causes sickle-cell anemia. Fortunately, variations in core vocabulary lists are usually less life-threatening.

ATG	GTG	CAC	CTG	ACT	CCT	GAG	GAG	AAG	TCT	GCC	GTT	ACT
ATG	GTG	CAC	CTG	ACT	CCT	GTG	GAG	AAG	TCT	GCC	GTT	ACT

The methodology for the first study begins with the compilation of equal-sized samples of text. Since the overall goal is the development of a methodology for creating core vocabulary lists, the primary focus should be on the higher frequency word types. Therefore, the next step is to create frequency lists for each of the samples that will be compared. The lists are then sorted in descending order of frequency. Similar to the DNA comparison, the two ranked word lists are laid side by side (See Table 3.2).

Table 3.2

Excerpt from the ranked list of tokens in descending frequency order for two 500,000-token samples of face-to-face conversation. Words present in one list, but not the other are highlighted.

Sample α		Sample β	
Word	Total	Word	Total
I	18,748	I	19,471
the	14,903	you	15,888
you	14,705	it	14,541
it	13,498	the	14,228
and	11,895	's	12,520
's	11,350	and	10,629
that	9,912	that	9,936
to	9,583	to	9,367
a	8,701	a	9,040
't	7,950	't	8,212
of	6,221	yeah	7,890
yeah	6,108	in	5,586
in	5,860	of	5,166
they	5,548	they	5,048
he	5,483	he	4,826
is	4,487	oh	4,558
no	4,179	no	4,492
was	4,167	is	4,456
we	4,142	what	4,426
know	4,069	we	4,228
oh	3,965	know	4,108
what	3,878	there	3,769
so	3,768	was	3,711
but	3,727	on	3,701

In collections of written texts, the most frequent word is almost always *the*, while in casual conversational texts *I* is typically the most frequent. Even if the two frequency lists agree on a certain word's rank in relation to the other words in the sample, there will inevitably be discrepancies in frequency between the two samples. In this example, *I* is the most frequent word type in both samples, so *I* has the rank of 1 in both samples. Looking at the frequencies, however, one sees that the word *I* occurs 18,748 times in sample α , while in sample β the word *I* occurs 19,471 times. There is a discrepancy of 723 occurrences, but this minor variation in frequency is irrelevant. This is one of the advantages of using rank rather than the actual frequency in comparing words, as Popescu and Dinu (2008: 91) pointed out.

Unlike the comparison of DNA samples, the precise sequence of words on each frequency list is not essential. More and more precise matches will be seen as sample sizes increase. (With very large samples that are drawn from the same text type, e.g. more than 20 million tokens of general written exposition, there are an uncanny number of precise matches. See Appendix F.) However, users of vocabulary lists are not typically interested in whether a particular word is the 15th or the 21st most frequent word, only that it is among the most frequent 100 or 1,000 words in a certain text type. Pedagogical decisions are usually made on the basis of a word's inclusion in particular range of word frequencies, e.g. the top-ranked 500 words in spoken conversation, rather than a word's exact place on that list.

The first study begins by comparing the most frequent 100 word types in two samples and finds the Dice distance between them. (Dice distance is simply one minus the Dice similarity.) In order to keep the table reasonably small, let us calculate the Dice distance for the top-ranked 25 words for the samples in Table 3.2. Each list agrees on 23 of the 25 words, so $\frac{2|23|}{|25|+|25|}$ which is $\frac{46}{50}$ and a Dice similarity of 0.92 and a Dice distance of 0.08.

After finding the Dice distance between the most frequent 100 word types in the two samples, the range is then enlarged to the 200 most frequent words, and the Dice distance is calculated again. The range is then further increased to 300, and the Dice distance is determined for that length. In this manner, the length of the comparison is steadily enlarged in increments of 100 ranks until the comparison includes the most frequent 10,000 words in each sample.

Figure 3.4

Illustration of the comparison of increasingly large portions of two frequency lists, calculating the Dice distance for the most frequent 100 word types, 200, 300, etc.

The Dice distance scores for each length can then be plotted on a graph to show how much the two samples differ at each point as more and more of the two frequency lists are compared. Each comparison is cumulative, since each comparison includes all the previously compared word types.

Figure 3.5

Graph of an example comparison of the Dice distance for the most frequent 100, 200, 300 word types.

For the sake of completeness, another set of comparisons was also made starting with the most frequent 1,000 word types in each sample and then increasing in increments of 1,000 ranks until it reached the length of the shortest frequency list (rounded to the last full millennium). The Dice distance scores are then plotted to show how the samples diverge as one moves from the most frequent word types to the inclusion of the lower frequency word

types in each sample. As will be seen, the pattern of the divergence for each text type is distinct, and is related to the nature of the text type.

Of course, comparisons of only two samples would not be very reliable. The planned methodology, therefore, called for cross-comparisons of ten equal-sized samples. The average among these 45 comparisons can then be plotted with their standard deviations, keeping in mind Gries' admonition that one "should never -- NEVER! -- report a measure of central tendency without a corresponding measure of dispersion" (Gries, 2009d: 111). This proved problematic, however, at larger sample sizes, since the quantity of data available did not always allow for the creation of 10 non-overlapping samples.

The Dice distances for the most frequent 100 words in the example below (Table 3.3) are relatively small despite the fact that the samples are a mere 100,000 tokens each. The mean Dice distance is 0.044 with a standard deviation of 0.015, which means that each list of 100 words differs by around four words plus or minus one or two.

Table 3.3

Dice distance matrix for the top-ranked 100 word types for ten 100,000-token samples composed of data from the BNC spoken demographic section.

	100-1	100-2	100-3	100-4	100-5	100-6	100-7	100-8	100-9
100-2	0.04								
100-3	0.05	0.03							
100-4	0.06	0.05	0.06						
100-5	0.06	0.04	0.04	0.06					
100-6	0.07	0.06	0.05	0.08	0.05				
100-7	0.03	0.02	0.03	0.05	0.05	0.05			
100-8	0.03	0.03	0.03	0.06	0.04	0.07	0.04		
100-9	0.06	0.04	0.04	0.07	0.02	0.05	0.05	0.04	
100-10	0.03	0.02	0.03	0.03	0.03	0.05	0.02	0.03	0.04

For the most frequent 200 items, the Dice distance would be expected to grow, as it does. The mean Dice distance for the most frequent 200 types (Table 3.4) has risen to 0.107 (standard deviation = 0.015) which means that, on average, the list of the 200 most frequent words in each sample differs by 10 or 11 words.

Table 3.4

Dice distance matrix for the top-ranked 200 word types for ten 100,000-token samples composed of data from the BNC spoken demographic section.

	200-1	200-2	200-3	200-4	200-5	200-6	200-7	200-8	200-9
200-2	0.085								
200-3	0.085	0.105							
200-4	0.090	0.105	0.100						
200-5	0.095	0.110	0.110	0.090					
200-6	0.100	0.100	0.100	0.110	0.110				
200-7	0.120	0.125	0.120	0.115	0.120	0.130			
200-8	0.095	0.115	0.100	0.110	0.120	0.110	0.155		
200-9	0.065	0.110	0.095	0.090	0.095	0.100	0.130	0.105	
200-10	0.110	0.110	0.095	0.120	0.115	0.115	0.120	0.125	0.105

The first research question asks how much data is needed to give consistent results for high-frequency vocabulary, and the Dice coefficient measures the stability of the high-frequency word types as the sample size is steadily increased. Initially, 100,000-token samples of conversation were compared. This minimal sample size was used to give a maximally divergent baseline figure. The sample sizes were then increased in increments of 100,000 tokens. In other words, the second round compared 10 samples of 200,000 tokens each. The third round compared 10 samples of 300,000 tokens each, etc.

The original hypothesis was that a point would be reached where there was no noticeable improvement in the Dice distance scores despite increases in the sample sizes. This hypothesis was disproved. Though improvement slowed, it did not stop. There were still visible improvements among the highest frequency items, even for comparisons of multi-million-token samples (See section 4.1.1.6).

The methodology was therefore revised to start with comparisons of 1 million tokens and increase the sample size in increments of 1 million.

Figure 3.6

Mean Dice distance scores between samples of conversation of 1 million (dotted), 2 million (dashed) and 3 million (solid) tokens. Dice distance was calculated for the most frequent 100 word types and increasing in increments of 100 words until the comparisons included the top-ranked 10,000 word types.

In Figure 3.6, the standard deviation is not shown since the 1, 2, and 3-million token comparisons are based on a diminishing number of samples. Only 7.05 million tokens of casual conversation could be collected for this study, so at most seven samples of 1 million tokens each could be compiled. Three samples of 2 million tokens each were possible for the next round of comparisons, and only two samples could be made for the 3-million-token comparison. Since the Dice coefficient requires the measurement of the distance between samples, it would be impossible to directly measure the amount of variation to be expected from a sample equal to the total amount of data in hand. Ideally, for this study one needs to have many times the amount of data one will eventually find is desirable.

3.3.2 Projections based on the Dice distance between non-overlapping samples.

For the written text types, eight or more comparisons were possible. For comparisons of two 8-million-token samples, one needs at least 16 million tokens for that text type. This number of parallel data points allows for tentative projections to be made. Using the Dice distance scores for the same length comparison, e.g. the top-ranked 1,000 words, from each round of the study, i.e. samples of 1, 2, 3 million tokens etc., one could extrapolate to larger sample sizes that cannot be measured directly in this study (See section 4.1.1.6).

For example, for the text type general written exposition comparisons were made of for samples of 1-5, 10, 15 and 20 million tokens. Taking the Dice distance score for the most frequent 1,000 words from each of these comparisons gives eight points from which to plot a trend line. Since Zipf's law is probably the most robust power law distribution known (Clauset, Shalizi & Newman, 2009: 26), a power law trend line gives an excellent fit ($R^2 = 0.9955$). Judging from this projected trend line, one could confidently expect there to be a Dice distance of only around 0.02 for the most frequent 1,000 words in 50-million-token samples of general written exposition. That would mean that frequency lists drawn from these 50-million-token samples would likely only differ by only two words.

Figure 3.7

Projection from the Dice distance measures for the top-ranked 1,000 word types for samples of general writing. (Diamonds indicate data points on which the power law projection is made.)

One caveat has to be made, though, since the last two data points (15 and 20 million) are from comparisons of only three or two samples respectively. For conversation extrapolating results is especially problematic since the amount of data available is very limited.

3.3.3 Dice distance study with replacement. In an effort to create a greater number of more consistent data points from which to project from, a script was written that used the

statistical technique of bootstrapping. Unfortunately, like the proverbial pulling oneself up by one's bootstraps, this effort proved less than practical.

The idea is to select multiple samples from the total data collection for a text type, but without eliminating those files from the pool available for the selection of the subsequent samples. The script was written to conduct this sampling with replacement to create 10 samples at each sample size and calculate the mean Dice distance and standard deviation for each 10x10 matrix (Appendix C.1). In this way, mean and standard deviation could be plotted and projections could be made from a larger number of consistent data points. It was also hoped that this bootstrapping would allow for larger comparisons, that is one could create samples that are slightly larger than 50% of the available data. This would mean that each of these large samples would inevitably overlap with some of the other samples. It was hoped that this would not greatly reduce the Dice distance between the samples.

One drawback of this methodology is that the script can take many days or weeks to run if the text type contains a diverse set of word types, as general writing does. It was also uncertain how sampling with replacement would affect the results, especially for samples larger than 50% of the available data.

3.4 How different are the text types?

3.4.1 One core vocabulary, or more? How different are the high-frequency words of each of the key text types? What is the clustering pattern of the four key text types? Is there a large gap between spoken and written texts? A large gap between oral and written texts or between any of the text types would show that there is a need to have individual core vocabulary lists.

Kilgarriff (1997), for example, created two sets of frequency bands for the LDOCE, one for spoken and one for written English. Coxhead (2000) demonstrated the usefulness of an academic word list as the next step beyond West's GSL. She saw that there is a general core vocabulary as well as a shared academic vocabulary. West's list was designed primarily for written English. He had long advocated a focus on EFL reading skills over conversation.

When West was in Bengal, he realized that most of the learners had little need for spoken English, but they did need access to informational texts that were not available in their own language. He also realized that most of the children would not stay in school long enough to develop fluency in spoken English. "Basic literacy skills in English . . . could be acquired much more rapidly and . . . be used later in life" (Howatt, 2004: 279). Therefore, West (1960: 18) concluded that "[a]t about the five hundred word level we believe that reading and

speech-teaching should be split.” This apparently means there should be two core vocabulary lists branching off of a nucleus of 500 words.

Some complained that the GSL does not include “the words needed by visitors to England or new residents . . . for use in everyday situations” (Howatt, 2004: 289-290). One could also argue that it lacks many items needed in everyday situations in East Africa, India, the Philippines, Singapore and many other contexts where English is widely spoken. This conception of the core vocabulary insists that it cover essential oral communication.

The taxonomy of the English language texts chosen for these studies contains four key types of text based on Biber (1995). Now, it is necessary to determine how different from each other these four text types are and discover if one core vocabulary unites them all.

3.4.2 Dice distance between text types. The difference between text types can be determined using the Dice coefficient in much the same way as the differences between samples of a single text type were measured. Ten samples of 10 million tokens each were randomly selected for each of the four key text types. In other words, there were 40 samples, 10 each for conversation, narrative writing, general writing and academic writing.

Except for conversation, these are the same samples that were used for the 10-million-token Dice distance study with replacement (See sections 3.3.3 and 4.1.2). The data collection for conversation only contains 7.05 million tokens, so the script was rewritten to allow true sampling with replacement for each sample (c.f. Appendix C.1: line 26). That is, each time a file is selected to be part of one of the samples it remained available to be selected for that sample again. In this way, some files were selected two or three times in order to accumulate a sample of 10 million tokens. [For the Dice distance study with replacement (Section 3.3.3), it was possible for a file to be part of more than one sample. For the conversation samples in this study, a file could be selected multiple times to be part of a single sample.] In this way, the conversation samples were bootstrapped to the same size as the other samples.

The sample size of 10 million tokens was selected as a compromise between what was feasible for conversation that had only around 7 million tokens and general writing at nearly 50 million tokens. It was thought that 10 million would be reasonably stable, while at the same time not stretching the conversation data too unrealistically.

Lists of the most frequent 500 words in each of the 40 samples were selected, and the Dice distance was calculated between the samples in this 40x40 matrix. The Dice distance comparison was then repeated for the most frequent 1,000 words in each of the 40 samples. This was repeated in increments of 500 until it included the most frequent 6,000 words in each of the samples (c.f. Appendix C.1: lines 80-113 and 150-166). The four text types would

not be a uniform distance from each other at every frequency level. It was expected that the text types would be closer together at the higher frequencies and then diverge to some degree.

The mean Dice distance was then calculated for each increment for each of the four key text types. In other words, the mean Dice distance was calculated between the ten lists of the most frequent 500 words in conversation and the ten lists of each of the other three text types. This gave an average Dice distance score between each group of samples: conversation vis-à-vis narrative, general and academic writing; narrative writing versus conversation, general and academic writing, and so on.

This cross-comparison was then repeated for each increment of 500 ranked words from 500 to 6,000. Each of these 12 comparisons was then plotted on a radar diagram with each text type occupying one corner of the diagram (Figures 4.56 and 4.57). (See the “Text type comparison matrix analysis” spreadsheet in the attached supplemental files for the matrix and all 12 diagrams.)

A cluster dendrogram (Figure 4.58) was also plotted to illustrate the relationships between the text types using their mutual Dice distances (Gries, 2009d: 317-318; Lucas, 2012). The dendrogram is a sort of family tree that graphically illustrates the relatedness of each of the four text types. Similar diagrams are created to illustrate genetic relationships of species or of individuals in a population.

A survey was then made of the word types which were common to all four text types and which were not. Proper nouns and marginal words found on Nation and Webb’s (2011) COCA/BNC base word lists 31 and 32 were not included in the sorting for this part of the analysis.

3.5 Which Unit Should Be Used For Counting Words?

“If the list is being made for productive purposes, that is, for speaking or for writing, then the word type or the lemma is probably the best choice because knowing how to use one member of the word family does not mean that you know how to use the other members. If the list is being made for receptive purposes, that is, for listening or for reading, then the word family at an appropriate level for the proficiency of the learners is the best choice.” (Nation, to appear)

Creators of core vocabulary lists have the option of choosing a smaller definition of a word that does not assume knowledge of any other word forms, or a definition that includes the inflectional endings words can have in English, or a broader knowledge of a family of

derivationally related words. Each option has advantages and disadvantages. The smaller definition(s) is not in danger of making false assumptions about the learner's morphological knowledge and does not burden the learner with any more word forms than are absolutely essential for initial learning. A larger definition helps the learner leverage a limited familiarity with some regular morphemes to vastly increase the number of word forms they recognize. How many more word forms would be known or would need to be learned for each definition of the word is not known. How much text coverage is achieved in each of the text types for each definition is also not known.

The present study skips the intermediary unit of lemma and compares lists of types and families to discover how much of a difference this makes in terms of text coverage and the number of word types that need to be learned.

3.5.1 Text samples

A single 10-million token sample was selected for each of the four key text types. The samples were composed of 1,000 randomly selected blocks of 10,000-tokens. Except for the general writing texts, these are the same text blocks that were text described earlier. For this study, however, digits were stripped from the general writing texts and the texts were then reassembled into 10,000-token blocks. In the first study (section 3.3), it was discovered that general writing contains an enormous number of letter-digit combinations, most of which occur very infrequently. All 10 of the 10-million-token samples of general writing used in the Dice distance with replacement study (3.3.3) contained more than 133,000 word types when letter-digit combinations were included. The large number of unique letter-digit combinations would not add any precision to this study, but would have made processing the data very cumbersome.

3.5.2 Family and type frequencies

Frequency lists of word types were first created for each of the four samples (c.f. Appendix C.1: lines 30-52). To create family frequency lists, the type frequency lists were merged with Nation and Webb's COCA/BNC lists (2011: Chapter 8). These lists contain headwords plus a list of the other family members indented on the subsequent lines, e.g.

```
SEE 0
SAW 0
SEEING 0
```

SEEN 0
 SEER 0
 SEERS 0
 SEES 0
 SEEST 0
 UNSEEING 0
 UNSEEINGLY 0
 UNSEEN 0
 SEEM 0
 SEEMED 0
 SEEMING 0
 SEEMINGLY 0
 SEEMS 0

A script was written to create a two column file in which the headword was duplicated in the column to the right of each of the types, including the type that is the headword (Appendix D.1). Additional headwords had to be created for new word types that were not found on the COCA/BNC lists. A pair of scripts then merged this family-type list with the type frequency lists for each of the four text type samples (Appendix D.2). This resulted in a type frequency list with a headword label for each type. This list was then copied onto an Excel spreadsheet and a pivot table was constructed that totaled the frequencies for each word family by summing the frequencies of each of its types, e.g.

SEE	29552
SAW	6881
SEE	16231
SEEING	1428
SEEN	4555
SEER	6
SEERS	2
SEES	285
UNSEEING	23
UNSEEINGLY	30
UNSEEN	111

The most frequent 3,000 word types were then selected from each of the four frequency lists and each list of word types was saved as a text file. The pivot table of word families was then sorted by the sum of the frequencies of its members, and the word types that

are members of the most frequent 3,000 families in each text type were selected and also saved as text files. This created two sets of word lists, one of the 3,000 most frequent types for each text type and one of the word types that are included in the 3,000 most frequent word families.

The number of word types and word families present on each list was then calculated and a comparison bar graph was plotted (Figures 4.59, 4.61, 4.63 and 4.65) The token coverage of each list for its own text type was also calculated by summing the frequencies of all of the word types included on the list (Figures 4.60, 4.62, 4.64 and 4.66).

Each of the eight lists (four word type lists and four word family lists) were processed using the Range 32H program (Nation & Heatley, 2002) and the GSL, AWL lists to discover the amount of overlap between each of the text type lists and the GSL and AWL. The eight lists were then scanned using Range again with the COCA/BNC lists (Nation and Webb, 2011: Chapter 8). This was done and to determine the number of marginal words, proper nouns and abbreviations on each list. (See the “Type Family Study” files for each text type in the attached supplemental files.)

3.6 Dispersion

In the text type comparison study (sections 3.4 and 4.2), instances of local repetition were discovered where one speaker repeated a word many times in a single utterance or conversation causing a couple of words to be listed in the top 6,000 word types for conversation that would otherwise have been very rare. A dispersion measure, e.g. Julliard’s D or Gries’ DP (Gries, 2010c), is able to identify items which are poorly dispersed, and a dispersion threshold can be used to eliminate poorly dispersed words. However, this will lead to frequent dilemmas. A word may be unevenly distributed, but that does not mean it should be automatically discarded. Its minimum frequency might place it further down the list or in a subsequent frequency band. A consistent solution to this problem is a usage coefficient, such as U (Lyne, 1985: 125) or U_{DP} (Gries, 2008e: 426; Matsushita, 2012: 99). A usage coefficient factors a word’s frequency by its dispersion score. The score for DP increases as distribution becomes more uneven, so U_{DP} is found by multiplying the word’s frequency by $1-DP$. A higher DP score leads to a greater reduction in the U_{DP} score.

To calculate the dispersion of each word type and its usage coefficient, each text type collection needs to be divided into several parts. This can be done artificially by simply segmenting the data by the number of words, as was done for the studies in this thesis, or one could use the existing text or category divisions of the corpus. If one is preparing a core list

for more than one variety, one could first adjust for dispersion within each variety, but since a decision will have to be made about a word's dispersion across the varieties, it is probably more efficient to calculate the dispersion jointly for all the sections of all the varieties.

If one wishes to divide each text type sample into 10 sections and the core list covers 10 different varieties of English, that would mean creating a frequency table for 100 files. This is beyond the capabilities of Nation and Heatly's (2002) Range program. The R-script in Appendix E.1, however, can theoretically accommodate any number of files. It has been written to be completely automatic beyond opening the R interface to execute it. After a joint frequency list has been created, it can be opened in a spreadsheet program, like LibreOffice Calc or MS-Excel.

Gries' DP (2010c) and the usage coefficient U_{DP} (Gries, 2008e: 426; Matsushita, 2012: 99) are used here for illustration. DP is calculated by first subtracting the observed frequency of each word from the expected frequency if it were uniformly distributed. If the total frequency of a word is 100 in a comparison of ten samples, then one would expect the word to occur 10 times in each sample. DP uses normalized frequencies, so this word would be expected to have a normalized expected frequency of 0.1 in each of the ten sample if the samples are all the same size. The normalized observed frequencies in each sample are subtracted from this expected frequency and the sum of the absolute values is divided in half. This yields a number that varies from 0 to 1. Zero indicates a perfectly even distribution. The maximum score of 1 is not actually possible, even if all the observed tokens were in a single sample, since at least some of the tokens would have been expected in that sample anyway.

If all 100 tokens of a word type are in a single sample, then the sum of the normalized absolute differences will be 1.8 (Table 3.5). Dividing by 2 (since the differences are being summed for both directions) gives a DP score of 0.9. Multiplying $1 - DP$ (i.e. $1 - 0.9 = 0.1$) by the word's frequency of 100 yields a U_{DP} score of 10. If the list is sorted by U_{DP} , rather than the raw frequency values, this word's placement on the list will be much lower. A script that can be used to calculate the frequencies, range and dispersion (DP) for each word type in a collection of texts is found in Appendix E.2.

Table 3.5

Example calculation of Gries' DP (2010c) for a word with very uneven distribution.

Sample #	1	2	3	4	5	6	7	8	9	10		
observed frequency	0	0	0	0	0	0	0	0	0	100	10	expected frequency
normalized	0	0	0	0	0	0	0	0	0	1	0.1	
difference	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	-0.1	0.9		
absolute difference	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.9	1.8	Sum
											/2	
											0.9	DP

In this next example (Table 3.6), the word type has the same total frequency, but it is much more evenly distributed. The expected frequency is the same 0.1, but only slight differences are seen in the first and last samples. This yields a very low DP score of 0.1. U_{DP} for this word will be $(1 - 0.1 = 0.9) \times 100 = 90$. So these two words which are identically frequent will be ranked very differently if the ranking is based on U_{DP} .

Table 3.6

Example calculation of Gries' DP (2010c) for a word with more even distribution.

Sample #	1	2	3	4	5	6	7	8	9	10		
observed frequency	20	10	10	10	10	10	10	10	10	0	10	expected frequency
normalized	0.2	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0	0.1	
difference	0.1	0	0	0	0	0	0	0	0	-0.1		
absolute difference	0.1	0	0	0	0	0	0	0	0	0.1	0.2	Sum
											/2	
											0.1	DP

To see how this works with real world data, 10 samples of conversation from several varieties of English are compared below (Table 3.7). For the sake of illustration, the samples have not been subdivided. Only a portion of the available data for the UK is used, so as not to obscure variations in the other varieties. This leaves a total of 3 million tokens for the comparison.

Table 3.7

Sample sizes used for a dispersion comparison of English varieties from the UK, New Zealand, North America, Hong Kong, the Philippines, Singapore, Jamaica, Ireland and a mixture from East Africa, Europe and elsewhere.

	UK	NZ	CA & USA	HK	IND	PHIL	SING	JA	IRL	Misc
% of total	21.7%	16.7%	12.7%	12.3%	6.3%	6.3%	6.3%	6.3%	5.7%	5.7%
tokens	650,000	500,000	380,000	370,000	190,000	190,000	190,000	190,000	170,000	170,000

Since the quantity of data available for each variety differs, the frequencies need to be normalized according to the size of each sample. (See the “DP demonstration” spreadsheet in the attached supplemental files.) The verbal particle *lah* (see section 4.2.2.2) has a rank of 259 in this comparison. But, as can be seen in Table 3.8, its distribution is highly uneven. (The two tokens in the North American sample are different marginal words that happen to be homographs.) The DP score for *lah* is 0.935 which means it would be considered to have a normalized frequency of only 26.8 per million which would give it a rank of 1,485. This is, however, still within the top 2,000 words, despite its having a range of only 2 (really 1). This is a good case for the use of a range threshold in addition to a dispersion coefficient to eliminate words from consideration that are only present in a small minority of varieties or samples. For example, one could eliminate any words from consideration if they occur in fewer than a set number of samples.

The minimal response *yeah* is much more evenly distributed. It is somewhat underrepresented in both the Singaporean and North American samples. In both of the corpora from these varieties, it was often transcribed as *ya* instead, which has a DP score of 0.758. The DP score for *yeah* is 0.216 which would adjust its normalized frequency down to 10,139.6 per million which would increase its rank by only one position from 12 to 13.

Table 3.8

Examples of words with even and uneven dispersions in a 3-million-word collection of conversation from several varieties of English. Frequencies are normalized per million tokens.

	Total	UK	NZ	CA & USA	HK	IND	PHIL	SING	JA	IRL	Misc
lah	412.7	0	0	0.5	0	0	0	650.5	0	0	0
yeah	12,932.7	906.5	1,974.4	666.05	2,056.8	1,401.1	1,485.3	30.5	723.7	1,328.2	2,187.1
ya	1,810.3	18.3	0.2	700.5	4.3	1.6	0.5	1,308.4	14.7	0.6	67.1
sir	117.3	10.5	3.80	0.53	0.5	104.2	20.0	1.1	8.4	1.2	5.0

In this example comparison, 46 words in the top 1,000 word types have a DP score of greater than 0.5. They fall into four categories: marginal words, terms related to language or nationality, contractions, and regional lexical variants.

Marginal words are phonetically variable and also do not have established spellings and so may have differing transcriptions in different corpora: *aha*, *aw*, *aye*, *bloody*, *eh*, *em*, *er*, *erm*, *hm*, *hmm*, *mhh*, *mhm*, *mmm*, *na*, *nah*, *ok*, *ooh*, *uhn*, *uhu*, *um*, *yah*, *yep*, *yup*, *ya*. Not all marginal words are poorly dispersed, however.

Of course, the frequency of proper nouns related to nationality and language vary according to region: *China*, *Chinese*, *Hong*, *Kong*, *India*, *Indian*, *Japan*, *Japanese*, *Maori*, *Singapore*, *Spanish*, *Zealand*.

Contractions common in speech are also typical of certain varieties. *Dunno*, is more commonly British, while *kinda*, *wanna* are more typical in North America and the Philippines. *Cause* is typical in the North American samples, since it is more often transcribed as *cos* in other corpora. *Gotta* is common in British and American speech, but not in the other varieties. *Gonna* is well dispersed across all the varieties, though.

Finally, there are typical cases of lexical variation or alternation: *lah* [a regional borrowing], *mom* [vs. *mum*], *pound(s)* [a local currency], *reckon* [vs. *guess*], *secondary* [vs. *junior high*], and *sir*. The final item, *sir*, is an example of what Kachru describes as “an extremely deferential lexical spread based on the politeness hierarchy of the L₁” (1992 :313). This already existing English lexical item is serving a sociolinguistic need in an Indian cultural context which is raising its frequency beyond what is typical in most other varieties of English.

3.7 Scope, limitations and assumptions

3.7.1 A tentative attempt at a core vocabulary for English as an International Language. One research question that is implied, but not addressed explicitly, is how much variation there is between the core of regional Englishes? Ideally, individual collections of each of the key text types should be compiled for several varieties. Unfortunately, that amount of data is not yet available. This is the same issue faced by Nelson (2006a). Pooling all the available varieties to form one collection for each text type was the best possible arrangement. This should provide the largest possible samples, while at least giving some amount of representation, if uneven, to many different varieties. As described in section 3.6, calculating each word's range (the number of samples a word occurs in) and dispersion (how evenly it occurs across the samples) can be used to filter out words that are not representative of the way English is spoken in most situations.

3.7.2 Orthographic words. The studies in this thesis are based on the operational definition of the word as an orthographic form, ignoring identically formed parts of speech, homographs and polysemes. Since corpus lines were collapsed and the words divided into uniform blocks of 10,000 tokens, sentences and utterances were often cut off from their original contexts. Words that are now contiguous in the data samples, may not have been originally, and words that were adjacent could now be in separate files. This makes semantic or part-of-speech tagging unreliable. A manual examination was made of the most frequent 1,500 word types for conversation, but no dramatic distinctions were found for differences in part of speech. For example, the verb and the adjective *clear* both share a core meaning. It does not seem reasonable to divide them simply because they are different parts of speech. A distinction might be of interest when the senses are more divergent or when the balance of each sense of a word differs for different text types. *Train*, for example, is used in the sense "to teach" in approximately 5% of cases in conversation, but in academic writing, in over 25%. For *train*, one could still argue that they share the core meaning of some sort of fixed path of progression, either rails or the steps in a course of learning.

More troubling differences were seen when word forms were homographs or polysemes. For example, the form *rest* occurs in the top 1,000 types of conversation. The sense of the majority of its uses, is "remainder" rather than "sleep." In these studies, the frequencies of both senses have been conflated. Palmer's (1934: 49) warning that these studies may be "straining at vocabulary gnats while swallowing semantic camels" is acknowledged. Nonetheless, Wang and Nation (2004) found that where there are homographs on the Academic Word List (Coxhead, 2000), one homograph usually has a much higher

frequency. For the few words that are homographs in the AWL, there is less than a 5% error. Kevin Parent (2012: 73) found a similarly small number of homographs on the GSL (ca. 10%) and that “for 72% of these homonyms the commonest meaning accounts for 90% or more of the total uses of the form.”

3.7.3 Multiword Units. One further limitation of these studies is that the focus is narrowly on individual words and does not recognize the existence of phrasal lexical items, like *get up* or *by and large*. The word is not always the unit which should be taught since much of language is prefabricated (Hunston, 2002: 191-192; Nattinger & DeCarrico, 1992). A limited exploration is made of multi-word patterns in conversation and general writing in Table 4.2, but a more thorough study is beyond the scope of this thesis.

CHAPTER 4

RESULTS

This chapter reports the results of the three main studies that make up this thesis. Section 4.1 contains the results of two versions of the Dice distance study that looked at how much data is necessary to produce stable high-frequency word lists. The methodology of the second version of that study, though superior in some ways, was undone by the quantity of data available.

Section 4.2 also made use of the Dice coefficient to measure the distance between the four key text types described in section 3.1.2. Section 4.2.2 examines the high frequency words which are shared or unique to each of these key text types. Section 4.2.3 compares samples of the four key text types to several sets and lists of words in order to determine where the core vocabulary is among the four text types.

Section 4.3 explores how the definition of a word as a word type or as a word family impacts text coverage and the number of word types that need to be learned. A preliminary comparison is also made with West's (1953) General Service List and Coxhead's (2000) Academic Word List.

Section 4.4 concludes the chapter with the construction and analysis of a trial core list. Objective steps in constructing a core vocabulary list are outlined in 4.4.1. This trial list is then compared to the West's (1953) GSL in section 4.4.2.

4.1 Dice distance between samples of the same text type

4.1.1 Dice distance without replacement

In order to answer the first research question, "How much data is necessary to give consistent results for high-frequency vocabulary?" the Dice coefficient was used to measure the distance between samples of each of the key text types: conversation, as well as narrative, general and academic writing. As explained in section 3.3.1, this application of Dice distance shows the percentage of words that vary from one sample of texts to another, so a lower Dice distance shows greater self-similarity among the texts of that text type. One compares a certain length of the frequency lists from a group of text samples. For example, comparing the most frequent 1,000 word types from a set of 1-million-token corpora of text type A, one finds the lists differ on average by 0.02. In other words, Texts of type A differ on average by around 20 words out of the most frequent 1,000. When comparing the most frequent 1,000 word

types from a set of 1-million-token corpora of text type *B*, on the other hand, one finds they differ by 0.09. That is, texts of type *B* differ on average by around 90 words out of the most frequent 1,000. This indicates that the texts in set *A* are more similar to each other than the texts in set *B*.

As was shown in Figures 3.5 and 3.6, how similar the texts of a given type are to each other can be plotted as one enlarges the scope of the comparison from the most frequent 1,000 words to the top 2,000, 3,000, etc. A comparison of the complete range of Dice distances (from the most frequent to the least frequent word types) for each of the four key text types at the same corpus size shows that each text type has a unique shape. As one compares more and more of the word types in the samples, the overall distance between the samples becomes greater and greater, but the rate and pattern of each text type's evolving difference is unique.

Figure 4.1 shows that for 3-million-token samples of each of the four text types, conversation starts out being the most uniform. That is, the highest frequency words in conversation are very similar in every sample. However, as one steadily includes words that are lower in frequency, the conversation samples soon become very diverse. For the most frequent 7,000 word types, conversation samples are less similar to each other than any other text type. In the space of the most frequent 7,000 word types, conversation goes from being the most homogenous to the most heterogeneous text type.

Two other text types, general and academic writing eventually surpass the Dice distance score for conversation, but only at a much higher rank. This is because these two written text types use a somewhat larger vocabulary than conversation. In conversation, however, the diversity of its vocabulary shows up more quickly, i.e. among the higher frequency word types.

Conversation has the smallest vocabulary of the four text types (seen in the shortness of its line in Figure 4.1). This confirms Biber's (1988) description of informative texts as having more extensive vocabularies. That letter-digit combinations were counted as words in this study only partially explains why the vocabularies of the written text types are so much larger. (In transcriptions of speech, numbers are always spelled out as words.) Equal-sized samples show that general and academic writing both contain a significantly greater number of word types than the two more dialogic text types, conversation and narrative writing (Figure 4.1).

Writing allows for careful consideration and editing to achieve a varied and interesting vocabulary. It also allows for the preservation of archaic word forms, such as *thee* and *thou* that were mentioned as being in the first thousand of Thorndike and Lorge's *The Teacher's Word Book of 30,000 Words* (1944). Even though, only modern corpora were used in this

study, two interesting examples of this phenomenon, *þæc* and *þeccean*, are still found in the general writing texts. *þeccean* is an Old English verb meaning “to cover” (cognate of German *decken*), and *þæc* is *thyself* (cognate of German *dich*). These two words would have been among the most frequent words of spoken English 1,000 years ago, but they resurfaced in the 20th century in a story about C. S. Lewis that was collected as part of the BNC. While teaching at Oxford, Lewis hosted beer and Beowulf nights at which he would teach his students chants for the declensions of Anglo-Saxon pronouns and verbs to help them make it through a paper on the history of English.

Figure 4.1

Dice distances for the four key text types for corpus samples of 3 million tokens each. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.

The highest point for any of these vectors does not exceed a Dice distance of 0.5. In other words, the samples differed from the other samples of their text type by around 50% at most. In Figure 4.1, the Dice distance only once exceeds 4.0. The maximum Dice distance is actually slightly greater than shown, but a comparison could only be made to the number of types in the sample with the smallest number of types. If the comparison were extended beyond the length of the shortest list, some samples would be judged to be more similar to each other simply because they were alike in *not* containing a particular word type. This study aims to quantify how much shared vocabulary exists between samples of a text type, so knowing that samples A, B and C are similar to each other in that they do not contain the

word type *zukunftsmusik* is irrelevant (Table 4.1). The comparison is also rounded to the last full increment of 1,000 word types for the sake of programming and graphing.

Table 4.1

Example of a partial comparison of types used to calculate Dice distance.

	A	B	C	D
<i>abandon</i>	1	1	1	1
<i>abridged</i>	1	1	1	1
<i>zodiac</i>	1	0	0	0
<i>zoologists</i>	0	1	0	0
<i>zukunftsmusik</i>	0	0	0	1
<i>zygote</i>	0	0	1	0

The reason the maximum Dice distance for each type is limited to about 0.5 is related to one formulation of Zipf's law. The types most likely not to be shared by other samples are the hapax legomena. The approximate quantity of hapax legomena follows the formula (Zipf, 1949):

$$\frac{1}{f(f+1)}$$

This means that the fraction of types in a corpus with a frequency of f should be equal to 1 over f times f plus 1. So, word types with a frequency of only 1 should equal:

$$\frac{1}{1(1+1)} = \frac{1}{2}$$

So, hapax legomena should make up about 50% of any corpus of natural language. The Dice distance for two corpora of the same text type will, therefore not likely exceed 0.50. This can be seen in Figure 4.4.

A close-up view of the Dice distances up to the top 10,000 ranked words (Figure 4.2) shows that narrative writing initially resembles conversation (the dotted and dashed lines respectively), while academic writing (the double line) roughly parallels its more general cousin (the solid line). A comparison of conversation and general writing is a good place to begin to decipher the dynamics that form the pattern of Dice distances for each text type.

Figure 4.2

Dice distances for the four key text types for corpus samples of 3 million tokens.

Measurements were made in increments increasing by 100 word types up to a rank of 10,000.

4.1.1.1 Conversation and general writing. Contrasting conversation and general writing, one sees that the vocabulary of conversation is significantly less variable than general writing for roughly the most frequent 2,000 words (Figures 4.3 and 4.4). Thereafter, the general writing texts are the more stable. In other words, lists of the most frequent 1,000 or 2,000 word types of samples of conversation will be more similar to each than the same length lists for samples of general writing, but lists of the most frequent 5,000 word types of general writing will be more similar to each other than the same length lists for conversation. Why that is the case can be better understood by looking at a Zipf graph of the two text types (Figures 4.6 and 4.7).

Figure 4.3

Dice distances for conversation (solid lines) and general writing (dashed lines) for samples of 1 million (light gray), 2 million (dark gray) and 3 million (black) tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.

Figure 4.4

Dice distances for conversation (dashed lines) and general writing (solid lines) for samples of 1-3 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.

On a Zipf graph the most frequent word type is plotted on the far left of the x-axis and its height on the y-axis indicates its frequency. That would place the most frequent word type close to the top-left corner of the graph. The rest of the word types are arranged in descending order of frequency from left to right on the x-axis in the same manner. The words' frequencies decline so rapidly, however, that a graph based on their actual ranks and frequencies would be quite useless. The line would be almost indistinguishable from the x and y axes (See Figure 4.5).

Figure 4.5

Graph of the word frequencies for the Brown Corpus (Kučera & Francis, 1967).

In order to be readable, Zipf graphs are plotted in double-logarithmic space. That is one finds the logarithm of each word's frequency and the logarithm of each word's rank and the position of each word is plotted based on these two quantities. As was seen in section 2.6.3, logarithms reduce a very broad range of quantities to a very compact length. Zipf graphs traditionally use a base 10 logarithm, so each interval on the scale increases by a factor of 10. The Richter scale uses the same logarithm. This is why a magnitude 6.0 earthquake is 10 times more powerful than a magnitude 5.0 earthquake. This is not just a mathematical trick, but is based in nature. Human aural perception, for example, is also logarithmic. That is why the decibel scale is also logarithmic. In this way, we are able to perceive a very wide range of sounds from a scream down to a whisper.

Plotting the line of word frequencies logarithmically has the effect of compressing the higher frequencies into shorter and shorter lengths of the y-axis bringing the higher points on the graph lower and lower. Likewise, the logarithmic scale of the words' ranks forces the large number of low-frequency words (that therefore have very high ranks) into ever shorter lengths of the x-axis. The far-left interval on the x-axis contains 10 word types, but an equal-length interval on the right side of the x-axis can contain thousands of word types. These two logarithmic scales have the effect of squeezing and bending the line from the L-shape that it was in Figure 4.5 into a nearly straight line descending from the top-left to the bottom-right of

the graph, e.g. Figure 4.6. A perfect diagonal line that is a 45° degree angle from both the x and y axes is said to have a -1 (negative one) slope.

Figure 4.6

Double-logarithmic graph showing the Zipfian distribution of a collection of general writing texts (47.65 million tokens). A dashed vertical line marks rank 6,000.

Figure 4.7

Double-logarithmic graph showing the Zipfian distribution of a collection of conversation texts (7.05 million tokens). A dashed vertical line marks rank 6,000.

General writing shows a nearly perfect -1 slope for the first approximately 6,000 types (Figure 4.6). Even if measured from the highest-ranked word, the line is exceedingly straight. Since the graph is double-logarithmic this means that each successive word has a far greater frequency than the next in a very vertical hierarchy. The Dice distance scores for the top 100 word types in general writing are initially not quite as low as those for conversation, i.e. the words are among the most frequent 100 words in general writing vary somewhat more than the top 100 words in conversation (Figure 4.4). Nonetheless, the Dice distance scores are still fairly small since the most frequent items are so incredibly common.

The shape of the line in Figure 4.6 is ultimately related to the informational purpose of general writing. Written texts are dominated by noun phrases (Biber & Conrad, 2009: 116). These noun phrases contain the majority of the information in these texts. The most frequent n-grams of four words or more in general writing include, and usually also introduce, a noun or noun phrase (Table 4.2). If one were to visualize the lexical tree that is behind general writing, its trunk would look like that of a pine with the definite article *the* as its base.

Table 4.2

The most frequent n-grams of four or more words for conversation (left) and general writing (right) in samples from the International Corpus of English. An apostrophe is considered a word division, so abbreviated forms count as two words. Voiced pauses and recasts have been omitted from the conversation data.

Conversation			General Writing	
rank	frequency	n-gram	frequency	n-gram
1	3,148	I don't know	205	per cent of the
2	951	I don't think	191	the end of the
3	634	I think it's	159	at the end of
4	402	but I don't	133	at the same time
5	396	don't know what	119	as a result of
6	352	don't want to	115	in the case of
7	333	I'm going to	115	on the other hand
8	314	you know it's	108	one of the most
9	307	I don't know what	94	the rest of the
10	284	I don't like	85	is one of the

The lexical tree of spoken language, on the other hand, seems to be more of a pohutakawa. Unlike a pine tree, a pohutakawa does not have a single dominant trunk. Instead it typically has a cluster of several trunks branching out from or near to ground level. In spoken language, there are a number of frequent structures that dominate the frequency list. In the Zipf graph (Figure 4.7), the top left portion of the line is rounded downward since the most frequent words are much more similar to each other in frequency than Zipf's law would predict. Each of the most frequent n-grams of four words or more in conversation is the core of a verbal clause (Table 4.2). There is a group of very common clauses that pervade speech rather than the dominant noun phrase pattern seen in general writing. All 10 of the four-word n-grams in conversation are also more frequent than any in of those in general writing. That shows that pre-assembled structures, e.g. *I don't know what...* make up a larger portion of speech conversation than general writing. This is to be expected since speech is composed in real time. The overall vocabulary size for conversation is smaller for the same reason, since speakers are speaking “off the top of their heads” and cannot edit their work to give it more lexical variety.

Together the most frequent words in conversation form a large mass of tokens that keeps the Dice distance for the most frequent words exceedingly low. However, the diverse clausal strands quickly diverge causing the Dice distance scores to rise rapidly. This is

reflected in the much steeper primary slope of -1.45 in the Zipf graph for conversation (Figure 4.7). Conversation may not have quite the vocabulary size of general writing, but what it has, it gets to quickly. This is understandable when one considers the context of conversation. Conversation is typically face-to-face interaction. One does not want to stand and listen to a well-formed dissertation. Instead conversation is soft assembled in real time between speakers who often have a large quantity of shared experience. In writing, one is less able to assume shared knowledge.

That conversation is interpersonal also contributes to the high degree of lexical variability between samples of conversation. Personal names vary in conversation depending on the speakers and their circle of acquaintances, even if the topics being discussed are the same. Lexical variation in general writing, on the other hand, is topic-based. Of the most frequent 3,000 word types in a 10-million-word sample of conversation, 7.8% were proper nouns (Table 4.25). Proper nouns made up 4.8% of the same-sized sample of general writing (Table 4.19).

In conversation, there are also frequent connections to other linguistic networks of which the speakers are a part, i.e. code-switching and borrowing. This is a natural part of speech and will only become more common as English continues to play a role as an international language. Artificially excluding this variation would be a misrepresentation of the data (See section 3.2.5).

Corpora of spoken English also suffer from a lack of conventional spellings for the transcription of hesitations and other marginal words. *Huh* and other such word types made up 2.0% of the most frequent 3,000 word types in conversation. In general writing marginal forms are basically nonexistent. Since written text is composed “offline,” there is no turn-taking. In speech, if a speaker pauses during conversation, but has not yet finished, the speaker can insert a vocalized pause, such as *hmm*. Writers have punctuation marks that can signal whether or not a pause is the conclusion of an utterance.

Smaller samples of general writing are characterized by a sharp spike in Dice distances at around rank 500 that then levels off into a very gentle rise (Figure 4.8). The abrupt rise for smaller samples is due to the inclusion of a few very frequent but still topic-specific terms. For larger samples there is much greater agreement. The top 500 ranked words in two samples of general writing (23.8 million tokens each) differed on only these 24 words: *am, b, began, board, capital, certain, David, de, expected, game, history, kind, mother, move, needs, role, royal, secretary, sense, Soviet, thing, thus, value, workers*. From this short list, one can already begin to see the range of topics common in this text type.

Figure 4.8

Dice distances for general writing samples of 1-5, 10, 15 (dotted), 20 (dashed) and 23.8 (solid) million tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.

Looking at the complete graph of the word types in general writing (Figure 4.9) one sees that general writing contains a very large number of word types. For the largest samples that were measured, 23.8 million tokens, general writing had a vocabulary of over 200,000 word types. Many of these are letter-digit combinations, but the total vocabulary size is, nonetheless, impressive. The maximum Dice distance between the samples (i.e. when comparing all the word types in each sample) remained approximately 0.4, i.e. about 40% of the word types in one sample were not found in the other. This study is interested in the high-frequency words, however. Encouragingly, the largest samples have a Dice distance of 0.05 or less for the most frequent 10,000 words. This means that very stable word frequency lists for general writing are possible with 20+ million tokens of data.

Figure 4.9

Dice distances for general writing samples of 1-5, 10, 15, 20 and 23.8 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded down to the last full 1,000 word types.

4.1.1.2 Academic writing. Academic writing closely parallels general writing except that academic writing makes more frequent use of generally academic vocabulary, as well as, a wide range of technical vocabulary. Of the most frequent 3,000 word types in a 10-million-token sample of academic writing, 694 (23%) of the most frequent word types were on Coxhead's (2000) Academic Word List. These word types were members of 411 out of the 570 families on the AWL (Table 4.21). This still does not include a large group of technical terms (section 4.2.3.4). This broad technical vocabulary stabilizes at a higher mutual distance than that for general writing (Figures 4.11 and 4.12). This diversity of vocabulary is also seen in the slightly steeper primary and secondary slopes in the Zipf graph for academic writing (Figure 4.10). The context of academic writing, e.g. in a discipline-specific journal, allows a writer to assume some degree of shared technical knowledge. It is not necessary to couch technical terms in as much general vocabulary. On the whole, though, academic writing does not use as many word types as general writing. General writing covers a broader range of topics and contains a wider variety of letter-digit combinations.

The slope between two word types in a frequency list is calculated by first finding the height of the slope by subtracting the frequency of the word with the lowest rank (i.e. the

highest frequency word) from the frequency of the word with the highest rank. Since the Zipf graph is double-logarithmic, the logarithm of each frequency is used. Then the length of the slope is calculated by subtracting the logarithm of the rank of the word with the lowest rank from the logarithm of the rank of the word with the highest rank.

$$\frac{\log(f_{start}) - \log(f_{end})}{\log(r_{start}) - \log(r_{end})}$$

The ranking function on a spreadsheet, such as MS-Excel, assigns the same rank to all words with the same frequency. The rank that is assigned is equal to the rank of the first word in that group of words with the same frequency. In other words, if 100 word types have already been listed (i.e. 100 words have a higher frequency), the next word type will be given the rank 101, as will all other types that have that same frequency. That means that when there are many ties, as there are at the low-frequency end of the distribution, the steps between ranks can become very large. To maintain a Zipfian -1 slope, the steps need to be as long as the frequencies are small. If too few words share the same frequencies, the steps are shorter creating a steeper incline. So a steeper slope means the number of words at those frequency levels is tapering off.

Figure 4.10

Double-logarithmic graph showing the Zipfian distribution of academic writing. A dashed vertical line marks rank 6,000.

Figure 4.11

Dice distances for samples of academic writing from 1-8 million tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.

Figure 4.12

Dice distances for samples of academic writing from 1-8 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.

4.1.1.3 Narrative writing. Narrative writing turned out to be a text type for all seasons. As noted earlier, it begins closely following the pattern for conversation (Figure 4.2). This is not terribly surprising since the majority of texts in this type are from novels and short stories, which contain a large quantity of simulated dialogue. The upper portion of the Zipf graph for narrative writing also closely resembles that of conversation (Figure 4.13). The seemingly unstable region at the very top of the Zipf graphs for conversation and narrative writing is apparently not as random as it appears (c.f. Figure 4.7). The primary slope for narrative writing is steeper than that of the other written text types, but not nearly as steep as for conversation. Narrative writing clearly starts out somewhere between general informational writing and authentic face-to-face conversation.

Figure 4.13

Double-logarithmic graph showing the Zipfian distribution of narrative writing. Dashed vertical lines mark ranks 6,000 and 15,000.

After following conversation for the most frequent 2,000 or so words, the rate at which narrative writing diverges suddenly slows and follows roughly the same path as academic writing. However, it continues to slow until it is actually lower than the mutual distance for general writing (Figure 4.2). The greater lexical diversity seen in academic writing stems from its rich academic and technical vocabulary. Numerous personal and place names, personal titles and other terms used for characterization and setting descriptions are, in a sense, the technical vocabulary of narrative writing. Of the 3,000 most frequent word types in a 10-million-word sample of narrative writing, proper nouns made up 8% of the types, just slightly more than for conversation (Table 4.25 and 4.27). Once the corpus samples grow large enough and contain the majority of these terms, the rate at which the Dice distance increases slows considerably (Figures 4.14 and 4.15).

There is always an influx of new low-frequency words, but beyond rank 9,000 or 10,000, there are not nearly as many new word types in narrative writing as Zipf's law would predict. The drop off is so extreme that in the Zipf graph (Figure 4.13) it was impossible to reasonably fit a line from rank 6,000 to the end, as was done in the other Zipf graphs. Instead, the secondary slope was measured from rank 15,000, which is approximately where the Zipf-Mandelbrot model estimates the second slope should begin (Mandelbrot, 1953; Sorell, 2012).

As was mentioned above, the hapax legomena in any sample are the least likely to be shared by another sample. For narrative writing, the point at which the word type frequency drops to 1 is more obvious than it is for the other text types. At that point, the angle of the Dice distances visibly shifts direction again since almost every new type is unique to the sample in which it is found. This creates the wing-like pattern seen in Figure 4.15.

Each text type shows a uniform incline at the end of the distribution when only hapax legomena are being compared. Though this pattern is most clearly visible in narrative writing, it can also be seen in the patterns in the other text types if one looks carefully. Each comparison is made of a fixed number of types, but corpus samples would not usually have exactly the same number of types at each frequency level. Word types are sorted by frequency and then alphabetically. Certain types that have the same frequency are inevitably left out of one comparison and bumped to the next simply because of alphabetization. The larger comparisons in this first study are of only two samples where all the available data is divided into two random groups. A slight dip in the Dice distance marks the spot where the two lists catch up with each other at the end of the dislegomena (Figure 4.15). At this point, all or very nearly of the shared vocabulary has been compared.

Figure 4.14

Dice distances for narrative writing for samples of 1-9 million tokens. Measurements were made in increments increasing by 100 words up to a rank of 10,000.

Figure 4.15

Dice distances for narrative writing for samples of 1-9 million tokens. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all of the compared samples rounded to the last full 1,000 types.

4.1.1.4 Conversation versus simulated dialogue. Probably the single greatest difficulty with the conversation text type is that transcripts of authentic, casual conversation are rare in comparison to the amount of data available for the other text types. Recording conversations is technically challenging, and the act of recording tends to alter what is said. Transcription is also costly and often inconsistent. What is probably the most important text type in the language is unfortunately the least accessible.

In section 2.5.5, it was seen that Vakar (1966) collected dialogue from Russian dramas to approximate spoken Russian. From what has been seen about the connection between narrative writing and conversation, this assumption is at least partially correct. The SUBTLEXus corpus (Brysbaert & New, 2009) taps another plentiful alternative in the form of subtitles for film and television. A Zipf graph of their SUBTLEXus corpus (Figure 4.16) shows that the primary slope is indeed very similar to authentic conversation. Fitting the secondary slope presented the same problem seen in narrative writing. This deserves further investigation. TV and film are visual literature rather than true interactions, so it is interesting that the secondary slope in SUBTLEXus resembles narrative writing more than authentic

conversation (c.f. Figure 4.7). However, it may also be that the conversation corpus is simply too small to show the true angle of its secondary slope.

Figure 4.16

Double-logarithmic graphs showing the Zipfian distribution of the SUBTLEXus corpus. A dashed vertical line marks rank 6,000 (left) and, in two graphs, rank 15,000 (right).

4.1.1.5 Graded writing. The ultimate goal of this study is to assist in the creation of vocabulary lists for language learning which could also be used to guide the writing of vocabulary-controlled texts. A comparison of these text types with graded materials is, therefore, also of interest. A Zipf graph of a corpus of Oxford graded readers (Figure 4.17) shows a pattern that is remarkably similar to narrative writing, the text type to which these graded readers belong. One interesting difference is the degree to which the top of the primary slope is rounded downward. This would indicate a high degree of homogeneity among the most frequent words of the texts. This shows the effect of the deliberate control of the vocabulary.

Figure 4.17

Double-logarithmic graph showing the Zipfian distribution of a corpus of OUP graded readers (763,507 tokens). A dashed vertical line marks rank 6,000.

4.1.1.6 Projecting to larger corpus sizes. The Dice distance methodology requires one to have at least twice the amount of data one wishes to measure. However, with a number of empirical comparisons, it is possible to tentatively extrapolate to even larger corpus sizes that have not been directly measured.

If a cross section of the data is plotted, e.g. each of the Dice distance scores for the top-ranked 3,000 words for all the samples of certain text type, one will see the rate of change declines according to a power law. The gap between the Dice distance for the top 3,000 words for the smallest sample and the next largest sample will be fairly large. The gap between the second and third largest samples will be much smaller than it was between the smallest two samples. The fact that this decline follows a power law is not surprising since Zipf's law shapes the overall distribution of word frequencies, and it is one of the most robust power laws ever discovered (Clauset, Shalizi, & Newman, 2009: 28). With several known Dice distance measurements and a predictable rate of decline, a reasonable projection can be made to at least the total quantity of data currently available and should be a reasonable estimate of how much improvement can be expected if additional data were available.

General writing. For general writing, eight data points were used to project to larger sample sizes (Figure 4.18). These eight points are the measurements for Dice distance for 1 through 5 million tokens in increments of 1 million. Thereafter, the measurements were made at 10, 15 and 20 million (c.f. Figure 4.9). The final measurement of 23.8 million was not used for projection since it was an odd-sized increment. The shift to increments of 5 million tokens was made to speed up processing. The script for processing one measurement of Dice distance at the larger samples sizes can take weeks to run. Looking at the line which shows the Dice distance for the most frequent 9,000 words (round points) in Figure 4.18, one sees that the samples of 1 million tokens have a Dice distance of just under 0.24, which means that nearly a quarter of the words on any list were different from the words on the other lists. By 20 million tokens this distance has dropped to about 0.055 or a difference of around 5.5%. Projecting forward to a sample size of 50 million tokens, one could expect a Dice distance of slightly less than 0.04.

The line for the top-ranked 6,000 word types consistently parallels the line for 9,000 words at a somewhat lower Dice distance. Something interesting happens with the line for the most frequent 3,000 words, however. This line matches the line for the top 2,000 words for the first 5 million and then actually falls below it. The projected Dice distance for the top 3,000 words actually matches that of the top 1,000 words when it gets to sample sizes of 50 million tokens. This “sweet spot” indicates that a frequency list for the top 3,000 words could be just as stable as a list for the top 1,000 words. Using all the general writing data that was assembled for this study (47.65 million tokens), the top-ranked 3,000 word types should vary by only around 2%. That should increase to a little more than 3% for the top 6,000. Even for the top-ranked 9,000, the variance should be around 4%. That means that the top-ranked 9,000 words on a frequency list from the 47.65-million-token collection of general writing should vary by only about 360 words from the list of top-ranked 9,000 words taken from another collection of general writing of the same size.

Figure 4.18

Projections from the Dice distance measures at various ranks for samples of general writing. (Points indicate samples sizes at which measurements were made. Trend lines are power law projections.)

Narrative writing. The Dice distances between samples of 1-9 million tokens of narrative writing were measured in increments of 1 million tokens. The lists become steadily more variable as one increases the length of the frequency list. Projecting to a size a sample size of 25 million tokens, one could expect the lists of the top 1,000 words to vary by around 2% or 20 words (Figure 4.19). With the narrative writing data used in this study (18.42 million tokens), a variation of close to 2% can be expected for the most frequent 1,000 word types. For 2,000 and 3,000, that rises to nearly 4% and 5% respectively.

Figure 4.19

Projections from the Dice distance measures at various ranks for samples of narrative writing. (Points indicate where measurements were taken. Lines are power law projections.)

Academic writing. The Dice distances between samples of 1-8 million tokens of academic writing were also measured in increments of 1 million tokens. Similar to narrative writing, the lists become steadily more variable as one increases the length of the frequency list. Projecting to a size a sample size of 25 million tokens, one could expect the lists of the top 1,000 words to vary by slightly more than 3% (Figure 4.20). The list for the top-ranked 2,000 and 3,000 lists are only slightly more variable at around 4% and 4.5%. A total of 17.91 million tokens of academic writing were assembled for this study. At this size, there would be an approximately 4% variation for the top-ranked 1,000 types. The variation for top 2,000 and 3,000 word types should be slightly above or below 5%.

Figure 4.20

Projections from the Dice distance measures at various ranks for samples of academic writing. (Points indicate where measurements were taken. Lines are power law projections.)

Conversation. The projected Dice distances shown in Figure 4.21 are based on only three measurements. The Dice study of conversation was originally done in increments of 100,000 from a size of 100,000 to 1.7 million tokens and then for 2, 3 and 3.5 million. Only the points for 1-3 million are shown in the current graph in order to keep the unit consistent with the other studies. The current 7.05 million tokens of data likely vary by 3.5% for the top-ranked 1,000 words and around 6% for the top 2,000. The top 3,000 items would be much more uncertain at around 8% variability. Projections have been made to a sample size of 25 million, but these projections should be considered very tentative. Having been based on only three measurements, these projections are far more uncertain than those for the three written text types.

Figure 4.21

Projections from the Dice distance measures at various ranks for samples of conversation. (Points indicate where measurements were taken. Trend lines are power law projections.)

All four key text types. Figure 4.22 compares the projected distances for all four key text types for the top-ranked 1,000, 2,000 and 3,000 word types. The Dice distances for conversation frame the range of the other three text types. Its Dice distances for the top 1,000 words are the lowest measured for any of the text types, and those for the top 6,000 words in conversation are have the highest recorded Dice distances for those ranks.

Figure 4.22

Projections from the Dice distance measures for the top-ranked 1,000, 2,000 and 3,000 word types for samples of all four key text types. (Points indicate where measurements were made. Trend lines are power law projections.)

The top-ranked 1,000 words. For the top-ranked 1,000 words, academic writing shows the most variability, followed by general and then narrative writing (Figure 4.23). Narrative writing would possibly catch up with conversation at a corpus size of 20 million tokens. The projection for conversation is tentative, however, since it is based on only three measurements.

Figure 4.23

Projected Dice distances for corpora of the four key text types for the top-ranked 1,000 words up to a sample size of 20 million tokens. (Points indicate where measurements were taken. Lines are power law projections.)

The top-ranked 3,000 words. For the top-ranked 3,000 words, general writing is significantly more consistent than the other text types. For general writing, the Dice distances at 10, 15 and 20 million tokens are empirical measurements rather than projections since the general writing collection is over twice the size of the collections for the other text types (Figure 4.24). For the top-ranked 3,000 words for a corpus of 20 million tokens, general writing should vary by around 3% (around 90 word types), whereas the other text types would likely vary by around 5% (around 150 word types).

Figure 4.24

Projected Dice distances for samples of the four key text types for the top-ranked 3,000 word types up to a sample size of 20 million tokens. (Points indicate where measurements were made. Trend lines are power law projections.)

The top-ranked 6,000 words. For the top-ranked 6,000 words at a sample size of 20 million tokens, general writing now varies by around 5% (Figure 4.25). The other text types show an additional 2% variability with conversation now trailing the other three text types.

Figure 4.25

Projected Dice distances for samples of the four key text types for the top-ranked 6,000 words up to a sample size of 20 million tokens. (Points indicate where measurements were taken. Trend lines are power law projections.)

Narrative writing and conversation. Narrative writing resembles conversation initially, but in comparisons of more than the first few thousand ranks, conversation becomes somewhat more variable (Figure 4.26). This is understandable since published narratives would be edited as opposed to the free ranging nature of conversation.

Figure 4.26

A comparison of Dice distance scores for narrative writing and conversation. (Points indicate where measurements were made. Trend lines are power law projections.)

General and academic writing. Academic writing is much more variable than general writing. For the top-ranked 3,000 word types and at a sample size of 20 million tokens, academic writing is projected to have a variability of around 5% or 150 word types (Figure 4.27). At that corpus size, the variability of general writing was measured at 5% for the top 6,000 ranks. Again, frequency lists for general writing are remarkably stable for many more word types than was initially expected.

Figure 4.27

A comparison of Dice distance scores for general and academic writing. (Points indicate where measurements were made. Trend lines are power law projections.)

General and narrative writing. Beyond the most frequent 1,000 word types, a larger number of proper nouns in narrative writing helps to make this text type less predictable than general writing (Figure 4.28).

Figure 4.28

A comparison of Dice distance scores for general and narrative writing. (Points indicate where measurements were made. Trend lines are power law projections.)

Narrative and academic writing. Narrative writing shows a similar degree of variability to academic writing beyond the top-ranked 1,000 words. This is especially interesting since the reading of narrative texts is often used as a transition to the study of more academic texts.

Figure 4.29

A comparison of Dice distance scores for academic and narrative writing. (Points indicate where measurements were made. Trend lines are power law projections.)

4.1.2 Dice distance with replacement

Section 3.3.1 describes how Dice distance was calculated between samples that were randomly selected from the available data for each text type. This methodology is limited, however, by the amount of data available for that text type. The largest sample size that can be measured is half the available data, and at that size only two samples can be compared. In order to generate more and larger samples, a common statistical technique called “sampling with replacement” was written into the Dice distance script.

Section 3.3.3 describes how samples were selected from all the data that had been collected for that text type. Each selection was made with replacement in the sense that all the files were available for each sample. Files selected for previous samples were not excluded as they were in the study described in 3.3.1 and reported in 4.1.1.

In this version of the study, 10 samples were selected for each comparison. This allows for the consistent reporting of the standard deviation (shown in dashed lines above and below the line showing the mean Dice distances). There was also a prospect of being able to compare slightly more than 50% of the available data.

Before sharing the results of the Dice distance study with replacement, some evaluation of the method is first necessary. Sampling with replacement unfortunately turned out to significantly lower the Dice distances between samples. The differences between the samples made with and without replacement correlate strongly with the size of the pool of data available for that text type. For smaller samples of 1 million tokens, the differences between the two studies are often negligible as can be seen in Figures 4.38, 4.41, 4.44, and 4.47. As sample sizes increase, a smaller pool of data, quickly leads to a significant overlap between samples, and therefore, a lower Dice distance. Since Dice distance is a measure of the similarity of texts, sampling with replacement is apparently not appropriate to statistically simulate a larger pool of available data.

To compare the results of the first version of the study (without replacement) and the second version (with replacement), the Dice distances for the top-ranked 100 to 10,000 words were determined for results that were available from both studies. In other words, the results of the two versions of the study were compared where the second version of the study replicated the first. The results were compared for the top-ranked 1-3, 6, 9 and 10,000 word types. A simple linear regression was then plotted to estimate the difference that would be expected for the largest sample measured with replacement. For example, for conversation, comparisons were made without replacement for samples of 1-3 million tokens, but samples of 1-5 million tokens were compared with replacement. Therefore, linear regression was used to extrapolate forward a further 2 million tokens to estimate the differences that would be likely if comparisons of 5 million tokens without replacement had been possible for conversation as well. This should approximate the amount of error in the projected Dice distances for the study with replacement.

The relatively small size of the conversation corpus resulted in a significant difference in the Dice distance of nearly 5.9 for the top 10,000 words (Figure 4.30). This means that if a long-range projection were to be made for the top 10,000 words of conversation from the Dice study with replacement, the projections would likely need to adjusted upward by a Dice distance of approximately 5.9. On the other hand, for general writing, with its 47.65 million tokens, the effect was minimal. Even for the top 10,000 words, the estimated error is just a little over 0.01. Narrative writing with 18.42 million tokens and academic writing with 17.91 million tokens produced differences between the other conversation and general writing.

These results imply that one should only use the technique of sampling with replacement to increase the number of available samples, not to enlarge the feasible size of those samples. The amount of likely error in the results is shown in Figures 4.30-35. The results of the Dice distance with replacement study in sections 4.2.1-4 should be interpreted in light of this information. Otherwise the results would sometimes appear far too optimistic.

Figure 4.30

Error estimates from sampling with replacement when measuring Dice distances for the top 10,000 words.

Differences for the top 1,000 word types vary from less than 0.005 for general writing to almost 0.02 for academic writing (Figure 4.31). Some of the mean Dice distance scores between the samples of academic writing with replacement are actually lower than those without replacement, hence the negative scores. In other words, the study with replacement found the samples of academic writing of 1 and 2 million tokens to be less similar to each other than the study without replacement had found. This is not very surprising. These small differences of less than 0.004 are within the mean standard deviation for samples of that size (Figure 4.45).

NB: The mean standard deviation is the mean of the standard deviations for each measurement of Dice distance. Dice distance is measured in increments of 100 or 1,000 ranked word types depending on the graph. Each measurement represents a cross comparison of 10 samples. The standard deviation was calculated for each increment and the mean for all of these standard deviations is what is reported as the mean standard deviation for that sample size.

Figure 4.31

Error estimates from sampling with replacement when measuring Dice distances for the top 1,000 words.

For the top 2,000 word types, general writing now differs by approximately 0.007 and narrative writing by 0.01 (Figure 4.32). Conversation and academic writing are highest at almost 0.023.

Figure 4.32

Error estimates from sampling with replacement when measuring Dice distances for the top 2,000 words.

For the top 3,000 word types, the effect of the small quantity of data for conversation is seen more clearly with a difference of 0.028 (Figure 4.33). Academic writing is just barely ahead of where it was for the top 2,000 at 0.023. The difference for narrative writing has moved up to 0.013, while general writing has remained just below 0.007.

Figure 4.33

Error estimates from sampling with replacement when measuring Dice distances for the top 3,000 words.

For the top 6,000 words, the difference for conversation has risen dramatically to almost 0.043 (Figure 4.34). Academic writing has barely changed at just above 0.024. Narrative writing is just shy of 0.02, and general writing has only grown to around 0.009.

Figure 4.34

Error estimates from sampling with replacement when measuring Dice distances for the top 6,000 words.

The Dice distances for the top 9,000 words is the final set of measures used for projections in section 4.1.2.5. Here, conversation differs by 0.054, academic writing by about 0.031, narrative by a little more than 0.022 and general writing just over 0.01 (Figure 4.35).

Figure 4.35

Error estimates from sampling with replacement when measuring Dice distances for the top 9,000 words.

4.1.2.1 Conversation. As noted in section 4.1.2.0 and seen in Figures 4.38 below, the measurement of the Dice distances for conversation suffers due to the small quantity of data available. One purpose in sampling with replacement is to allow for the consistent reporting of the standard deviation. For samples of 1 million tokens of conversation, the standard deviation is 0.007 (Figures 4.36-37). The standard deviation can be seen to narrow as samples sizes increase, but this may be partly due to the increasing overlap of the samples. Of course, the standard deviations for larger sample sizes should not exceed the standard deviation for the samples of 1 million tokens.

Figure 4.36

Mean Dice distances for 10 samples of 1-5 million tokens of conversation with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.37

Mean Dice distances for 10 samples of 1-5 million tokens of conversation with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types. Dashed lines indicate $\pm$ the standard deviation.

In each case, the mean Dice distances for the samples with replacement are markedly lower than for those without replacement (Figure 4.38). The distances for the 2-million-token samples with replacement correspond almost perfectly with the distances for 3 million tokens without replacement.

Figure 4.38

A comparison of the Dice distance measurements for conversation for samples of 1, 2 and 3 million tokens with replacement (dashed lines) and without (solid lines).

4.1.2.2 Narrative writing. For samples of 1 million tokens of narrative writing, the mean standard deviation is 0.0062 (Figures 4.39-40). The final comparison of samples of 10 million tokens is slightly more than 50% of the available 18.42 million tokens that were collected for narrative writing. The discrepancy at a sample size of 9 million tokens of narrative writing (Figure 4.41) is still within the standard deviation of for samples of 1 million tokens.

Figure 4.39

Mean Dice distances for 10 samples of 1-10 million tokens of narrative writing with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.40

Mean Dice distances for 10 samples of 1-10 million tokens of narrative writing with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.41

A comparison of the Dice distance measurements for narrative writing for samples of 1, 2, 4 and 9 million tokens with replacement (dashed lines) and without (solid lines).

4.1.2.3 General writing. For samples of 1 million tokens of general writing, the mean standard deviation is 0.0055 (Figures 4.42-43). With its larger pool of available data, general writing is only text type for which sampling with replacement does not cause substantial differences in the estimated Dice distances (Figure 4.44). A script that calculates Dice distance with replacement is then a convenient way of randomly generating a 10x10 comparison.

Figure 4.42

Mean Dice distances for 10 samples of 1-10, 15, 20 and 25 million tokens of general writing with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.43

Mean Dice distances for 10 samples of 1-10, 15, 20 and 25 million tokens of general writing with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.44

A comparison of the Dice distance measurements for general writing for samples of 1, 5, 10 and 20 million tokens with replacement (dashed lines) and without (solid lines).

4.1.2.4 Academic writing. For samples of 1 million tokens of academic writing, the mean standard deviation is 0.0084, the largest mean standard deviation of the four key text types (Figures 4.45-46). For the 1- and 2-million-token samples, the discrepancy between the two methods of sampling is not pronounced for the top 3,000 words (Figure 4.33). Academic writing had the second smallest data collection (17.91 million tokens), which becomes evident at larger sample sizes. At 8 million tokens, the Dice distance is already underestimated by around 3% (Figure 4.47). Samples of 9 and 10 million are beyond 50% of the available data and would be even less reliable.

Figure 4.45

Mean Dice distances for 10 samples of 1-10 million tokens of academic writing with replacement. Measurements were made in increments increasing by 100 words up to a rank of 10,000. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.46

Mean Dice distances for 10 samples of 1-10 million tokens of academic writing with replacement. Measurements were made in increments increasing by 1,000 words up to the maximum vocabulary size shared by all 10 samples rounded to the last full 1,000 types. Dashed lines indicate $\pm$ the standard deviation.

Figure 4.47

A comparison of the Dice distance measurements for academic writing for samples of 1, 2, 4 and 8 million tokens with replacement (dashed lines) and without (solid lines).

4.1.2.5 Projections of Dice distances with replacement. Projections have been plotted to a sample size of 50 million tokens for each of the four key text types based on the results of the Dice distance study with replacement. These estimates shown should be adjusted higher based on the analysis in section 4.1.2.0.

Conversation. The projection to 25 million tokens seen in Figure 4.21 shows the mean Dice distances for conversation ranging from around 0.02 for the top 1,000 words to 0.085 for the top 9,000. In the study with replacement, all the projections have been compressed to below 0.03 at the 25-million-token level (Figure 4.48).

Figure 4.48

Projected Dice distances for samples of conversation up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.

Narrative writing. The projection to 25 million tokens seen in Figure 4.19 shows the mean Dice distances for narrative writing ranging from around 0.02 for the top 1,000 words to about 0.07 for the top 9,000. In the study with replacement, the projections fall in a narrower band between 0.01 and 0.04 at the 25-million-token level (Figure 4.49).

Figure 4.49

Projected Dice distances for samples of narrative writing up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement.

Projections have been plotted for the top 1-3, 6 and 9,000 words.

General writing. The projection to 50 million tokens seen in Figure 4.18 shows the mean Dice distances for general writing ranging from around 0.02 for both the top 1,000 and 2,000 words to just under 0.04 for the top 9,000. In the study with replacement, the projections fall between 0.015 and 0.03 (Figure 4.50).

Figure 4.50

Projected Dice distances for samples of general writing up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.

Academic writing. The projection to 25 million tokens seen in Figure 4.20 shows the mean Dice distances for academic writing ranging from slightly higher than 0.03 for the top 1,000 words to about 0.075 for the top 9,000. In the study with replacement, the projections still start around 0.03, but only rise a little above 0.04 for the top 9,000 at the 25-million-token level (Figure 4.51).

Figure 4.51

Projected Dice distances for samples of academic writing up to a size of 50 million tokens based on the mean distances between 10 samples selected randomly with replacement. Projections have been plotted for the top 1-3, 6 and 9,000 words.

Based on the estimates shown in Figure 4.31, the projections in Figure 4.52 need to be adjusted upwards by approximately 0.005 for general writing, 0.008 for narrative writing, 0.017 for conversation and almost 0.02 for academic writing.

Figure 4.52

Projected Dice for top 1,000 words for the four key text types up to a size of 50 million tokens.

Based on the estimates shown in Figure 4.33, the projections in Figure 4.53 should probably be adjusted upwards ranging from almost 0.007 for general writing, to 0.013 for narrative writing, to 0.23 for academic writing, to almost 0.027 for conversation.

Figure 4.53

Projected Dice for top 3,000 words for the four key text types up to a size of 50 million tokens.

The projections in Figure 4.54 show all three written text types clustered together with conversation being even more homogeneous. This is unrealistic. General writing should be adjusted upwards by almost 0.009, narrative writing by 0.02, academic writing by about 0.025, and conversation needs to be raised by almost 0.043 based on the estimates shown in Figure 4.34.

Figure 4.54

Projected Dice for top 6,000 words for the four key text types up to a size of 50 million tokens.

Figure 4.55 shows an even tighter overlap of all three written text types. Rather than being the most variable text type, as it is in Figure 4.21, conversation is projected as having an average Dice distance of only 0.015. This estimate should be raised by 0.054, academic writing by 0.03, narrative writing by 0.021, while general writing should only need to be adjusted by a 0.01 based on the estimates shown in Figure 4.35.

Figure 4.55

Projected Dice for top 9,000 words for the four key text types up to a size of 50 million tokens.

4.2 How different are the text types?

4.2.1 Dice distance of the key text types from each other

The Dice coefficient was also used in this study to answer the research question, “How different are the high-frequency words of each of the key text types?” By looking at the amount of overlap between the text types, list designers can determine if distinct oral and written core lists are needed, or if some other configuration is desirable.

In the first study, samples were compared with other samples of the same text type. In this study of the differences between the four key text types, 10 samples of 10 million tokens each were used in a 40x40 cross-comparison that is described in section 3.4.2. Samples of 10 million tokens were chosen as a reasonably stable quantity that was feasible for all four text types. (See section 3.4.2 in regard to conversation.) Frequency lists were created for each of

these 40 samples (10 each per text type), and the Dice similarity was calculated between each of them (780 cross-comparisons) beginning with the most frequent 500 word types in each text type. Twelve comparisons were made in all from the most frequent 500 word types in each text type to the most frequent 6,000 in increments of 500 word types.

Figure 4.56 shows four dimensions – one for each of the key text types. The average Dice similarity scores for the four text types is plotted for the most frequent 1,000 words in each of the four text types. (See the “Text type comparison matrix analysis” spreadsheet in the attached supplemental files for the matrix and all 12 diagrams.)

Dice similarity is the complement of Dice distance. Two items which have a Dice distance of 0.2 have a Dice similarity of 0.8. That is, they differ by 20%, and are therefore 80% similar. A point closer to one of the four corners of the diagram, indicates greater similarity to that text type. A point at the exact center would indicate 100% dissimilarity.

Figure 4.56

Radar diagram showing the Dice distances between the top 1,000 words of each of the key text types.

4.2.1.1 Conversation. The most frequent 1,000 word types in samples of conversation are obviously very similar to each other so one of its points is very close to the corner representing conversation. They also share approximately 70% of the words found in the top

1,000 words in narrative writing texts, and nearly 60% of those in the general writing texts. Conversation shares only a little more than 40% of the word types in the top 1,000 in samples of academic writing.

4.2.1.2 Narrative writing and conversation are roughly the same distance from general (green diamond) and academic writing (purple circle). Notice the line from conversation to narrative writing is parallel to the grid. Conversation and narrative writing themselves differ by around 30%. Many of the differences between narrative writing and conversation stem from the fact that word types rather than lemmas or family lists were used for this comparison. Simple present tense verbs are the more common in conversation, while simple past forms predominate in narratives. So conversation and writing are quite similar despite one text type being oral and the other written.

4.2.1.3 General writing is truly general being around 58% similar to conversation and narrative writing. It is closest, however, to academic writing overlapping by around 70%. General writing is reasonably close to conversation and narrative writing, but even more similar to academic writing with which it shares an informational focus.

4.2.1.4 Academic writing is a specialized informational text type. It is the most distant from narrative writing and conversation.

Biber's (1988, 1995) multidimensional studies found "a fundamental opposition between 'oral' registers and 'literate' registers" (Biber & Conrad, 1990: 256). This dichotomy can be seen in the way the two dialogic text types gravitate to one side, while the two that represent written exposition pull to the other. Notice how the lines for conversation and narrative writing cross closest to their shared side of the matrix in Figure 4.56, while the lines for general and academic writing cross closest to their shared side of the matrix.

This contrast between informational and interactional texts can be seen in the top few words of each text type (Table 4.3). General and academic writing represent carefully edited texts and their communicative purpose is largely informational. Conversation, on the other hand, is interactional taking place in real time. In between, these text types are texts that are written-to-be-spoken or contain written representations of speech. Brysbaert and New's (2009) SUBTLEXus corpus (discussed in section 2.5.12) is composed of subtitles from TV and film. The OUP graded readers and the collection of imaginative narrative texts used in this study contain large quantities of simulated dialogue.

Together, these six collections form a spectrum from a largely interactional to an overwhelmingly informational purpose. The ranking of personal pronouns and possessives illustrates this gradation. The graded readers and subtitles have one more personal pronoun or possessive than authentic conversation, but lack response forms like *yeah* and *oh*. General

writing has only two such items, and academic writing has none. The pronouns *I* and *you* are most indicative of interaction. These two occupy the top two ranks in conversation and the subtitle corpus. In the graded readers, they slip to third and eighth place. In ungraded narrative writing, they are sixth and twelfth. In general writing, *I* is in twenty-second place and *you* is somewhere below that.

Table 4.3

The top-ranked 25 word types on the complete frequency lists for the collections of the four key text types, a collection of OUP graded readers and Brysbaert and New's (2009) SUBTLEXus corpus.

rank	Academic	General	Narrative	OUP readers	SUBTLEXus	Conversation
1	the	the	the	the	you	I
2	of	of	and	and	I	you
3	and	to	to	I	the	the
4	to	and	a	to	to	it
5	in	a	of	a	's	's
6	a	in	I	he	a	and
7	is	is	he	was	it	that
8	that	for	was	you	't	to
9	for	that	in	of	that	a
10	be	was	she	in	and	't
11	as	it	it	it	of	yeah
12	it	on	you	that	what	of
13	by	's	her	's	in	in
14	with	with	that	she	me	they
15	was	as	his	said	is	he
16	this	be	's	but	we	oh
17	are	by	had	his	this	is
18	on	at	on	her	he	know
19	which	he	with	at	on	what
20	or	are	't	had	for	was
21	not	from	for	't	my	no
22	's	I	as	for	'm	we
23	from	this	at	me	your	but
24	an	have	but	on	don	have
25	have	but	said	with	have	well

By rank 6,000, all four text types have moved toward each other, but the basic configuration is maintained. Conversation now shows the greatest internal variation as can be seen in its distance from its own corner of the diagram.

Figure 4.57

Radar diagram showing the Dice distances between the top 6,000 words of each of the key text types.

A dendrogram helps to better visualize the overall relationship between the text types by reducing the relationships to two dimensions (Figure 4.58). Conversation and narrative writing are the closest. General writing is a step further away from these two text types, and academic writing is even more distant from this grouping. A dendrogram of the top 6,000 words is almost identical except that on the whole the text types are somewhat closer together, while academic writing is relatively more distant. The four text types, in some ways, resemble a set of Russian Matryoshka dolls with conversation at the center, followed by narrative, general and academic writing.

Figure 4.58

A dendrogram showing the hierarchical clustering of the key text types using the complete-linkage or furthest-neighbor method based on the Dice distances between the top 1,000 words of each of text type.

4.2.2 Analysis of word types shared by or unique to each text type

The analysis presented in 4.2.1 shows the quantity of word types shared by each of the text types. An examination of the actual word types that are shared by different text types or that are unique to one text type can give depth to that analysis. For this exploration of the vocabulary of each key text type, a list was made for each text type of all the word types which occurred among the most frequent 6,000 word types in at least one of the 10 samples of a that text type. These are the same samples that were used in the study reported in section 4.2.1.

In order to move this analysis closer to a pedagogically useful form, marginal words and proper nouns were first removed from each text type vocabulary list (See section 1.4.7). An analysis of the distribution of marginal words is found in section 4.2.2.1. Section 4.2.2.2 will examine word types that were only found among the most frequent 6,000 word types in one of the key text types. Section 4.2.2.3 will explore the overlap in vocabulary of these four text types. Section 4.2.2.4 will return to a numerical look at the overlap between each of the four text types.

After sorting out marginal words and proper nouns, a total of 11,305 word types remained that were among the most frequent 6,000 word types for at least one sample of one text type. (See the “Text type top 6000” spreadsheet in the attached supplemental files.) Of these, 2,011 word types are common to all four key text types. These common words would not, however, make a good core vocabulary list. This list would still lack a large number of very basic words which simply do not occur in high frequency in the more literate text types.

4.2.2.1 Distribution of marginal words. Marginal words are a typical feature of spoken language making up 2% of the top 3,000 types in conversation (Table 4.25). This is somewhat circular logic, however, since the fact that they do not usually have established written forms or do not fit the lexical categories developed primarily from the study of written language are key reasons for considering them marginal. Marginal types are also common in narrative writing, but more consistently spelled versions predominate. The type *oh* occurs 700 times per million and *ah* 134 times per million in narrative writing.

Marginal words are present in general and academic writing, but they are very infrequent. They typically occur in articles from the humanities that are discussing features of spoken language. The lists of the most frequent types in general and academic writing do not include any true marginal words. One word type, *mm*, that is a member of the family *AH* in the BNC/COCA lists, is found in the most frequent 2,000 types of academic writing and the top 3,000 types of general writing. However, in general writing *mm* is the abbreviation for *millimeter*. In academic writing, it is also frequently the abbreviation for *millimolar*.

One problem with this method of marking marginal words is how frequently these very short word forms can have multiple interpretations, especially when case is ignored. *ERM*, for example, is used in academic writing in a discussion of fillers in speech, but elsewhere it is an abbreviation for *Exchange Rate Mechanism*. In general and academic writing, *HM*, *ER* and *UM* do not indicate hesitations, as they do in conversation, but are abbreviations for *Her Majesty's*, the *All England Law Reports* and μm , the abbreviation for *micrometer* when a Greek font is not available. *WAH* is sometimes an interjection, but in the academic writing portion of ICE Hong Kong, it is a proper noun. For this reason, possible marginal forms had to be investigated manually using Laurence Anthony's *Antconc* concordancer (Anthony, 2011).

If marginal words were not eliminated from consideration, a second problem would arise. Grouping marginal words into families is extremely problematic, so Nation and Webb (2011) opted for the use of a few very large marginal word families. This was a practical solution to dealing with many amorphous and ambiguous marginal word forms. *AH* can be spelled with any number of *A*'s and *H*'s, and it can function as an interjection, a voiced pause,

a response form, etc. The conversation data collection in this study even contained six additional marginal word types not found in Nation and Webb's (2011) extensive list: *MHH*, *MHHM*, *UHM*, *UHN*, *UHU*, *UNHUNH*. Placing these many marginal types in a few very large bins is much more convenient, but such large word families quickly accumulate a large number of tokens. If they are not excluded, they quickly rise to the top when sorting families by frequency.

NB: The letters of the alphabet are also part of Nation and Webb's (2011) base word 32. These were removed from the counts of marginal words. As an interesting test of coreness, only conversation contains all 26 letters in the top 1,000 types.

4.2.2.2 Words unique to the high-frequency word lists for each text type. (See the "Words unique to each text type" file in the attached supplemental files.) Word types that are among the most frequent word types in a particular text type, and yet are unique to that text type, should give a good indication of the nature of that type of language use.

Conversation includes an eclectic group of 735 words that were not among the 6,000 high-frequency words of the other text types. There are still many proper nouns and some expletives that are not on Nation and Webb's (2011) marginal word list. The focus of conversation is domestic or everyday life, as seen in such words as *dishwasher*, *kitten*, *loo*, *pram*, *landline*, *Google*, *email*, *internet*, *bookshop*, *chemist*, *prescription*, *postoffice*, *playschool*, *downtown* and *milkman*. Many terms are related to leisure, e.g. *mahjong*, *javelin*, *karaoke* and *volleyball*. There are some also academic terms related to school or university: *dormitory*, *fractal*, *triangle*, *tyrannosaurus*, *undergraduate*, *tuition* and even *dissertation*. By far the largest category are words related to food, such as *banana*, *carrot*, *cider*, *hamburger*, *onion*, *pizza*, *mango*, *marzipan*, *marmite*, *saucepan*, *tomato*, *vinegar*, *yam* and *yoghurt*.

Some of the terms unique to conversation are noticeably regional. There are several terms even among the top 1,000 of conversation which highlight the need for a dispersion measure, such as Gries' DP or Jullian's D (Gries, 2010c). *Ta*, for example, is listed in Nation and Webb's BNC/COCA lists as an abbreviation, but in conversation *ta* is more often used in the sense "thanks" than as "T.A." Another UK regional type is the nonstandard contraction *int* (*isn't*). The particle *lah* in Singapore English is a special case. It is found in only one corpus, but it is so frequent that it is among the top 500 words in 7 of the 10 random samples for conversation. Indian English also contributes the tag question *na*, the response form *haan* "yes" and *yaar* "mate." *Ano* "what?" and *kasi* "because" are among the top 1,000 due to their frequency in the Philippine corpus.

Narrative writing. The 511 words unique to the high-frequency list for narrative writing are unmistakable. Besides some additional proper nouns and two foreign titles, *Herr*

and *Fräulein*, the list contains a wide range of verbs relating to speech or other vocalizations: *bellow, bicker, chuckle, confess, giggle, glance, glare, overhear, exclaim, grunt, interrupt, mock, mumble, wail, weep* and *yell*, among others. Verbs describing bodily movement are also frequent, e.g. *clench, flinch, fling, grin, heave, kneel, quiver, nod, shudder, stoop, stride, tug, tremble, twitch* and *wince*. Words related to characters, their emotions, clothing and properties abound: *abbot, brigadier, elf, friar, matron, senator, warrior, eyelid, forehead, skull, thigh, arrogant, ecstasy, horrified, solemn, briefcase, dagger, rifle, robe*. Descriptive adjectives are understandably common: *exquisite, feeble, ragged, scarlet*. There is also an interesting assortment of settings: *compartment, doorstep, graveyard, hallway, mantelpiece, meadow, saloon, submarine, tavern* etc.

General writing earns its title having only 168 words that are unique to its most frequent 6,000. They cover a diverse range of topics: business (*businessman, dairy, dividend, fisherman, freight, takeover, petroleum, privatise, salesperson*), hobbies (*aquarium, tablespoon, weave, yarn*), medicine (*casualty, outbreak, syphilis*), military (*artillery, brigade, generals, missile, regiment*), politics (*ballot, delegate, deficit, faction, premier, presidency, shah, summit*), science and technology (*beta, desktop, enthalpy, mainframe, upgrade, quantum, spreadsheet, workstation*), sports (*tournament, goalkeeper, innings, midfield, referee, scrum, wicket*), religion (*papal, theological*) and the natural world (*dolphin, foliage, lava, whale, wildlife*).

General writing's unique high-frequency word types include 14 letter-digit combinations: the ordinals *1st* plus *17th* through *20th* (the latter set usually referring to centuries), the measures *1m* through *8m*, and one related to scheduled time, *30pm*. This last item was left after the R-script deleted the hour and colon from sequences, such as *1:30pm*, since there are no letters connected to the digit *1* when the colon is counted as a word boundary. As was seen section 2.5.1, writing systems were almost universally developed for accounting and bureaucratic record keeping. Here there is a focus on centuries. Section 4.2.3.3 will show a focus on the annual calendar. In general writing, there is also a contrasting focus on exact hourly times.

Academic writing is second only to conversation with 665 unique terms. About 10% are basic academic vocabulary found on the Academic Word List (Coxhead, 2000), such as *analogy, dynamics, hypothetical, parameter* and *simulate*. The others range from reasonably common academic words, like *ecology, microscope, prose, pronoun, quantify, rainfall* and *textbook* to truly technical terms, like *deixis, eigenvalue, fiduciary, indomethacin* and *oligonucleotide* that are specific to one discipline, c.f. Nation (2001: 198). Three combinations of letters and digits are among the highest frequency types in academic writing:

NCp7, ATF1 (both designating proteins) and *1940s*. All the other decades from the 1920s through the 1990s are present among the most frequent words of general writing. When arranged by frequency, the decades also occur in chronological order from the present to the past, except for *1990s* which postdates most of the corpus data used in this study. The decade designated as *1940s* may seem a curious item to be uniquely academic, except when one considers that each of the other decades is frequently referred to numerically: the 1920s, 1950s etc. Each evokes a recognizable set of cultural and historical associations, but one does not typically refer to the 1940s as “the 40s.” One event obscures everything else – the War. Only academic writing regularly refers to the war years dispassionately as the 1940s.

4.2.2.3 Vocabulary overlap among the high-frequency words of each text type.

The words that are unique to each text type exemplify the nature of each text type. Words that are shared by more than one text type offer a glimpse into how these four large branches are connected in the lexicon. This section will examine the number and type of words shared by more than one text type and discuss what that indicates about the relationship between the text types.

The most frequent 6,000 word types of each of the four key text types comprised a combined vocabulary of 11,305 word types. Of those, nearly a quarter were common to all four text types. This collection includes very basic vocabulary, but since this comparison spanned the top 6,000 words, it also includes many markedly academic terms as well.

The 800 items shared by conversation, narrative and general writing are much more uniformly basic vocabulary (Table 4.4). One then sees two major branches separate. There are another 726 words that are shared by conversation and narrative writing. These are still predominately colloquial dealing with domestic concerns. Contractions are frequent in both the vocabulary shared by conversation and narrative writing and in the vocabulary shared by conversation, narrative and general writing.

Branching off in a more educated, literate direction are the text types of general and academic writing. This is a good-sized list of 1,147 essential terms from the world of business and policy decisions. A smaller group of 369 words is shared by all three written text types. These are terms describing situations, individuals, their actions and emotions. These are all terms that are useful in relating narratives. A further 119 slightly less academic words are shared by only general and narrative writing.

If Table 4.4 were three-dimensional, the 201 words common to conversation and general writing would branch off from the 800 words shared by conversation, narrative and general writing. These terms are part of the general vocabulary of an educated adult. There is

noticeable focus on leisure activities in this group of words, which is typical content for many genre of general writing.

Finally, there are two groups that could be described as examples of intertextuality. Narrative writing shares 51 words with academic writing where narratives adopt a slightly more academic tone. Academic vocabulary also finds its way into casual conversation. A suspicious number of the 87 words shared by conversation and academic writing have to do with linguistics. This may be related to how the conversation data was collected (See section 3.2.6).

Table 4.4

Samples of the vocabulary shared by more than one text type. (The number of word types shared by that group of text types is given in bold print.) Examples of words unique to each text type have already been discussed above in section 4.2.2.2.

Conversation	Narrative	General	Academic
2,696 a about after all already always and another are as at back because been before best between both but by called can children come could course day did different do down each end even every fact family far few find first for from full give good great hand has he her here high him his home house how I if in into is it just know later least left life like little long make man many may might more most much must not now once other out over part people place point quite rather real right same say school second see she should since small so something still such take than that the their them then there these they this those though thought three through time times under until up us use used very was way we well were what when where which while who whole why work profile previously owners pursuit rejected suspended scenes possibilities governor reserved fluid ignored waves surrounded premises urgent deny slope superior threatening encouraged curve arrest domestic intimate behalf protected severe pupils modest depression ensure elaborate increasing sole beings fraction delayed merchant defeated obscure referring represented documents largest deputy believes encouraging witnesses columns			
800 couldn please nice tea wouldn maybe dear sorry haven thank tomorrow pretty phone dinner stuff dad tonight lovely aren weren mum christmas guy yesterday everybody funny hasn Saturday anybody Friday guess bet kitchen walked hadn cup afternoon wine thanks picked busy watching lunch liked worry bag terrible bottle bedroom tired laugh shouldn cool shoes holiday fun wonderful stick lucky bread excuse awful yours ok silly stupid bother weekend likes brilliant birthday shopping definitely bigger cake chicken sing eighteen mate chocolate hi guys dollars coat wash cheese movie missed cream afraid wedding kid wearing breakfast			
726 yeah hello okay aye mummy gonna cos ninety dunno bye daddy hey darling upstairs forgot reckon horrible pardon bloke eighty quid rang bothered weird phoned mom honestly telly stairs laughing downstairs fridge grandma asleep cupboard biscuits lad toast joking scared pie salad biscuit ate terribly socks hers shout mustn whoever skirt jeans pudding pity dare chat sandwich smells nanny raining swim conversations rung woke cups kettle jam anyhow awake sandwiches auntie teddy shouting tidy wipe slept bite remind bunch jumping lately lads cakes		1,147 elections democratic profits associated provides applications officials markets estimated proposals legislation billion sources reform leadership additional provision strategy suggests substantial establish governments revenue distribution expenditure developments requirements readers proportion effectively representatives negotiations currency consumer unions assets ruling supreme supporters organisations losses obtained principles corporate parliamentary decade directors defined urban decline gulf monetary contemporary crucial increases recovery	

	369 thrust senses alien distress butler rounded revealing believing deliberate respectable enquiries refuge destination paths confronted examining practised positively chains protective shield doubtful acknowledge beneath anger remained unable repeated deeply firmly demanded moments distant sharply begun bishop returning observed arrival enemy fought servants chamber emerged reflected		
	119 gesture lightly relieved clouds instantly companion assured precious detective delicate halt fierce passengers charm elegant mysterious guards safely lovers lieutenant wounded dignity passenger fled manor canvas commander exploded magnificent		
	51 imagined sensation evidently veins unconscious owl heir →		fragments traced subsection indicating unaware irony caution masculine
201 rugby batteries skiing chip battery pays lectures recipe choir chemistry garlic comedy kilometres guitar cancelled tremendous freeze trips flavour ballet chopped →		anniversary surgeon flour cheques adjust gardening drivers celebrate litre comic housework cruise scored youngsters knit soccer helicopter teenagers greenhouse calendar organise	
87 dictionary linguistics faculty eleventh patois assignment essays syllabus vacuum →			graduates dialect injection transcription corpus tertiary analyse slides vitamin gel sociology symptom technically
735 unique types (section 4.2.2.2)	511 unique types (section 4.2.2.2)	168 unique types (section 4.2.2.2)	665 unique types (section 4.2.2.2)

4.2.2.4 Quantifying the overlap of vocabulary between the key text types. One key issue in constructing a core vocabulary list is how much coverage it will afford in different situations (See Lee, 2001: criterion #5 in section 2.4). As has been seen in the previous section, there is some amount of overlap between each of the text types. This section will detail how many of the most frequent word types in each text type are covered by each of the other text types and how many of those word types are missing. It will then describe the way the vocabulary of each text type and groups of text types overlap. The pattern that will be seen in tables 4.5-8 is related to the dendogram seen in Figure 4.58 above.

In Table 4.5, the dark-shaded central line shows the number of word types which occurred in at least one of the ten samples of 10 million in that text type. The total vocabulary for a single text type can exceed 6,000 types since the word list for each text type is the sum of words from 10 samples. Conversation interestingly has the largest vocabulary (6,521 word types) even after marginal words and proper nouns have been removed. Academic writing is close behind with 6,507 word types. Next is general writing with 6,388 word types. Narrative writing at 5,943 word types, actually falls below 6,000 after the removal of proper nouns.

In regular type in the bottom half (below the central diagonal line) of the Table 4.5 are the number of words common to each row-column pair of text types, i.e. their shared vocabulary. Below these is the quantity (**shared vocabulary – x**) i.e. “How many word types are lacking in y (row) to cover x (column)? That means that of the 6,521 word types found in the samples of conversation, 4,254 of those were also in samples of narrative writing. An additional 2,267 word types would then be needed to cover the vocabulary of conversation. As an additional example, one can see that the vast majority of the 6,388 word types in general writing are covered by academic writing. Besides the 4,828 word types shared by general and academic writing, only an additional 1,560 word types would be needed to cover the vocabulary of general writing.

In the upper half (above the central diagonal line) of the Table 4.5 in italics is each *pair-wise sum* (the total vocabulary of each pair of text types). Below each sum in bold print is the quantity (**y – the shared vocabulary**), which answers the question “How many word types would y (row) add to x (column)? For example, the combined vocabulary of conversation and narrative writing is 8,210 word types with narrative writing adding an additional 1,689 word types. Being very dissimilar conversation and academic writing would have a combined vocabulary of 9,596 word types, almost a third of which (3,075 word types) would be word types added by academic writing.

Table 4.5

Matrix showing the quantity of shared or differing vocabulary for pair-wise comparisons of the 6,000 most frequent word types in each of the four key text types.

	Conversation	Narrative	General	Academic
Conversation	6,521	8,210	8,596	9,596
		1,689	2,075	3,075
Narrative	4,254	5,943	8,347	9,302
	-2,267		2,404	3,359
General	4,313	3,984	6,388	8,067
	-2,208	-1,959		1,679
Academic	3,432	3,148	4,828	6,507
	-3,089	-2,795	-1,560	

The largest pair-wise shared vocabulary is that of general and academic writing with 4,828 words in common. Conversation and narrative writing are next, but would have been much larger if proper nouns had been included. One again sees the lexical trunk of the language dividing between these two major branches of information and interaction.

The next three tables (Tables 4.6-8) explore how one might combine the vocabularies of these text types. If the two text types with the smallest Dice distance (conversation and narrative writing) were joined, the combined high-frequency vocabulary would be 8,210 words (Table 4.6). This grouping is then closest to general writing, which would add another 1,587 words for a total of 9,797 word types. General writing shares 4,801 word types with this conversation-narrative writing pair, but would need an additional 3,409 word types to cover this pair's combined vocabulary.

Table 4.6

Matrix showing the quantity of shared or differing vocabulary for general and academic writing vis-à-vis the combined vocabulary of conversation and narrative writing.

	Conversation & Narrative	General	Academic
Conversation & Narrative	8,210	9,797	8,596
		1,587	2,075
General	4,801	6,388	8,067
	-3,409		1,679
Academic	3,852	4,828	6,507
	-2,208	-1,560	

As seen in Table 4.5, since proper nouns are not included, the vocabularies of academic and general writing are even slightly closer to each other than conversation and narrative writing are to each other. Notice that general writing only adds 1,679 word types to academic writing, and academic writing only lacks 1,560 of the word types in general writing. If one were to pair general and academic writing (Table 4.7), this pair would be slightly further away from the other two text types.

Table 4.7

Matrix showing the quantity of shared or differing vocabulary for conversation and narrative writing vis-à-vis the combined vocabulary of general and academic writing.

	Conversation	Narrative	General & Academic
Conversation	6,521	8,210	10,155
		1,689	2,088
Narrative	4,254	5,943	9,943
	-2,267		1,876
General & Academic	4,433	4,067	8,067
	-3,634	-4,000	

As mentioned above, if one were to combine the three more general text types (conversation, narrative and general writing), this grouping would have a combined vocabulary of 9,797 word types, and it would lack only 1,508 of the most frequent academic words (Table 4.8). Though these are word types rather than word families, this total seems

close to the number of words Nation & Newton (2009: 133) and Nation (2006) suggest are necessary for ungraded reading and especially listening.

Table 4.8

Matrix showing the quantity of shared or differing vocabulary for academic writing vis-à-vis the combined vocabulary of conversation, narrative and general writing.

	Conversation, Narrative & General	Academic
Conversation, Narrative & General	9,797	11,305 1,508
Academic	4,999 -3,211	5,943

One can now see a regular pattern of 1,500-1,700 words (out of the most frequent 6,000) separating conversation from narrative writing (1,589 word types, Table 4.5), and this pair from general writing (1,587 word types, Table 4.6). Combining these three text types is a similar distance from academic writing (1,508 word types, Table 4.8). This shows the origin of the dendrogram in Figure 4.58.

4.2.3 Where is the core?

In this section, it remains to test the hypothesis that the high-frequency words of the conversation text type are roughly equivalent to the core vocabulary. In section 4.2.3.1, the 10 samples of each text type used in section 4.2.1 above will be compared with a survival vocabulary (Crabbe & Nation, 1991) and in 4.2.3.2 with a kernel vocabulary (Swadesh, 1971: 283) used for historical linguistic research. Section 4.2.3.3 will examine the distribution of the words in several lexical sets across the text types, and 4.2.3.4 will conclude by looking at the words mentioned as being problematic in Nation (1990: 20-21).

In calculating the Dice distance between the text types (section 4.2.1), the 500 most frequent words in 10 samples of each text type were compared with each other. The most frequent 1,000 words were then compared, 1,500 etc. up to the most frequent 6,000 words. At each stage a word could be present in up to 10 samples in any given text type. In the analyses below, word types were sorted according to the number of samples in which they were present in that particular text type. For example, if a word occurred in all 10 samples of the most frequent 500 words in general writing, then it would be present in all the subsequent

samples for general writing as well. That would equal 120 samples in total. A word that occurred in 2 samples of the most frequent 5,500 and 3 in the most frequent 6,000 would have a total of only 5. The word type list for each text type was sorted in descending order according to this score, and a rank was assigned to each word using the rank function in Excel. Words with the same score also share the same rank. A pivot table was constructed in which the sum of these scores was calculated for each word family. The family list for each text type was then sorted in descending order according to how many samples each family occurred in and ranks assigned to each word family in the same manner.

4.2.3.1 Survival vocabulary. One good test of the hypothesis that conversation is the primary source of core vocabulary is to compare each text type to a survival vocabulary. Crabbe and Nation's (1991) survival vocabulary is a collection of key phrases and sentences, so a selection of 35 word types was extracted for use in this comparison. The American term *bathroom* was added to the list as a parallel to *toilet* for good measure.

Conversation is the only text type to contain all of these words, and most have the highest frequency in conversation as well. All the survival vocabulary are in the top 2,000 families in conversation (Table 4.10), and all but three are in the top 2,000 types (Table 4.9). The narrative family list also performs nicely. The type list for narrative writing suffers since past tense forms of verbs, e.g. *thanked*, *spoke*, and *called*, are more common in narratives than the simple forms of the verbs.

Table 4.9

Ranks in each text type for a selection of words from Crabbe and Nation's (1991) survival vocabulary. The lowest rank (= highest frequency) for each word type is in bold print and shaded. *NA* indicates the word type did not occur among the 6,000 words with the highest frequencies in that text type.

	Conversation	Narrative	General	Academic
hello	1	1,510	NA	NA
thank	1	508	3,416	NA
morning	1	1	539	3,623
tomorrow	1	514	1,720	NA
speak	1	514	1,578	1,594
call	1	1	539	1,072
repeat	2,843	3,359	3,459	3,427
excuse	499	1,013	4,647	NA
sorry	1	1	3,507	NA
matter	1	1	539	1
delicious	2,765	3,413	6,096	NA
cheaper	927	NA	3,283	NA
expensive	499	1,761	1,541	2,893
price	499	1,799	1	1,072
half	1	1	1	531
entrance	2,241	1,451	3,507	NA
exit	4,426	3,677	5,537	NA
closed	1,021	1	1,377	2,255
help	1	1	1	460
street	499	1	445	1,844
station	998	514	935	3,235
toilet	998	3,548	5,613	NA
bathroom	1,021	1,476	4,738	NA
hotel	1,021	514	1,014	5,960
airport	1,978	2,262	2,555	NA
train	499	514	1,485	3,891
bus	499	1,013	1,919	5,875
post	499	1,013	539	531
bank	499	997	1	1,364
straight	1	514	1,514	3,002
ahead	1,021	514	1,014	4,423
slow	1,464	997	1,578	2,115
down	1	1	1	1
stop	1	1	539	2,115
sick	499	987	2,764	5,139

Table 4.10

Ranks in each text type for a selection of word families from Crabbe and Nation's (1991) survival vocabulary. The lowest rank (= highest frequency) for each word family is in bold print. *NA* indicates the word family did not occur among the 6,000 words with the highest frequencies in that text type.

	Conversation	Narrative	General	Academic
hello	100	333	3,053	NA
thank	279	217	991	NA
morning	512	700	1,253	2,184
tomorrow	876	1,137	1,579	NA
speak	13	38	81	44
call	63	49	82	168
repeat	1,945	726	858	665
excuse	1,056	1,125	2,651	NA
sorry	876	947	2,216	NA
matter	449	261	438	405
delicious	1,976	2,262	3,310	NA
cheap	183	1,736	956	2,612
expensive	1,068	1,550	1,547	1,923
price	421	1,421	210	713
half	876	947	1,023	1,257
entrance	1,740	1,434	2,216	NA
exit	2,790	2,373	3,053	NA
closed	596	446	806	1,621
help	54	38	72	130
street	185	140	196	435
station	774	1,137	581	1,481
toilet	703	2,312	3,092	NA
bathroom	1,156	1,458	2,695	NA
hotel	671	924	806	3,132
airport	1,608	1,766	1,863	NA
train	78	208	166	428
bus	487	1,292	1,303	3,091
post	523	1,292	702	875
bank	541	592	204	853
straight	876	561	1,536	1,955
ahead	1,233	1,137	1,372	2,496
slow	403	221	715	935
down	850	939	968	1,060
stop	80	109	438	1,034
sick	1,040	1,266	1,911	2,514

4.2.3.2 Kernel vocabulary. Swadesh's kernel vocabulary (1971: 283) has been used by historical linguists in looking for historical links between languages. This list of 100 words represent central concepts in most every human society and so are less likely to be replaced by borrowing. In related languages, these kernel words tend to be cognates. One caveat is that, again, orthographic forms are used for comparison, so number 27 *bark*, for example, does not distinguish *bark* n. "the skin covering a tree" from *bark* v. "a sound made by a dog."

Comparing these 100 words with the sample of the four key text types (Table 4.11), only the word type *louse* is missing from all four. An additional 3 word types are missing from conversation and narrative writing. The information-focused text types of general and academic writing are missing an additional 11 and 16 word types respectively. The vast majority of these word types are most frequent in narrative writing. In 79 cases, they are ranked lowest (= highest frequency) in narrative writing versus 59 times in conversation. (The total is greater than 100 due to ties in rank.)

When grouped as families, narrative writing is even more impressive, having the highest frequency for 61 of the word families (Table 4.12). Conversation has the highest frequency for 35 of the word families and is often a close second for others. The other two written text types barely figure, having the highest frequency for only 3 word families each.

Table 4.11

Ranks in each text type for a selection of words from Swadesh's (1971: 283) kernel vocabulary. The lowest rank (= highest frequency) for each word type is in bold print and shaded. *NA* indicates the word type did not occur among the 6,000 words with the highest frequencies in that text type.

word type	Conversation	Narrative	General	Academic
<i>top-ranked word types</i>	55	79	32	26
1. I	1	1	1	1
2. You	1	1	1	1
3. we	1	1	1	1
4. this	1	1	1	1
5. that	1	1	1	1
6. who	1	1	1	1
7. what	1	1	1	1
8. not	1	1	1	1
9. all	1	1	1	1
10. many	1	1	1	1
11. one	1	1	1	1
12. two	1	1	1	1
13. big	1	1	1	2,592
14. long	1	1	1	1
15. small	1	1	1	1
16. woman	1	1	539	971
17. man	1	1	1	1
18. person	1	514	445	1
19. fish	499	940	903	2,809
20. bird	1,508	1,013	2,729	4,956
21. dog	469	514	1,812	3,173
22. louse	NA	NA	NA	NA
23. tree	966	514	1,998	1,594
24. seed	3,388	5,059	3,773	4,765
25. leaf	5,625	3,465	4,312	5,093
26. root	2,182	5,613	4,396	2,749
27. bark	NA	4,364	NA	NA
28. skin	1,508	514	1,812	3,475
29. flesh	4,771	1,013	4,911	NA
30. blood	1,021	510	1,030	895
31. bone	2,488	1,936	3,906	2,478
32. grease	4,830	NA	NA	NA
33. egg	1,021	2,761	2,994	5,236
34. horn	NA	3,892	NA	NA
35. tail	2,034	1,979	3,906	5,801
36. feather	4,886	4,071	NA	NA
37. hair	1	1	1,428	5,360
38. head	1	1	1	531
39. ear	1,793	1,013	4,220	4,045
40. eye	1,021	503	1,064	1,792
41. nose	1,021	514	3,559	NA

42. mouth	915	1	2,684	3,679
43. tooth	2,455	5,613	NA	NA
44. tongue	1,978	1,013	4,865	NA
45. claw	5,327	NA	NA	NA
46. foot	499	514	1,485	3,799
47. knee	1,829	1,761	4,135	NA
48. hand	1	1	1	1
49. belly	3,678	2,418	NA	NA
50. neck	1,343	514	2,622	NA
51. breasts	NA	1,510	NA	NA
52. heart	1,021	1	539	1,553
53. liver	3,706	NA	NA	2,311
54. drink	1	514	1,961	5,189
55. eat	1	514	1,961	4,206
56. bite	1,978	2,798	NA	NA
57. see	1	1	1	1
58. hear	1	1	1,064	3,105
59. know	1	1	1	1
60. sleep	499	488	1,961	5,293
61. die	1,021	514	1,961	3,315
62. kill	1,021	514	2,729	4,693
63. swim	1,793	2,701	NA	NA
64. fly	1,433	1,375	2,310	NA
65. walk	1	488	1,377	5,293
66. come	1	1	1	1
67. lie	1,414	514	2,431	2,115
68. sit	1	514	1,919	4,693
69. stand	499	1	1,064	1,594
70. give	1	1	1	1
71. say	1	1	1	1
72. sun	966	499	539	2,478
73. moon	1,621	1,392	3,609	3,623
74. star	2,710	1,476	1,812	6,508
75. water	1	1	1	492
76. rain	1,021	514	2,274	3,891
77. stone	1,508	514	1,428	1,792
78. sand	1,942	1,510	3,283	3,475
79. earth	1,316	514	1,030	1,425
80. cloud	5,796	1,510	4,396	6,152
81. smoke	3,956	3,756	6,389	6,508
82. fire	499	1	539	2,478
83. ash	3,819	2,262	6,096	6,298
84. burn	2,211	2,376	5,013	6,508
85. path	2,034	514	1,860	2,346
86. mountain	1,829	1,906	2,274	5,960
87. red	469	1	539	1,594
88. green	499	1	539	2,032
89. yellow	952	925	2,349	5,189
90. white	1	1	1	531
91. black	1	1	1	531

92. night	1	1	1	2,376
93. hot	499	514	1,064	3,378
94. cold	1	1	1,064	2,749
95. full	1	1	1	1
96. new	1	1	1	1
97. good	1	1	1	1
98. round	1	1	1	1,594
99. dry	927	514	1,485	2,809
100. name	1	1	1	531

Table 4.12

Ranks in each text type for a selection of word families from Swadesh's (1971: 283) kernel vocabulary. The lowest rank (= highest frequency) for each word family is in bold print. NA indicates the word family did not occur among the 6,000 words with the highest frequencies in that text type.

word type	Conversation	Narrative	General	Academic
<i>top-ranked word types</i>	35	61	3	3
1. I	16	10	23	106
2. You	9	9	135	331
3. we	23	23	54	85
4. this	259	280	321	363
5. that	259	280	321	363
6. who	96	62	137	157
7. what	259	311	395	447
8. not	259	280	215	363
9. all	876	947	1,023	1,060
10. many	876	947	1,023	1,060
11. one	233	311	395	447
12. two	198	349	438	638
13. big	133	332	243	1,845
14. long	217	301	349	405
15. small	352	443	395	447
16. woman	292	280	359	424
17. man	292	280	321	363
18. person	80	188	131	138
19. fish	477	526	638	1,353
20. bird	518	475	806	1,662
21. dog	233	404	715	1,908
22. louse	NA	NA	NA	NA
23. tree	398	349	660	785
24. seed	1,422	3,059	1,764	1,923
25. leaf	3,427	2,283	2,527	2,766
26. root	1,185	2,218	1,645	880
27. bark	3,088	1,884	3,453	3,401
28. skin	1,435	1,137	1,605	2,131
29. flesh	2,961	1,292	2,768	3,401
30. blood	1,233	1,125	1,381	1,319

31. bone	793	574	1,883	1,050
32. grease	2,894	3,059	3,453	3,401
33. egg	410	833	911	2,082
34. horn	3,883	2,474	3,453	3,401
35. tail	1,394	1,669	2,370	3,063
36. feather	2,414	1,615	3,453	3,401
37. hair	363	337	1,497	2,879
38. head	229	82	163	480
39. ear	541	455	1,911	2,347
40. eye	366	124	484	759
41. nose	1,233	1,125	2,234	3,401
42. mouth	1,156	386	1,911	2,205
43. tooth	580	1,112	2,120	2,322
44. tongue	1,608	1,292	2,753	3,401
45. claw	3,261	3,032	3,453	3,401
46. foot	325	311	369	1,319
47. knee	625	514	2,464	3,401
48. hand	167	51	183	359
49. belly	2,414	1,844	3,453	3,401
50. neck	1,344	1,137	1,883	3,401
51. breasts	2,188	609	2,987	3,132
52. heart	793	537	900	1,540
53. liver	2,432	3,498	3,453	1,732
54. drink	46	21	647	2,537
55. eat	34	83	770	1,815
56. bite	1,284	1,266	3,453	3,401
57. see	27	16	30	48
58. hear	153	137	221	386
59. know	8	8	18	27
60. sleep	118	126	1,023	2,850
61. die	132	152	303	894
62. kill	173	106	295	1,374
63. swim	493	597	2,193	3,401
64. fly	205	126	602	3,401
65. walk	62	67	179	2,398
66. come	34	27	43	88
67. lie	178	70	424	490
68. sit	58	66	301	1,944
69. stand	83	52	106	285
70. give	25	15	19	22
71. say	34	24	39	80
72. sun	430	194	1,253	1,799
73. moon	1,454	1,389	2,254	2,184
74. star	635	550	558	2,283
75. water	874	542	638	924
76. rain	322	896	1,755	2,283
77. stone	605	446	799	863
78. sand	850	839	2,107	2,131
79. earth	1,325	1,137	1,381	1,481
80. cloud	3,502	561	2,527	3,225

81. smoke	238	377	1,536	1,879
82. fire	661	219	684	1,799
83. ash	2,489	1,749	3,310	3,298
84. burn	233	120	1,701	3,225
85. path	1,641	857	1,605	1,344
86. mountain	814	648	977	3,091
87. red	679	947	1,253	1,556
88. green	1,040	947	1,222	1,676
89. yellow	1,185	1,211	1,783	2,802
90. white	876	947	1,023	1,194
91. black	720	682	738	901
92. night	325	404	671	1,762
93. hot	1,068	1,137	1,394	2,094
94. cold	661	526	1,394	1,879
95. full	469	404	395	392
96. new	876	597	321	579
97. good	223	499	1,007	1,044
98. round	786	648	835	1,374
99. dry	352	306	1,173	1,845
100.name	194	198	150	165

4.2.3.3 Lexical sets. One of the criticisms of frequency lists is the fact that basic lexical sets, such as the days of the week, months or basic color terms, do not occur as a complete set or in a similar frequency range (c.f. Gougenheim, Michéa, Rivenc, & Sauvageot, 1964: 102, 112).

In conversation, the names of all the days of the week fall within the most frequent 500 words (Table 4.13). Those associated with the weekend are among the over 400 conversation word types tied for rank 1. In narrative and general writing, the days are more scattered. Only Sunday is found frequently in academic writing, due to the names of several Sunday newspapers and also legal disputes over Sunday trading in the UK.

Table 4.13

Ranks in each text type for the names of the days of the week. The lowest rank (= highest frequency) for each word type is in bold print. *NA* indicates the word type did not occur among the 6,000 words with the highest frequencies in that text type.

	Conversation	Narrative	General	Academic
Sunday	1	1,013	935	4,693
Monday	492	1,936	1,761	NA
Tuesday	499	3,596	2,431	NA
Wednesday	499	2,856	1,860	NA
Thursday	499	3,159	2,499	NA
Friday	1	1,817	1,919	NA
Saturday	1	1,410	1,030	NA

The focus in general writing is much more on the year than the week, with the names of all the months well within the most frequent 1,000 word types (Table 4.14). The names of the days, however, would logically precede the months in a beginning curriculum. The months of the year are all in the second thousand word types for conversation, except for May and June that are homographs, one with a modal verb and the other with a popular given name.

Table 4.14

Ranks in each text type for the names of the months. The lowest rank (= highest frequency) for each word type is in bold print.

	Conversation	Narrative	General	Academic
January	1,508	4,898	539	1,553
February	1,433	5,329	539	2,115
March	1,364	2,347	1	1,072
April	1,248	3,307	1	1,464
May	1	1	1	1
June	499	3,121	460	1,594
July	1,248	4,453	504	1,594
August	1,288	3,892	877	2,115
September	1,248	3,756	539	1,446
October	1,621	4,071	539	1,072
November	1,760	4,022	539	1,594
December	1,508	5,371	539	1,594

The seasons have a much more even pattern (Table 4.15). *Summer* is the most popular, with the others following in the order of the seasons in the three non-specialized text types.

Table 4.15

Ranks in each text type for the names of the seasons. The lowest rank (= highest frequency) for each word type is in bold print.

	Conversation	Narrative	General	Academic
Summer	499	514	539	2,551
Fall	1,021	514	539	1,072
Winter	1,021	1,013	1,296	3,105
Spring	1,671	1,425	1,400	2,853

Color terms occur in a very logical order and, except for *purple*, within the top 2,000 words in both narrative writing and conversation (Table 4.16). The color terms have their lowest rank (highest frequency) in narrative writing. *Brown*, is also slightly out of order since it benefits from being a common proper noun.

Table 4.16

Ranks in each text type for color terms. The lowest rank (= highest frequency) for each word type is in bold print. *NA* indicates the word type did not occur among the 6,000 words with the highest frequencies in that text type.

	Conversation	Narrative	General	Academic
black	1	1	1	531
white	1	1	1	531
grey	1,288	514	2,461	5,051
red	469	1	539	1,594
green	499	1	539	2,032
blue	499	1	1,064	3,002
yellow	952	925	2,349	5,189
brown	927	514	1,064	2,032
pink	1,021	1,013	3,416	<i>NA</i>
orange	1,021	1,761	3,232	<i>NA</i>
purple	2,670	2,464	5,939	<i>NA</i>
gold	1,433	514	1,030	3,002
silver	2,088	514	1,919	4,910

4.2.3.4 Out of order words. Of the words Nation (1990: 21-24) gives as examples of words that should be in the top 2,000, but are often not, only *bath* is within this range in the most frequent types of conversation and narrative writing (Table 4.17). *Soap* (2,269) is close in conversation. *Stomach* is well placed in narrative writing and almost so in conversation. *Damage* is arguably somewhat academic. Perhaps, *break* or *hurt* could replace it. *Chalk* is only found among the frequent word types of conversation and narrative writing, and there just barely.

If these word types are grouped and ranked by family frequency, all but *chalk* are within the top 2,000 of conversation, and many are within the top 2,000 families of narrative writing (Table 4.18). One could argue that *chalk* is specialized term for the classroom, so a list designer wanting to make a list for classroom use, should explore a corpus of spoken English in the classroom, a project which Sorell, Chen and Lin (2013) have already begun piloting.

Of those that should be in the top 3,000 according to Nation (1990), all are within the desired range either when counting types or families for conversation: *tidy* (1,942 | 1,535), *stupid* (499 | 1,068), and *behavior* [spelled *behaviour*] (2,765 | 1,976). *Behavior* is still decidedly academic. The verb form is slightly more common in conversation. *Act* and *action* may be useful alternatives here.

Table 4.17

Ranks in each text type for the words (word types) mentioned in Nation (1990: 21-24). Some alternatives are listed in italics at the bottom of the table. The lowest rank (= highest frequency) for each word type is in bold print. *NA* indicates the word type did not occur among the 6,000 words with the highest frequencies in that text type.

	Conversation	Narrative	General	Academic
bath	499	1,510	2,555	5,001
soap	2,269	3,548	5,285	<i>NA</i>
stomach	2,034	1,013	4,701	3,841
damage	2,356	2,418	1,064	994
chalk	5,139	5,010	<i>NA</i>	<i>NA</i>
tidy	1,942	3,596	<i>NA</i>	<i>NA</i>
stupid	499	1,013	6,096	<i>NA</i>
behaviour	2,765	1,979	916	1
<i>behave</i>	2,488	<i>NA</i>	5,013	<i>NA</i>
<i>act</i>	1,464	940	1	1
<i>action</i>	1,862	1,013	1	1
<i>break</i>	499	514	948	1,792
<i>hurt</i>	927	514	3,333	<i>NA</i>

Table 4.18

Ranks in each text type for the words (word families) mentioned in Nation (1990: 21-24). Some alternatives are listed in italics at the bottom of the table. The lowest rank (= highest frequency) for each word type is in bold print. *NA* indicates the word type did not occur among the 6,000 words with the highest frequencies in that text type.

	Conversation	Narrative	General	Academic
bath	747	1,458	1,863	2,728
soap	1,754	2,312	2,932	<i>NA</i>
stomach	1,641	1,292	2,675	2,263
damage	1,688	1,577	461	355
chalk	3,165	3,032	<i>NA</i>	<i>NA</i>
tidy	1,535	2,124	<i>NA</i>	<i>NA</i>
stupid	1,068	1,292	3,310	<i>NA</i>
behaviour	1,976	1,669	1,323	713
<i>behave</i>	1,855	1,125	2,811	2,322
<i>act</i>	104	40	8	10
<i>break</i>	51	57	93	363
<i>hurt</i>	243	924	2,139	<i>NA</i>

Chicago was mentioned in Nation (1990) as being inappropriate for a core list. Being a proper noun, it was excluded from this analysis. It would have been ranked 4,117 since there are fewer American sources in this study. *Hong Kong* would have been rank 467, however, and *London* 538. The decision to include or exclude proper nouns can be done in a principled

way. Including local place names in English instruction is logical though, especially if the local language(s) use a non-Latin written script. It may also be wise to include proper nouns that are widely dispersed internationally as London and New York were in this study.

Nation (1990) also cites *his* which is the 74th word on Thorndike and Lorge's list and its feminine parallel *hers* which is the 4,151st. To be fair, though, *his* parallels both the pronominal usage *hers* and the more frequent determiner *her*. In conversation, *hers* is, nonetheless, at rank 1,621. In narrative writing, *hers* ranks an impressive 514. In general and academic writing, *hers* is a distant rank 6,389 and 6,508. One can see that the rank of 4,151 in Thorndike and Lorge's study resulted from a mixture of written narrative and informational text types.

From these comparisons, one could conclude that the conversation text type is central to the language, but does not have a monopoly on the core vocabulary. It is very closely related to narrative writing. Further research needs to be done in how to splice these two text types, as well as how to weave in general and academic vocabulary.

4.3 Which Unit Should Be Used For Counting Words?

The scope of the definition of a word as type or family can influence the resulting frequency list, as mentioned in section 4.2.2.1. On the one hand, the word type as a unit of counting provides a specific and targeted form for learning or investigating the behavior of words in different text types, e.g. Chung (2003). The word family, on the other hand, offers the possibility of quickly extending knowledge of word types, and thereby gaining valuable text coverage. This study compares word type and word family list sizes and their text coverage for each of the four key text types in order to weigh the benefits of each unit.

For this study, a word type frequency list was created for a 10-million-token sample of each of the key text types. The individual word type frequencies were summed for the most frequent 3,000 word types in each of the text types to determine how many of the running words of those texts would be covered by those 3,000 word types. That is coverage was only calculated for the text type from which the frequency list had been created.

The word types in each text type were then grouped as word families (see section 3.5.2). The frequency of each word family is equal to the combined frequencies of the word types in that family. The families were sorted in descending order of frequency, and the most frequent 3,000 word families were selected. Their combined frequencies were used to calculate how many of the running words of the texts in their text type these 3,000 word

families would cover. The total number of word types contained in these 3,000 word families was also calculated.

Lists of the most frequent 1,000-3,000 word types and word families were then compared to the General Service List (West, 1953) and the Academic Word List (Coxhead, 2000) in order to see how the lists from each text type compared to a well-known core vocabulary list and a supplemental academic vocabulary list.

4.3.1 General writing

Figure 4.59 demonstrates that choosing the word family as a unit can have a dramatic effect on how many types are included. The most frequent 1,000 word families in general writing cover over six times the number of types that the most frequent 1,000 type list covers. The number of types included in the most frequent 2-3,000 families increases by almost the same factor.

Figure 4.59

Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of general writing.

Figure 4.60

Tokens covered by the most frequent 3,000 types and families for a 10-million-token sample of general writing.

This dramatic increase in type coverage results in a significant, but much more modest gain in token coverage (Figure 4.60). The family lists for general writing cover close to 10% more of the tokens in the general writing texts. Fortunately, these gains can be made without necessarily increasing the learning burden six fold. The word families on the COCA/BNC lists are based on the principles described in Bauer and Nation (1993). Most word family members are included because they are frequent, regular, created using productive morphological processes, or are transparently predictable. That means that learning a few morphological patterns can dramatically increase the number of types that can be understood.

Not all family members, however, deserve the attention of beginning learners of English. Some of the included types are common only in certain text types or varieties of English. Some rare, archaic and nonstandard types were also included for the sake of completeness (Table 4.19).

Table 4.19

Sample family headwords and members with their respective frequencies in a sample of 10 million tokens of general writing.

BE (family)	TL 460,183
ain	85
am	2,625
are	49,129
aren	291
be	65,006
been	25,642
beens	6
being	9,646
beings	208
innit	6
is	104,521
isn	740
m	7,439
re	3,879
s	75,934
tis	29
twas	5
was	81,886
wasn	876
were	32,007
weren	211
wert	11
wuz	1

NAME (family)	TL 4,368
misnamed	1
misnaming	1
name	2,361
named	768
nameless	16
namely	177
names	862
naming	68
rename	11
renamed	50
renaming	15
unnamed	38

West's GSL contains approximately 2,000 word families, so a comparison with the most frequent 2,000 word families of a text type should be a good measure of how well the GSL fits each text type. The most frequent 3,000 word types in the sample of general writing contains 1,869 word families, and so is also a close equivalent. Whether one considers the top 2,000 families of general writing (Table 4.20) or the top 3,000 types (Table 4.21), there is approximately a 65% overlap with the GSL.

Table 4.20

Analysis of the most frequent word 3,000 families from a 10-million-token sample of general writing.

	top 1000 families				top 2000 families				top 3000 families			
	types		families		types		families		types		families	
GSL 1000	2,806	42.8%	714	71.4%	3,682	30.9%	931	46.6%	3,899	24.1%	983	32.8%
GSL 2000	478	7.3%	102	10.2%	1,500	12.6%	363	18.2%	2,374	14.7%	610	20.3%
Total not on GSL	3,273	49.9%	184	18.4%	6,748	56.6%	706	35.3%	9,904	61.2%	1,407	46.9%
AWL	677	10.3%	123	12.3%	1,842	15.4%	362	18.1%	2,438	15.1%	479	16.0%
proper nouns	94	1.4%	32	3.2%	201	1.7%	93	4.7%	411	2.5%	223	7.4%
marginal words	0	0.0%	0	0.0%	59	0.5%	2	0.1%	59	0.4%	2	0.1%
abbreviations	1	0.0%	1	0.1%	8	0.1%	7	0.4%	15	0.1%	14	0.5%
Total words	6,557		1,000		11,930		2000		16,177		3,000	

Table 4.21

Analysis of the most frequent 3,000 word types from a 10-million-token sample of general writing.

	top 1000 types				top 2000 types				top 3000 types			
	types		families		types		families		types		families	
GSL 1000	830	83.0%	597	79.3%	1,407	70.4%	899	63.9%	1,797	59.9%	914	48.9%
GSL 2000	49	4.9%	47	6.2%	194	9.7%	169	12.0%	396	13.2%	305	16.3%
Total not on GSL	121	12.1%	109	14.5%	399	20.0%	339	24.1%	807	26.9%	650	34.8%
AWL	68	6.8%	61	8.1%	226	11.3%	184	13.1%	430	14.3%	306	16.4%
proper nouns	31	3.1%	26	3.5%	82	4.1%	68	4.8%	143	4.8%	125	6.7%
marginal words	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%	0	0.0%
abbreviations	2	0.2%	2	0.3%	6	0.3%	6	0.4%	9	0.3%	8	0.4%
Total words	1,000		753		2,000		1,407		3,000		1,869	

4.3.2 Academic writing

The most frequent families in academic writing cover almost as many word types (Figure 4.61) as those of general writing with similar gains in token coverage (Figure 4.62).

Figure 4.61

Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of academic writing.

Figure 4.62

Tokens covered by the most frequent 3,000 word types and 3,000 word families for a 10-million-token sample of academic writing.

Not surprisingly, the top 3,000 word types and word families of academic writing include a large percentage of the words on Coxhead's (2000) Academic Word List, more than

any of the other text types. There is also a 55% or more overlap with the GSL for both the top 2,000 families (Table 4.22) or the top 3,000 word types (Table 4.23). The top 3,000 word types in academic writing are members of slightly fewer word families (1,739) than the top 3,000 word types of general writing.

Table 4.22

Analysis of the most frequent 3,000 word families from a 10-million-token sample of academic writing.

	top 1000 families				top 2000 families				top 3000 families			
	types		families		types		families		types		families	
GSL 1000	2,469	38%	614	61%	3,353	29%	842	42%	3,719	23%	939	31%
GSL 2000	425	7%	94	9%	1,094	9%	265	13%	1,781	11%	447	15%
Total not on GSL	3,585	55%	292	29%	7,281	62%	893	45%	10,537	66%	1,614	54%
AWL	1,223	19%	218	22%	2,469	21%	484	24%	2,806	17%	556	19%
proper nouns	48	1%	14	1%	125	1%	44	2%	245	2%	113	4%
marginal words	12	0%	1	0%	12	0%	1	0%	12	0%	1	0%
abbreviations	2	0%	2	0%	4	0%	4	0%	13	0%	12	0%
Total words	6,479		1,000		11,728		2,000		16,037		3,000	

Table 4.23

Analysis of the most frequent 3,000 word types from a 10-million-token sample of academic writing.

	top 1000 types				top 2000 types				top 3000 types			
	types		families		types		families		types		families	
GSL 1000	748	75%	517	71%	1,262	63%	712	57%	1,648	55%	808	46%
GSL 2000	62	6%	55	8%	171	9%	125	10%	321	11%	217	12%
Total not on GSL	190	19%	161	22%	567	28%	417	33%	1,031	34%	714	41%
AWL	146	15%	120	16%	429	21%	292	23%	694	23%	411	24%
proper nouns	13	1%	12	2%	33	2%	27	2%	69	2%	57	3%
marginal words	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
abbreviations	1	0%	1	0%	4	0%	4	0%	4	0%	4	0%
Total words	1,000		733		2,000		1,254		3,000		1,739	

4.3.3 Narrative writing

Though the number remains very large, a steady drop is seen in the number of types covered by the most frequent families from general to academic to narrative writing. The most frequent 3,000 word families in general writing covered 16,177 word types (Figure 4.59), in academic writing 16,037 (Figure 4.61) and in narrative writing 14,845 (Figure 4.63). The gain in token coverage when using the word family as a unit has also dropped to around 6.5% (Figure 4.64).

Figure 4.63

Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of narrative writing.

Figure 4.64

Tokens covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of narrative writing.

For the top 2,000 families (Table 4.24) and 3,000 types (4.25) of narrative writing, there is 70% or more overlap with the GSL. This is the closest any of the text types comes to

the GSL. For this sample of narrative writing, the top 3,000 word types were members of 2,001 families.

Table 4.24

Analysis of the most frequent 3,000 word families from a 10-million-token sample of narrative writing.

	top 1000 families				top 2000 families				top 3000 families			
	types		families		types		families		types		families	
GSL 1000	2,689	46%	690	69%	3,459	32%	874	44%	3,702	25%	941	31%
GSL 2000	898	15%	221	22%	2,215	21%	570	29%	2,928	20%	774	26%
Total not on GSL	2,304	39%	89	9%	5,122	47%	556	28%	8,215	55%	1,285	43%
AWL	74	1%	17	2%	548	5%	108	5%	1,108	7%	220	7%
proper nouns	42	1%	26	3%	194	2%	139	7%	367	2%	293	10%
marginal words	74	1%	2	0%	87	1%	4	0%	112	1%	7	0%
abbreviations	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Total words	5,891		1,000		10,796		2,000		14,845		3,000	

Table 4.25

Analysis of the most frequent 3,000 word types from a 10-million-token sample of narrative writing.

	top 1000 types				top 2000 types				top 3000 types			
	types		families		types		families		types		families	
GSL 1000	831	83%	571	78%	1,310	66%	779	56%	1,651	55%	871	44%
GSL 2000	121	12%	113	15%	399	20%	336	24%	679	23%	516	26%
Total not on GSL	48	5%	46	6%	291	15%	274	20%	670	22%	614	31%
AWL	6	1%	6	1%	47	2%	43	3%	112	4%	97	5%
proper nouns	25	3%	25	3%	120	6%	119	9%	237	8%	233	12%
marginal words	2	0%	2	0%	5	0%	3	0%	7	0%	5	0%
abbreviations	0	0%	0	0%	0	0%	0	0%	0	0%	0	0%
Total words	1,000		730		2,000		1,389		3,000		2,001	

4.3.4 Conversation

The most frequent families in conversation cover the smallest number of tokens of the four key text types (Figure 4.65). In fact, this is the only time where the most frequent 1,000 families of a text type does not include all of the types in its most frequent 1,000 types list. The most frequent 1,000 types not among the top 1,000 families are the 30 types: *northern*, *EC*, *nor*, *alone*, *James*, *advice*, *cash*, *attention*, *x*, *firms*, *extra*, *competition*, *queen*, *knowledge*, *evening*, *majority*, *wrong*, *William*, *county*, *j*, *mid*, *Richard*, *anti*, *dead*, *beyond*, *Smith*, *cold*, *Sunday*, *radio*, and *IBM*. The top 2,000 and 3,000 word type lists are completely included in the corresponding word family lists.

The increase in token coverage gained by using families drops to around 4% for conversation (Figure 4.66). The overlap of the GSL with the top 2,000 families (Table 4.26) or 3,000 types (Table 4.27) of conversation is similar to that of general writing at or around 65-68%.

The top 3,000 word types of conversation are members of 2,029 families, slightly more than those of narrative writing. Interestingly, the number of word families included in the top 3,000 word types follows the same pattern seen in the dendogram (Figure 4.58) of the relatedness of the four key text types: conversation (2,029), narrative writing (2,001), general writing (1,869), academic writing (1,739).

In their study of working-class, Australian speech, Schonell et al. (1956) found that 2,000 word families provided a coverage of 99% of the half-million running words they had collected. In a larger and demographically broader study of British and Irish conversation, Adolphs and Schmitt (2003) found 2,000 word families provided only 94.76% coverage of the 5 million words of the CANCODE corpus. The most frequent 3,000 families achieved a coverage of 95.91%. In this study, using an internationally diverse collection of around 7 million tokens of conversation, token coverage was nearly identical to that found by Adolphs and Schmitt. The most frequent 2,000 word families had a coverage of 95.9% and 3,000 word families covered 97.2% of the running words in the collection (Table 4.66).

Figure 4.65

Types covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of conversation.

Figure 4.66

Tokens covered by the most frequent 3,000 types and 3,000 families for a 10-million-token sample of conversation.

Table 4.26

Analysis of the most frequent 3,000 word families from a 10-million-token sample of conversation.

	top 1000 families				top 2000 families				top 3000 families			
	types		families		types		families		types		families	
GSL 1000	2,337	50.2%	660	66.0%	3,185	37.6%	877	43.9%	3,461	29.8%	947	31.6%
GSL 2000	725	15.6%	197	19.7%	1,710	20.2%	486	24.3%	2,397	20.7%	693	23.1%
Total not on GSL	1,593	34.2%	143	14.3%	3,584	42.3%	637	31.9%	5,749	49.5%	1,360	45.3%
AWL	77	1.7%	21	2.1%	592	7.0%	138	6.9%	1,076	9.3%	257	8.6%
proper nouns	72	1.5%	37	3.7%	227	2.7%	152	7.6%	429	3.7%	313	10.4%
marginal words	138	3.0%	9	0.9%	155	1.8%	12	0.6%	162	1.4%	13	0.4%
abbreviations	1	0.0%	1	0.1%	2	0.0%	2	0.1%	5	0.0%	5	0.2%
Total words	4,655		1,000		8,479		2,000		11,607		3,000	

Table 4.27

Analysis of the most frequent 3,000 word types from a 10-million-token sample of conversation.

	top 1000 types				top 2000 types				top 3000 types			
	types		families		types		families		types		families	
GSL 1000	779	77.9%	534	74.5%	1,249	62.5%	763	54.7%	1,572	52.4%	864	42.6%
GSL 2000	116	11.6%	102	14.2%	349	17.5%	289	20.7%	599	20.0%	455	22.4%
Total not on GSL	105	10.5%	81	11.3%	402	20.1%	342	24.5%	829	27.6%	710	35.0%
AWL	15	1.5%	13	1.8%	53	2.7%	47	3.4%	133	4.4%	115	5.7%
proper nouns	19	1.9%	19	2.6%	120	6.0%	112	8.0%	233	7.8%	217	10.7%
marginal words	24	2.4%	5	0.7%	45	2.3%	12	0.9%	61	2.0%	15	0.7%
abbreviations	0	0.0%	0	0.0%	2	0.1%	2	0.1%	2	0.1%	2	0.1%
Total words	1,000		717		2,000		1,394		3,000		2,029	

Many text strings in the conversation texts were labeled as abbreviations by Nation and Webb's (2011) BNC/COCA lists. However, marking abbreviations in conversational texts is ambiguous, and these needed to be investigated individually using a concordancer. Of those marked in the top 3,000 types for conversation, only *PHD* can reliably be labeled as an abbreviation. *MA* is sometimes *M.A.* but also represents many other meanings, e.g. *mother* and *Ah ma* in Chinese. *BA* is sometimes *B.A.* (Bachelor of Arts) but more often a marginal vocalization [ba] or a Tagalog borrowing. *AL* is more often a proper noun than an abbreviation. *HH* is a response form. *INT* (*isn't*) is a nonstandard tag question in British English. *NA* is tag question in Indian English. *TA* is more often used in the sense *thanks* than *T.A.* *TE* is often a Maori particle or sometimes a word fragment. *YO* is often an attention getter or fragment, but sometimes the lexical noun *yo yo*. Despite being an abbreviation for everything from *White House* to *Women's Health*, in transcripts of conversation *WH* is usually a transcribed word fragment.

4.4 Constructing objective core vocabulary lists

In this section, practical steps are outlined for constructing core vocabulary lists and a trial list is analyzed which partially illustrates this process.

4.4.1 Steps in constructing objective core vocabulary lists

4.4.1.1 Purpose. The first step in making a core is list is to decide the purpose the core vocabulary list (Nation, to appear). One common reason for disappointment in core

vocabulary lists is probably that they are often expected to do more than is realistic, or they are being used for purposes they were not designed for.

Chapter one described this research question as coming from a university-level, EFL teaching context in the “expanding circle” of English where most of the students are learners of English as a foreign language. The purpose of this trial list is to guide the explicit English vocabulary learning of first and second year students who are EFL learners. Most will have had a considerable amount of English instruction in secondary school, and so, a goal of at least 3,000 word families should be reasonable by the end of the second year. Since the students will be engaging regularly in conversation with English-speaking classmates and teachers, conversation should be a priority, even if it were not a major source of basic vocabulary. During these first two years, students will also be taking content courses in which English is the medium of instruction and / or the language of the course texts. For this reason, academic vocabulary cannot be delayed.

4.4.1.2 Data collection. The next step would ideally be to collect the amount of data for each of the key text types for each variety of English one needs to include based on the needs of the prospective users and the amount of variation in the lists that can be tolerated. Collections of 50 million running words have been shown to give very stable results, but that quantity is not yet feasible for conversation. One should also assemble collections of any supplemental text types that are needed, e.g. EFL primary school classroom language, if the core list is for primary school use.

4.4.1.3 Frequency lists. A composite word type frequency list can then be made for each text type being examined. After one has determined which word types are high-frequency, one can create a normalized frequency list for the word types in each text type. In doing so, one should take into account their range and dispersion (section 3.6).

4.4.1.4 Lemmas or word families. If one wishes to use lemmas or word families as the unit of counting, one should merge a lemma or word family list with the type frequencies list of each variety’s text type frequency list. (In this study, lists were merged using the R-script in Appendix D.2, but this and the steps below could be accomplished using a database, as well.)

4.4.1.5 Family or lemma frequency. Next, a family or lemma frequency list can be created for each text type of each variety using a pivot table (Table 4.28). (The 64-bit version of MS-Excel 2013 allows for large pivot tables that are limited only by the amount of RAM installed on your computer. In this study, pivot tables were constructed using a 32-bit version of MS-Excel 2007.)

Table 4.28

A portion of a pivot table showing the seven most frequent word families in a 10-million-token sample of general writing.

THE	656,031
the	656,031
BE	460,183
ain	85
am	2,625
are	49,129
aren	291
be	65,006
been	25,642
beens	6
being	9,646
beings	208
innit	6
is	104,521
isn	740
m	7,439
re	3,879
s	75,934
tis	29
twas	5
was	81,886
wasn	876
were	32,007
weren	211
wert	11
wuz	1
OF	332,456
of	332,456
TO	272,604
to	272,604

AND	270,995
and	270,995
A	269,722
a	231,763
an	37,959
IN	214,966
in	214,381
inner	494
innermost	9
inward	44
inwardly	10
inwardness	1
inwards	27
HAVE	122,386
d	4,397
had	37,252
hadn	240
hadst	1
has	31,442
hasn	216
hast	4
hath	19
have	42,178
haven	394
haves	8
having	3,486
ve	2,749

4.4.1.6 Core list length. Based on the needs of the prospective users, one must determine the desired length and number of bands for the core vocabulary list. If for example, one wishes to set vocabulary learning goals for tertiary-level students who will likely pursue post-graduate degrees at English-medium universities, one will need a list of at least 9,000 word families (Nation, 2006). For a list of that length, one should probably divide the next sorting process into at least three phases so as not to mix together words which are vastly different in rank order.

4.4.1.7 A joint table of the text types. One could then arrange the frequency lists for each of the text types in adjacent columns on a spreadsheet. In this trial, the rank lists from the text type comparison study (sections 3.4 and 4.2.1) were used instead of frequency lists. These only included the most frequent 6,000 word types in each text type. Instead of the total frequency, the total number of frequency ranges were totaled for each word type.). (See the “Text type comparison matrix analysis” spreadsheet in the attached supplemental files.)

4.4.1.8 Prioritizing text types. The words can then be sorted within a certain range giving priority to particular text types. One should almost certainly give conversation first priority since this text type is itself the core of the language. Narrative, general and academic writing seem to follow in logical order. Since the trial list is intended for academic and broad general use, a sorting was made to find all of the most frequent word families that are common to all four text types. A total of 1,786 word families were common to all four key text types. Thereafter, the list was extended to 3,000 word families by including the highest frequency word families in conversation that were not already among those common to all four key text types. This mean that the trial list of 3,000 word families included the most frequent word families from conversation plus all the word types that were among the most frequent 6,000 families of all four text types.

In retrospect, this span of the 6,000 most frequent words is really too great. The sorting should probably be done in phases, such as three bands of 3,000 word families each. This will avoid giving precedence to rarer words which are common to all four text types over very frequent words in conversation or the more general text types. Section 4.2.2.4 also suggests a gradual overlapping pattern might be useful. Further research, should be done on exactly how to prioritize the key text types in vocabulary instruction.

4.4.2 A trial core vocabulary list

The trial list contains all but 35 of the first thousand words of the GSL and is missing 252 of the words from the second thousand. Some are minor differences, such as *hurray*

instead of the GSL *hurrah*. The major differences between the lists are due to several factors: the time of their compilation, the definition of core vocabulary, the purpose of the list, the varieties of English included, the implicit value judgments present in subjective word choice, and the fact that the trial list is still very much a work-in-progress.

4.4.2.1 Time of compilation. The usage of some words has declined in the past 50 years or more, e.g. *beast, haste, inn, merry*. These words were already antiquated when the GSL was compiled. *Thou*, on the other hand, is unexpectedly included on the trial list. Sometimes it is used humorously. Sometimes it is a word fragment in conversation, and sometimes it is a clipped form of *thousand*. This is a good example of where dividing homographs would be useful.

Technology. Many of the differences relate to the half-century difference in the times the lists were created. *Brass* (GSL 2,000) was once a commonly used material, but today the metal has been replaced by lighter and cheaper materials, such as plastic. Other words related to dated technologies or economic systems are: (GSL 1,000) *colony, flesh, sail, sword, vessel*; (GSL 2,000) *canal, carriage, oar, shield, shilling, telegraph, weave*. The trial list now has *battery, cassette* (which is already outdated again), *dishwasher, email, fax, headphone, helicopter, internet, jet, laser, movie, microphone, microwave, refrigerator, remote* (i.e. away from the TV), *rocket, stereo, television, teletext, vacuum, and video*. Some of the differences are subtle. The GSL *rug* is replaced by *carpet*, which is roughly synonymous, except that *rug* usually implies a single movable piece, whereas *carpet* is often wall-to-wall. At the time the GSL was compiled most floor coverings were handmade, whereas most modern homes contain factory-made carpet that is cut to fit the room.

Demography. At the time the GSL was being created, the earth's population was largely rural and many people were engaged in agriculture. Since 2008, and for the first time in history, the majority of human beings live in urban settings. Therefore many of the words related to life in the countryside are no longer as common: *agriculture, axe, cattle, cultivate, ditch, grease, grind, harvest, hay, deer, plough, rope, sow, strap, tame, tide, wheat*. *Wheat* would have been included in the trial list, if the list were extended a couple hundred more word families, but in modern usage *wheat* is more often an ingredient than something people are planting.

The newer list does not include *agriculture*, but a typically urban term, *architecture*, does make the list. The words *countryside* and *rural* are now included since that is a place one talks about rather than where one lives. The new list contains a number of items related to an urban setting: *airport, alley, apartment, balcony, cement, concrete, greenhouse, lawn, marina,*

mall, motorway, museum, supermarket, traffic, warehouse. Fox is on the newer list, but as the news channel rather than the animal.

4.4.2.2 Definition of core vocabulary. Some of the differences between the GSL and this trial list stem from a difference in the conception of what a core vocabulary list should be.

Superordinates. The GSL was constructed on the principle that broadly applicable superordinates should be selected, and that unnecessary synonyms should be avoided. The list of the first thousand words of the GSL includes *car*. The newer list also includes *automobile*, and many car parts: *aerial, bonnet, boot* and the American English terms *hood* and *trunk*. The GSL has both *boat* and *ship*, while the new list also adds the hyponym *ferry*. The GSL uses the general term *berry*, while the trial list only includes its hyponym *strawberry*. *Tobacco* on the GSL has been replaced in the trial list with the more common and more specific *cigarette*. Both the full forms *aeroplane* and *airplane* are on the GSL, while the trial list has only the clipped form *plane*. *Bath* is found on both the GSL and the new trial list, but only the GSL has the verb *bathe*. *Bathe* was likely included as the action clearly related to the noun *bath*. The verb *bathe* sounds formal, and is, therefore, rare in conversation. One would normally say *take a bath* or *take a shower*, but these expressions are arguably less transparent to a beginning learner of English. In creating a core list one has to make a choice between what might be most useful and what is most often used. Though West and company may not have liked the comparison, the GSL is really not unlike Ogden's BASIC in that they are both designed to be a minimalist code in which the learner can communicate adequately and quickly understand specially composed texts. Unlike Ogden, they saw their list as transitional. The objective methods outlined in this thesis seek to provide learners with the words most often used and expected by other speakers of English. It seeks to discover the core vocabulary that has already been (and is being) negotiated by speakers of the language.

Affectively neutral. Another commonly cited criterion for selecting core words is that they should be affectively neutral. Words that are marked in some way are often unavoidable, however. The GSL contains *insect*, which is formal, while the trial list has *bug*, which is informal. This tendency to err on the side of formality can be seen in many of the GSL words that are missing from the trial list: (GSL 1,000) *accountable, adopt, arise, beneath, declare, distinguish, experiment, latter, manufacture, moreover, necessity, numerous, numerical, poverty, possess, provision, quantity, remark, republic, scarce, substance, thus, yield*; (GSL 2,000) *compose, decay, descend, decrease, enclose, forbid, inward, limb, mineral, moderate, neglect, omit, onwards, ornament, postpone, reproduce, rejoice, scorn, shelter, solemn, upwards*.

4.2.2.3 Purpose of the lists. As mentioned in section 4.4.1.1, the proposed use of a core vocabulary list can and should influence the construction of a core word list.

Reading as a priority. The GSL was intended more for reading than for speaking. Of the four key text types, the GSL had the greatest coverage of narrative writing (See section 4.3.3). Many of the GSL words not on the trial list would be useful in composing or reading graded narratives: (GSL 1,000) *enemy, noble, shadow, laughter*; (GSL 2,000) *cape, cave, cliff, ceremony, conquer, explode, explore, feast, fierce, meanwhile, mystery, pearl, revenge, roar, sacrifice, slave, tribe, veil*.

Academics. With its implicitly academic focus, the GSL includes useful metalanguage: *plural, verb, vowel* etc. Similar to general writing, the GSL also includes an array of weights and measures.

The trial list deliberately included the academic text type, but failed to include the above words. The trial list does, however, cover 295 of the 570 words from the Academic Word List (Coxhead, 2000). The trial list also includes many other words related to academia that are not on the AWL or GSL: *atmosphere, biology, curriculum, classroom, classmate, dean, exam, faculty, geography, graduate, gymnasium, homework, laboratory, linguistic, mathematics, medieval, physics, professor, scholarship, semester, seminar, syllabus, tertiary*.

4.2.2.4 Varieties included. The GSL includes dollar and pound (and the now outdated *shilling* mentioned above). The trial list adds the informal term *quid* and the more international terms *euro* and *rupee*. The word types *ano, kasi, lah, haan, and yaar* (discussed in section 5.3.3) would arguably make good additions to a core list of ASEAN English varieties. Appropriately, the trial list also includes the term *bilingual*.

4.2.2.5 Tacit values. The GSL includes very many words that relate to positive or negative character traits: *conscience, courage, confess, coward, envy, greed, loyal, mercy, patriotic, prejudice, prompt, punctual, sincere, vain, virtue*. Many of these words would be useful for narrative writing, but they also seem to reflect the tacit value judgments of the list's creators. They are absent from the trial list, however. Instead there is *abuse, rape, suicide* and a worrying number of terms related to, and including, *alcohol: beer, booze, brandy, cocktail, gin, whiskey, vodka*. Objective methodologies apparently tend toward Realism rather than Romantic idealism.

4.2.2.6 A work in progress. This trial list is still flawed in many respects. It was compiled using too little data. The fact that the conversation data was bootstrapped to samples of 10 million words meant local repetition of types was replicated in all ten samples defeating any measure of dispersion. The representation of English varieties was also highly imbalanced. Additional work is needed on the lists of marginal words and proper nouns. More

experimentation is also required in how to balance and prioritize the text types so that the resulting core lists fit the needs of their users. Language teachers should obviously not hand over control of their syllabus to a programming script. Still the results of this study are promising. Core vocabulary lists based on transparent and replicable methodologies are feasible.

CHAPTER 5

DISCUSSION AND CONCLUSION

In the past, creators of core vocabulary lists have had to rely on intuitive methodologies or a mixture of objective and intuitive methods that were difficult to document and impossible to replicate. When these methodologies were corpus-based, the creators were restricted in how much and what type of data they could acquire. When they had amassed a certain quantity of data, there was no sure way of knowing how reliable the derived frequency list might be or how it would relate to lists compiled from other types of texts. All core vocabulary lists were, therefore, based on intuition either at the level of word selection or at the level of text selection, or often some combination of both.

The studies in this thesis have outlined practical steps that can produce intuitively *prima facie* core vocabulary lists that are derived using a clear and objectively replicable methodology. This final chapter will discuss the findings and implications of the three studies in this thesis in sections 5.1-4 and then conclude with a summary of the findings and some possible applications in section 5.5.

5.1 A Taxonomy of English Language Texts

Even today, when multi-billion-token corpora can be compiled, the types of texts included are largely determined by what texts are available rather than by any linguistically validated taxonomy of the texts of the language. To fill this gap, this thesis proposed a simplified taxonomy based on Biber's (1988, 1995) multidimensional analysis of English texts. The studies in this thesis, in a sense, have validated Biber's studies by adding a vocabulary dimension to the description of each of the key text types. These studies are limited, however, since a multidimensional analysis was not duplicated for the actual sources used in this thesis.

The source texts selected for the studies are from equivalent genre to those text that were characteristic of each of the key text types that Biber (1995) identified. Since face-to-face conversations made up all of text type 1 *intimate interpersonal interaction* and the vast majority of text type 2 *informational interaction* in Biber's study, texts of face-to-face interactions were used to represent the combined *interactional text type* in the studies in this thesis. For that reason, the interactional text type has usually been referred to in this thesis simply as *conversation*. The majority of texts in Biber's text type 6 *general reported exposition* were various genre of non-academic writing and reportage. A collection of

biographies, editorials, essays, humor, instructional writing, persuasive writing, popular lore, popular writing and reportage were extracted from a number of different corpora to represent what this thesis referred to as the *general writing* text type since each of these genre were typical of Biber's *general reported exposition* text type. Biber's text type 5 *imaginative narrative* consisted primarily of the many genre of fiction. Corpus sections labeled "novels and short stories," "fiction and drama," "general fiction," "mystery and detective fiction," "science fiction," "western and adventure fiction," and "romantic fiction" from various corpora were collected to represent the *narrative* text type in this thesis. A collection of technical and academic prose from both the social and natural sciences were combined to represent the *academic* text type in this thesis. Biber's study had divided academic prose into the distinct text types since the texts in text type 3 *scientific exposition* contained an even greater number of passive constructions than those in text type 4 *learned exposition*.

Despite the lack of a multidimensional analysis of the specific texts used in this study, the vocabulary profiles discovered for each text type fit very well with the linguistic characteristics and purposes of each text type as described by Biber. The first and second person pronouns are the most frequent words in conversation. Conversation also features numerous words from everyday life, especially food-related terms. Narrative writing contains a wealth of words for describing settings and the actions and emotions of characters. General writing has the widest vocabulary of all the text types, but still reflects its roots in the bureaucratic and accounting documents for which writing was first developed. Academic writing ranges from the shared academic vocabulary described by Coxhead (2000) to highly specialized terms, like *eigenvalue* and *oligonucleotide*. Further research could include a replication of Biber's 1988 and 1995 studies with an even larger group of texts than he originally used.

Biber's study placed some spoken texts in *general reported exposition* and *imaginative narrative*, while some personal correspondence was included in *informational interaction*. In the studies in this thesis, the choice was made to avoid mixing spoken and written texts since, without a multidimensional analysis, that would have greatly increased the risk of placing texts in the wrong text type categories.

5.2 How much data is necessary to give consistent results for high-frequency vocabulary?

5.2.1 A lot!

The answer to this research question, of course, depends on how stable a list one wants and for what quantity of high-frequency words. The results of the Dice distance study (sections 3.2 and 4.1) revealed that there is good news and there is bad news. The good news is that surprisingly stable results can be achieved for a much larger group of words than was expected. The bad news is that, to achieve the best results, much more data is needed than was previously expected.

These studies began with the hope of creating core vocabulary lists of 2,000-2,500 words. The Dice distance study revealed that the results for the most frequent 3,000 words can, in some cases, be superior to shorter lists. Reasonably stable lists can also be made for the most frequent 6,000 or even 9,000 words. This correlates well with the recommendations of Schmitt and Schmitt (2012) to call the 4th 1000 to 9th 1000 "mid-frequency vocabulary" in order to focus attention on them. They suggest that language programs and curricula set vocabulary learning goals that include the mid-level vocabulary between the highest-frequency 3,000 words and the low-frequency words beyond rank 9,000 since this range is currently neglected in EFL pedagogy.

With text type collections of 50 million tokens, one should be able to expect a variation of less than 5% for the most frequent 3,000 words of any of the key text types (Table 5.1). Even the most frequent 9,000 words should only vary between 4-7%. That means that 9,000-word vocabulary lists for a single text type should only vary by a few hundred words. Word lists with this level of reliability would command considerable respect and would be a trustworthy resource for language learners, teachers and researchers.

Table 5.1

Approximate amount of variation that should be expected for various-sized frequency lists for each of the key text types using a corpus of 50 million tokens for each text type.

most frequent words	Conversation	Narrative	General	Academic
1,000	≥2%	2%	2%	3%
3,000	4%	3%	<3%	4.5%
6,000	<6%	4.5%	<4%	5%
9,000	7%	5%	4%	6%

Collecting 50 million tokens for a single text type is feasible in many cases. For English-language conversation, that quantity is not currently feasible. More work needs to be done in comparing simulated dialogue from TV and films with authentic conversation. However, if one wishes to create parallel regional collections for each text type rather than mixing different varieties of English, gathering that quantity of data for any of the text types

may be currently unrealistic. Pooling the available data from different varieties and using a range filter and a usage coefficient (Section 3.6) are the best options at this time.

Brysbaert and New (2009) considered corpora of 16-30 million to be cost-effective for psycholinguistic studies. The Dice distance study found collections of 20 million words should still vary by around 5% for the first 3,000 words, and less than 10% even for the first 9,000 (Table 5.2).

Table 5.2

Approximate amount of variation that should be expected for various-sized frequency lists for each of the key text types using a corpus of 20 million tokens for each text type.

most frequent words	Conversation	Narrative	General	Academic
1,000	2.0-2.5%	>2.0-<3.0%	3.0%	>3.5->4.0%
3,000	4.5-5.0%	≤4.5%	3.0-3.5%	≥5.0%
6,000	7.0%	6.0-<6.5%	<5.0%	≥6.5%
9,000	8.5-9.0%	7.0-7.5%	5.5-6.0%	8.0-8.5%

Even collections of 15 million, just below the cost-effective range recommended by Brysbaert and New (2009), are only somewhat more variable (Table 5.3).

Table 5.3

Approximate amount of variation that should be expected for various-sized frequency lists for each of the key text types using a corpus of 15 million tokens for each text type.

most frequent words	Conversation	Narrative	General	Academic
1,000	≤2.5%	≥2.5-3.0%	≤3.5%	4.5%
3,000	5.0-<6.0%	5.0-5.5%	4%	6.0%
6,000	7.5-8.0%	≥7.0%	5.5-6.5%	≥7.5%
9,000	9.5-10.5%	8.0-8.5%	6.5-7.0%	9.0-9.5%

Dividing the vocabulary learning goals in a curriculum by at least this number of bands seems advisable. Mixing words with too wide a span of frequencies would reintroduce the same problems with the ordering of words on core lists that Nation (1990: 20-21) identified and that this study has been trying to solve.

A nucleus of under 2,000 word families could launch learners into the language. A vocabulary of 3,000 words would enable learners to deal with most graded readers. Beyond that, learners need guidance in prioritizing mid-level vocabulary learning. This also highlights the need for well-written mid-level graded readers. Quality core vocabulary lists of around 9,000 word families are needed to guide their composition. Many common ESL textbooks use a vocabulary of 6-9,000 words (Hsu, 2009), but the mid-level vocabulary was almost certainly

based on each author's intuition and preferences. A vocabulary of 6-9,000 word families would allow for participation in normal conversation and give learners access to ungraded written texts (Nation & Newton, 2009: 133; Nation, 2006).

In all, Dice distance proved a useful tool for exploring variations in the vocabulary of texts. Using sampling with replacement should probably only be used when at least 50 million tokens of data are available. Even then, the results should be adjusted up slightly.

Brysbaert and New's estimate of 16-30 million tokens as a cost effective corpus size is a reasonable amount. For studies that require more precision, especially the creation of longer core vocabulary lists, a sum closer to 50 million tokens is very desirable. For the three written text types, compiling collections of 50 million tokens is quite feasible. For spoken conversation, that is not yet possible.

In future studies, digits could probably be deleted. Though they provided some insights, they greatly increased processing times. Part-of-speech tagging and word sense tagging would add precision to the resulting lists, and should probably precede the publication of a core list (McEnery, Xiao, & Tono, 2006: 37-38).

5.2.2 The Size and definition of the core vocabulary

The results of the first study (the Dice distance between samples of the same text type) led to the realization that the core vocabulary can be considered much larger than it often is. As was just mentioned, defining a total core vocabulary of 6,000-9,000 word families is not unreasonable. In the past, definitions of the core vocabulary have tended to limit the core to around 2,000 words, which was also the perceived practical limit of core vocabulary lists (Fries & Traver, 1960: 26). Hindmarsh's Cambridge English Lexicon (1980) and the more recent 5,000 list derived from COCA (Davies, 2012) are unusual in this regard. The relatively small size of core vocabulary lists left a large gap between the perceived core vocabulary and the thresholds for practical use of the language. Nation and Newton (2009: 133) estimated that 6,000 word families are needed as a minimum for comprehending ungraded speech while around 9,000 are needed for ungraded reading (Nation, 2006). Learners have essentially been left to fend for themselves in this region between beginning vocabulary learning and full participation in the language. Not surprisingly many foreign language learners have gone missing rather than achieving a functional command of the language.

The band of frequent vocabulary that fits the semantic, syntactic and sociolinguistic definitions described in section 2.4 would perhaps be better termed "the nucleus" – a

vocabulary at the center of the core. The core vocabulary of English is actually much larger than this, however.

Zipf diagrams of each of the key text types (Figures 4.6, 4.7, 4.10, 4.13) show that the thresholds identified by Nation and Newton for ungraded speech and writing correspond roughly to observed changes in the mathematical distributions of words in collections of texts. The Zipf diagrams show two distributions, one with an incline of approximately -1 and a second with a much steeper incline. Ferrer-i-Cancho and Solé (2001b: 4) argued that this is due to the network structure of the lexicon. The first frequency regime is the core vocabulary, while the second is the peripheral vocabulary. The few thousand words that make up the core are both highly connected and highly interconnected. The more specialized words that make up the second distribution are connected in a small world configuration to the words in the core. The words in the core are also closely connected to each other. Each core word is connected on average to 24% of the other core words (Ferrer-i-Cancho and Solé, 2001b: 5). Such an arrangement allows for mutual comprehension even when specialist terminology is not shared. Core words can be used to circumlocute around or elaborate on specialist terms. If this small world configuration reflects how words are arranged in the mental lexicon, it also explains how speakers are able to rapidly access a rich vocabulary that includes many thousand rarer words. With a highly interconnected core vocabulary, no two words in the network of a speaker's mental lexicon are more than a few jumps from each other. This is truly exciting since it means this frequency distribution that has fascinated mathematicians and physicists is intimately related to the work of many language teachers and linguists through the centuries. Core vocabulary lists are not just useful pedagogic tools, they are key to understanding the structure of a language.

5.3 How different are the text types?

5.3.1 A core or cores?

The Dice coefficient was also used to measure the distance between the four key text types in the proposed model of the English language. What was discovered is that each text type is approximately equidistant from its neighbor with conversation at the center, followed by narrative, general and academic writing in that order (section 4.2, especially 4.2.2.4). The nucleus of the lexicon seems to be largely concentrated in the two central text types of conversation and narrative writing (section 4.2.3). In a sense, the four key text types represent successive socio-linguistic domains spreading from the home to the wider society. One could

also picture this arrangement of the text types as tracing their historical evolution. Face-to-face conversation is the first and most basic use of language. Narratives are an essential part of conversation (Hutchby & Wooffitt, 1998; Jaworski & Coupland, 1999; Labov & Waletzky, 1967), and the oral narrative naturally became an early form of literature (Fludernik, 1996).

Narratives were eventually recorded in writing, but writing systems were not developed to record narratives. Instead, “the early civilizations of Mesopotamia, Egypt, China, Mesoamerica, and (probably) India all developed a system of writing” for use in administration, trade and religion (Gnanadesikan, 2008: 2). The earliest written texts are, thus, informational in purpose. The narrative is a naturally oral text type, but the evolution of the written narrative is also tied to the information-focused recording of historical annals (Scholes, Phelan, & Kellogg, 2006: 211), and so, it is not surprising that narrative writing stands between conversation and the other written text types. Modern English is like a binary star system. Of the four key text types, narrative writing feels the pull of the dual oral-written gravitational fields most strongly.

Figure 5.1

Visualization of the relationship of the four key text types. (The area of each circle is not related to the vocabulary size of that text type, but simply its relatedness to the other text types.)

In light of its origins, general writing’s emphasis on the annual calendar, scheduled time and measurement is understandable (section 4.2.3.3). The original themes of administration, business and religion have now been expanded to include more democratic

interests, such as hobbies, sports and less technical articles on science, technology and the natural world.

Since it is a specialized use of language, academic writing is the one text type that could arguably be separated from the other three. However, the distance between academic writing and general writing is roughly the same as the distance between general writing and narrative writing. It is also the same as the distance between narrative writing and conversation. Creating a core vocabulary list also probably implies an academic setting, in which case, this text type is essential. The structure of the curriculum and the language environment would have to be considered in deciding how soon to focus on academic vocabulary.

Writing by definition creates a certain amount of diglossia between the written and the spoken language. In many societies, the written and spoken languages are noticeably distinct. In German-speaking Switzerland, the written language is referred to as *Schriftdeutsch* (written German), as opposed to the variety of German one uses for everyday conversation. Marked cases of diglossia between writing and speaking exist for written Tamil in southern India, for Putonghua in many parts of China, and for standard Arabic in most of the Middle East, just to name a few. Looking only on the vocabulary variation between the text types seen in this study and ignoring the sometimes substantial phonological variation between the Englishes of the world, one could still say that English still has one core vocabulary.

At one time, access to literacy required wealth and power. Today, millions not only have access to written texts, but have published their own voice in the form of blogs and other online genre. Greater and greater numbers of people are able to hear and be heard in all four text types. In the past, writers sometimes resorted to writing over the text of an existing book since writing materials were so scarce, as was the case for one of Archimedes' most important works the only known copy of which was unfortunately written over by a Byzantine scribe. The publishing of academic texts is no longer limited by how many animal hides are available to be made into parchment. Many quality journals can now be accessed for free online, such as *Reading in a Foreign Language* or the *Public Library of Science* journals. The fact that so many speakers regularly participate in the consumption and production of all of these key text types binds the text types together. Because literacy is so widespread, the second study in this thesis found fairly academic vocabulary, e.g. *analyse*, *sociology*, *symptom* and *tertiary*, among the most frequent words in conversation (Table 4.4).

One difficulty arises when one tries to create a single vocabulary list for all four text types. Vocabulary acquisition, like the lexicon of a language, is multidimensional. In the course of a single day, one may learn new words in the context of reading a novel, a

newspaper, an academic journal or while talking with a friend. How does one convert this multidimensional vocabulary growth into a single linear vector? Conversation and its nearest neighbor narrative writing should probably take priority followed by general and then academic writing. Exactly how to smoothly splice these four strands into a single list remains a problem. There may not be a single universal solution.

There may be a place for creating four free-standing lists. Each learner has a unique experience with the language which will be a unique mix of these four text types. Some learners may never venture outside casual conversation in English while others may be dealing with academic writing long before they are exposed to much spoken English. Even if there is one core vocabulary, the order in which learners acquire these several thousand words will no doubt vary considerably.

5.3.2 Narratives and proper nouns

In the first study, narrative and academic writing show a similar degree of variability beyond the top-ranked 1,000 words. This is, at first, surprising (or perhaps not for many EFL students in literature courses). A large portion of the variation in narrative writing is caused by an abundance of proper nouns. Of the top 3,000 types in narrative writing, 8% are proper nouns – 12% if they are grouped as families (Tables 4.24-25). This is higher than for any other text type.

Literature has long been a mainstay for EFL education, and proper nouns are generally not regarded as a significant hindrance to comprehension. The technical terminology in academic writing is certainly more challenging since it requires extensive content knowledge. However, proper nouns do add to reading difficulty, especially if they are phonologically unfamiliar (Brown, 2010: 357-358; Kobeleva, 2012). In order to comprehend a narrative, the listener or reader must associate successive actions or characteristics with particular individuals and places. If the names of people and places in the narrative are unfamiliar, organizing this information and, therefore, following the narrative becomes much more difficult.

5.3.3 Word types in spoken English, including EIL

Many of the words categorized in this study as marginal words are markers or inserts that are an essential feature of spoken language. *Okay* (or *OK*) is typically thought to mean one agrees with something or that it is acceptable. In conversation, where it is most frequent,

okay is probably more commonly used for signaling the closing or pre-closing of a topic or conversation, e.g. “OK. See you tomorrow. Bye.” One could argue that they are essential features of language and that they have been neglected because of an inordinate focus on written language. More recent ESL grammars (Biber, Conrad, & Leech, 2002; Carter & McCarthy, 2006) devote key chapters to spoken conversation and the functions of these often amorphous words, since they facilitate the flow of interaction and understanding in face-to-face conversation.

Specifically oral word forms become problematic for corpus analysis though, because they often lack standardized spellings or are indistinguishable in written transcription when they are separated from facial expression, intonation, voice quality and context (Biber, Conrad, & Leech, 2002: 430; Carter & McCarthy, 2006: 164). Since such words are important to a learner’s success in the context of a new language or variety, marginalizing these words in ESL/EFL instruction would not be wise. Even taboo words, which are also categorized as marginal words in this study, need to be addressed, so that learners are conscious of their social ramifications. These two issues have been side-stepped by this study. What have been termed “marginal words” in this thesis deserve further investigation. Future research could suggest a more extensive and detailed classification for these word families.

Linguistics texts usually cite nouns as examples of borrowings from other languages, e.g. Crystal (2003: chapter 20). This study has demonstrated that response forms, vocatives, tag questions and other spoken inserts are very frequently borrowed in spoken English as an International Language. Even though the samples of spoken English from India, the Philippines and Singapore made up only a very small portion of the spoken data in this study, numerous discourse markers from Hindi, Tagalog and more than one variety of Chinese were among the most frequent word types. In these contexts, it would be important to understand the Hindi borrowings *haan* “yes,” *na* “isn’t it” and *yaar* “mate,” the Tagalog *oo* “yes,” *ano* “what?” and *kasi* “but,” or *hor* “good” from Southern Min and *lah* a verbal particle indicating completed action from Chinese that is used for emphasis Singapore English. The same is really true for similar terms, like *ta* or *cheers*, for speakers moving from one inner-circle variety to another.

5.3.4 Beyond the four key text types

The four text types of conversation and narrative, general and academic writing are a useful set of text categories, but they are really just the beginning of a true taxonomy of

English language texts. Beneath and between these categories are many other types of texts that should be investigated.

“Conversation is the most basic register of human language Other spoken registers are . . . much less basic” (Biber & Conrad, 2009: 86). In this thesis, the interactive text type was considered synonymous with conversation. However, there are interactive written texts, such as some personal letters and written-to-be-spoken simulated interactions. Further research should examine the simulated dialogue of TV and film (Brysbaert & New, 2009) or the transcripts of interviews and talk shows (Davies, 2010) to discover how similar they are to authentic conversation.

An investigation should also be made of the vocabulary differences between casual conversation and more context-driven speech. A comparison of casual conversation, public dialogues, as well as scripted and unscripted monologues would be useful in validating the use of these texts to represent spoken English. Context-governed speech tends to make up most of what is included as spoken language in many corpora. This probably has to do with availability since meetings, interviews, speeches are more likely to be recorded and the participants are more willing to have them recorded. The vocabulary of these more formal speech situations is likely to be quite different from that of casual conversation. Google Glass™ and other recording devices raise the somewhat frightening possibility that recorded casual conversation may soon be widely available whether or not we are willing to be recorded.

Some personal correspondence was included in text type 2 *informational interaction* of Biber’s study, but written texts were excluded from the collection that represented the joint *interactive* text type in this thesis. The *interactive* text type is a combination of Biber’s first two text types. A subsequent study of the vocabulary of informal personal correspondence, would be of interest, especially writing that is typically less edited, e.g. emails and text messages. This electronically-mediated, often interactive writing should be similar in many ways to casual conversation. For example, in a comparison of a corpus of Twitter messages and the ukWaC (Ferraresi, Zanchetta, Baroni, & Bernardini, 2008), a billion-token corpus from UK websites, Herdağdelen (2013: 15) found 80-90% of the total occurrences of *me*, *my*, *I* and *you* in the Twitter corpus rather than in the more general web-based texts. The dozen most frequent word forms in a corpus of SMS messages sent from mobile phones were *I*, *to*, *you*, *the*, *u*, *haha*, *it*, *a*, *me*, *is*, *for* and *my* (Chen & Kan, 2013: 320). Such texts would also introduce numerous new word types due to typos and the use of acronyms. It could be a challenge to revise the word family lists to include such new members as *c* and *u* in the word families for *SEE* and *YOU*.

Specific Dice distance studies of the high-frequency vocabulary of the context-governed speech of academia could also be conducted using resources such as the British Academic Spoken English corpus and the Michigan Corpus of Academic Spoken English. Unfortunately, even combined, these two corpora would form a fairly small collection.

The differences between the high-frequency vocabulary of the natural and social sciences, that were conflated in this thesis, could also be studied in relation to a more general academic vocabulary. Using the TOEFL 2000 Spoken and Written Academic Language corpus (T2K-SWAL) Biber (2006: 41ff) found sizable differences between the vocabularies of different academic disciplines. Biber et al. (2004) lists words that are typical of written or spoken academic English or those that are neutral in regard to medium. Knowing how the vocabulary of academia fans out into specific disciplines would be of significant use to developers of EAP and ESP learning materials and tests. The T2K-SWAL was developed to ensure that the TOEFL tests are representative of spoken and written academic English in the US. For that reason, the corpus is also not accessible to the wider research community.

5.4 How does the scope of the definition of a word influence the resulting list?

The third study in this thesis compared two units for counting words: word types and word families. The one makes no assumptions about learners' knowledge of morphology and treats every surface form as an isolated word to be learned. The other unit, the word family, encourages learners to leverage a limited familiarity with some regular patterns in the morphology of English to substantially expand their known vocabulary. This third study explored how much the use of word families would actually expand token coverage in each of the text types. It was found that word families do offer a significant advantage in text coverage for each of the written text types and a noticeable boost in the coverage of conversation. There is also some indication that grouping by word families results in a more logical ordering of words since they accumulate the frequencies of closely related forms (c.f. Nagy et al., 1989).

5.4.1 Coverage and comprehension

The effect of using the word family as the unit of counting is most dramatic for general and academic writing with an approximately 10% increase in token coverage. Narrative writing gained an additional 6.4%, while conversation garnered around 4% more coverage. The motivation for considering the word family as a unit is based on the assumption

that “once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort” (Bauer & Nation, 1993: 253).

Of course, the goal of greater text coverage is increased comprehension. Laufer (1989) found that 95% token coverage typically allowed for adequate comprehension of texts, which Laufer defined as 55%. Hu and Nation (2000) found that most learners with a coverage of 100% scored 85.7% on a multiple-choice comprehension test and 56.5% on a recall test. Averaging these two scores they believed indicates that having some knowledge of all the words in the text should result in approximately 71% comprehension. Using linear regression Hu and Nation concluded that for unassisted comprehension ca. 98% coverage is needed.

Laufer (1992) approached comprehension from the angle of how many word families are known. In other words, she assessed approximately how many total word families were known by each learner rather than evaluating how many of the running words of a particular text were known. She found a minimum of 3,000 word families are needed for reasonable comprehension (56%) and 5,000 are required for 70% comprehension, which is similar to the level set by Hu and Nation (2000). Each of these studies assumed that proper nouns were known and were, therefore, included in the number of tokens covered.

In the word type / word family comparison in this thesis, proper nouns were included in the frequency lists as ordinary word types. Excluding proper nouns from the high-frequency lists and considering them to be known vocabulary would increase the coverage of each family list somewhat (Table 5.4). There are also some proper nouns that are not included in this percentage (Table 5.5) since they are not on Nation and Webb’s (2011) base word 31 list. (NB: the percentage of proper noun families in Table 5.4 cannot simply be added to the coverage figures in Table 5.5 since the number of tokens covered by the proper noun families in each list was not calculated. Adding 4-10% more word families to the known vocabulary of each text type would, of course, increase the number of tokens covered, but many of the proper noun families may represent only a few instances of that name.) Nonetheless, the coverage afforded by the most frequent families of each text type appears to be beyond or at least very close to a level for adequate comprehension. Taking into account the difference in how proper nouns were handled, the findings of the word type / word family study appear to correlate well with the results of the three studies above. The coverage for the most frequent 3,000 word types (Table 5.6), on the other hand, falls consistently below the level for reliable comprehension, even with the addition of proper nouns (Table 5.7), except in the case of spoken conversation.

Table 5.4

Token coverage for the most frequent 3,000 word families from a 10-million-token collection of the respective text types.

Token coverage	
General	90.1%
Academic	92.6%
Narrative	92.8%
Conversation	97.2%

Table 5.5

The percentage of proper nouns among the most frequent 3,000 word families from a 10-million-token collection of each text type.

Proper noun families	
General	7.4%
Academic	4.0%
Narrative	10.0%
Conversation	10.4%

Table 5.6

Token coverage for the most frequent 3,000 word types from a 10-million-token collection of the respective text types.

Token coverage	
General	80.9%
Academic	83.6%
Narrative	86.6%
Conversation	94.2%

Table 5.7

The percentage of proper nouns among the most frequent 3,000 word types from a 10-million-token collection of each text type.

Proper noun families	
General	4.8%
Academic	2.0%
Narrative	8.0%
Conversation	3.7%

With a more conservative design, Nation (2006) showed that 3,000 families derived from the BNC cover approximately 90% of the London Oslo Bergen corpus, which also represents British English. The most frequent 3,000 word families in each of the text types of this study covered a somewhat greater number of tokens (>90%) since token coverage was calculated for the one specific text type collection from which the 3,000 word families were

derived and not for all key four text types. Combining the proper nouns and the most frequent 3,000 word families from a text type should usually equal 95% coverage which Laufer (1992) considered a minimum for reasonable comprehension.

Nation (2006) argues that if proper nouns are considered known words, one still needs 5,000 word families to reach 98% coverage. The seemingly small difference between 95% and 98% coverage can make a significant difference in comprehension, however. Interestingly, since the top 3,000 word families from each text type overlap, combining all four lists would yield a total of 4,916 word families – remarkably close to Nation’s (2006) estimate. The amount of text coverage that would be achieved by this combined list has not been determined.

Table 5.8

The number of word families from each text type included in a combined list of the most frequent 3,000 word families from a 10-million token collection of each text type.

word families	
Conversation	3,427
Narrative	3,242
General	3,325
Academic	3,146
Total	4,916

Laufer and Ravenhorst (2011) looked at the correlation between vocabulary size and reading comprehension and estimated that a vocabulary of 5,000 word families are needed for minimal comprehension (95%) and around 8,000 word families for optimal comprehension (98%). One interesting result of Laufer and Ravenhorst’s (2011: 24) study is not specifically mentioned. Their Figure 1 shows an S-shaped pattern (or two) indicative of a dynamic system (c.f. Larsen-Freeman & Cameron, 2008: 85-88). Dynamic systems do not necessarily respond to changes in exactly the same magnitude as the changes that are made. Gradual changes to the environment or some system parameter may occur over time with little or no noticeable change in the system. Then, an additional small amount of change can push the system into a completely new state. This is the logic behind the warnings by climatologist that additional seemingly small changes to the earth’s climate could push it past a tipping point where there would be catastrophic changes to the state of the system, such as the interruption of the global jet stream.

Rather than seeing a strictly linear increase in comprehension with greater text coverage, Laufer and Ravenhorst's (2011) found a marked increase in the region of 2,000 to 3,000 known families. It seems that at that quantity learners reach a critical mass of vocabulary that allows for far greater comprehension. Therefore, 3,000 word families seems to be a good starting point for a basic-level core vocabulary. What appears to be a second S-curve peaks at around 6,000 word families, which would make a convenient break point for a second band. This level corresponds to the quantity Nation (2006) found was sufficient for regular conversational English. That would leave one additional band from 6,000-9,000 word families that should allow for comprehension of unsimplified general texts.

5.4.2 Family size and other issues

The third study also examined the number of types included among the high-frequency families of each text type and found the numbers to be possibly intimidating. The most frequent 1,000 word families of general writing comprised 6,557 word types (Table 5.9). That number dropped slightly in academic writing and more so in narrative writing and conversation.

Table 5.9

Types covered by the most frequent 3,000 families for 10-million-token samples of each of the four key text types.

text type	1,000 families	2,000 families	3,000 families
general	6,557	11,930	16,177
academic	6,479	11,728	16,037
Narrative	5,891	10,796	14,845
conversation	4,655	8,479	11,607

As was seen in Table 4.17a, the word families in the BNC/COCA lists (Nation & Webb, 2011) do not adhere strictly to the principles described in Bauer and Nation (1993). Some regional and archaic types were included in some common word families, since leaving them out would have created additional isolated families. These lists are not intended as a vocabulary learning resource for beginners. Therefore, core list creators should carefully examine the word families and edit the family list they use by frequency or a stricter following of the principles outlined in Bauer and Nation (1993).

Further work could be done in refining sets of word families designed for teaching and learning that avoid the presentation of rare, archaic or regional word family members. Further research could also be done on the validity of the assumptions behind word families. How transparent are the relationships between word family members? The coverage afforded by the lemma, the unit between the word type and the word family in size, has also not been explored.

The studies in this thesis were limited by their reliance on the orthographic word without any distinctions for part of speech or sense. Whereas, on the one hand the definition of words as orthographic forms mimics the experience learners have when encountering words in print, it may obscure distinctions that are important in production. The use of the orthographic word also ignores one aspect of the definition of the word family since word family members should be semantically related. The word form *bank*, for example, includes the noun referring to a business that offers financial services, but also the homograph that refers to the side of a river, an inclined section of road, a mass of clouds etc.

5.5 Summary

5.5.1 What this study has shown:

This thesis has demonstrated a relatively simple methodology based on the Dice coefficient for measuring the amount of variation between samples in regards to the frequency lists derived from them. What was discovered is that far larger vocabulary lists are feasible than was previously imagined. The first study (sections 3.3 and 4.1) showed that large quantities of data (20-50 million tokens) can be used effectively to produce very reliable vocabulary lists of many thousands of words. At times longer core vocabulary lists can even be more reliable than shorter ones. These “sweet spots” seem to indicate some related grouping of words has been artificially cut short in the shorter lists. The reliability of lists based on smaller-sized corpora can also be quantified using this methodology. An easy-to-use tool for measuring distance between texts or corpora has been provided in the form of an R-script in Appendix C.1-2.

A practical framework of four key text types for English has also been proposed (section 3.1.2), and it has been demonstrated that the frequent vocabulary of these four text types form an overlapping continuum (section 4.2). The most basic words are more frequently found in conversation and narrative writing, while the ranking of these words in conversation often better fits an appropriate order for teaching or learning. For example, the ranking of the

days of the week fell nicely within the top 1,000 of conversation and the months of the year within the top 2,000. Conversation should probably be the cornerstone of a core vocabulary list, followed closely by narrative writing which also describes many actions or events that are not often mentioned explicitly in conversation. General writing should likely come next since it contains more vocabulary items from the wider adult, educated, workplace-related world. Academic writing is even more specialized with a large quantity of specialized and technical vocabulary. Exactly how these four should be spliced together to form a single core vocabulary list is a matter for further experimentation, but each text type blends into the next rather than standing apart as a separate language. Each text type represents a varied expression of a single language.

The mathematical distribution of words in each text type has been shown to adhere to a Zipfian distribution with subtle differences that are indicative of each text type (section 4.1). Conversation contains a cluster of very frequent clauses at its core, while general and academic writing with their informational focus are dominated by the noun phrase as a structure. Narrative writing shows a pattern midway between conversation and general writing reflecting its dual oral and written provenance. The power-law distribution of word frequencies fits the overall vision of this thesis that the mental lexicon forms a network. A core vocabulary list is possible since the mental lexicon has a core of several thousand closely interlinked items. This network no doubt reflects the underlying neural network of the brain and is also mirrored in the macro-scale social networks which language makes possible.

The third study sections 3.5 and 4.3 demonstrated that that grouping word types as families can make a sizable difference in the coverage of spoken and written texts and should help learners to reach important comprehension thresholds. It has also been observed that increases in comprehension appear to follow a dynamic pattern. This and the observed word frequency distributions strongly imply that the lexicon is part of a complex adaptive system.

5.5.2 Applications

The results of these studies have numerous applications, most of which have to do with the teaching and learning of English.

First of all, these studies imply that ESL/EFL vocabulary learning priorities should and can be described for approximately 9,000 word families using some combination of the frequent words from each of the key text types. These should probably be in the form of 1,000-word bands. These lists could also lead to the refinement of vocabulary knowledge and size assessments.

Such lists should also guide the composition of graded listening and reading materials, especially materials in the region between 3,000 and 9,000 words in which so little is currently available (Nation, 2006; Schmitt & Schmitt, 2012). At present most series of graded readers end with a vocabulary of 3,000 words.

The word type list for conversation should be used in the creation of English language learning materials and teacher training materials that incorporate spoken forms or lexical preferences from different regional varieties of English. Educators and learners of English as an International Language need to recognize both the communicative value of common core vocabulary and the identification value of regional variants. A language teacher should, on the one hand, be ready to embrace local variants, while also helping students to be aware that these items may not be understood internationally. Such an awareness should be part of the education of teachers of English in all contexts, including English-dominant regions.

This thesis has also shown that researchers in psycholinguistics (Brysbaert & New, 2009; Burgess & Livesay, 1998) were on the right track when they looked for more conversation-like sources for reaction-time studies. A comparison of resources like Subtlex_{US} (Brysbaert & New, 2009) and authentic conversation using the Dice coefficient should lead to a refinement of these lists and their correlations in psychological familiarity research.

The taxonomy of English texts used in this thesis should also stimulate and guide the creation of text-type-specific learning resources or help to highlight the importance of these text types in the curriculum. The narrative-imaginative text type is at the heart of any good extensive reading program. Learners may question the value of reading stories, however, if their goals are the use of English in business or science. This thesis has shown that the narrative-imaginative text type is an excellent source of core vocabulary that is essential in interpersonal interaction. Making successful connections at conferences or tradeshow relies at least as much on the forming of personal connections as it does on accurately communicating technical information.

The findings of this thesis indicate that the vocabulary of narratives is at least two steps away from academic writing. This should be a valuable insight to EFL curriculum designers. EFL students are often asked to make the leap from graded conversation and narratives to ungraded academic articles. The lack of graded informational texts and courses that use them usually makes the transition to academic English abrupt and painful. In ESL and EFL contexts, it is quite common to hear teachers or learners make comments such as “I thought my/his/her English was pretty good, but after that assignment...” A learner who can hold a fluent casual conversation may still find academic reading and writing a daunting challenge.

Courses with titles like “newspaper English” are common in EFL curricula. In order to bridge spoken English and academic writing, the focus of such courses or textbooks should be expanded to general informational writing. Reportage makes up the lion’s share of general writing, but the vocabulary of the general writing text type is far more important than a name like “newspaper English” implies. Many learners may dismiss it thinking “But, I’m not interested in reading the newspaper.” General writing is, however, the original and most basic written text type and an important stepping stone toward the development of more technical informational prose. Courses that focus on genre from the general writing text type should be an essential part of any EFL academic curriculum. An examination of English L1 vocabulary learning and literacy could also benefit from this text-type approach. If teachers and learners in all contexts are aware of the overlapping, and yet differing nature of each text type, this can help to avoid unnecessary frustration or false expectations as learners move from one type of text to another. The application of Biber’s multidimensional study of English as a model of how to classify the texts of the language should inspire MD analyses and similar applications in other languages. MD analyses have already been made of Somali, Korean and Nukulaelae Tuvuluan, Taiwanese and Spanish (Biber, 1995; Jang, 1998). These analyses “uncovered dimensions relating to interactiveness/involvement versus informational focus, the expression of personal stance, and narrative versus non-narrative discourse” (Biber, 1995: chapter 7; 2004: 16). The first dimension is a key distinction between the *conversation* and *general writing* text types used in this thesis. The expression of an objective stance is a key feature of the *academic* text type, especially the natural science texts. The *narrative* text type is, of course, typical of the narrative dimension. This implies that a similar basic set of text types should be found in these languages as well.

Comparisons of the text types for other languages and the vocabulary in them could highlight universal patterns in the structure of the lexicons of human languages. Language learning resources that highlight commonalities or differences in the way text types are realized in different languages would be highly useful to those finding their way in a new language. One could also envision exciting multi-lingual language learning resources. For example, the *Longman Student Grammar of Spoken and Written English* (Biber, Conrad, & Leech, 2002: 108) lists “[t]he twenty most common ‘mental’ verbs in conversation, fiction, newspaper writing and academic prose.” This corpus-based approach to prioritizing learners’ focus on vocabulary learning and grammar is exactly what is needed in language learning materials. This could be further improved by showing learners the relative ranking of each of these terms in each of the four key text types. Which terms would sound more natural in conversation and which in a journal article? Placing this information side by side with a study

of the most common mental activity verbs in the texts types of Chinese, for example, would be another possible step. I look forward to studying (or co-authoring) such resources!

BIBLIOGRAPHY

- Adolphs, S., & Schmitt, N. (2003). Lexical coverage of spoken discourse. *Applied Linguistics*, 24(4), 425–438 doi:10.1093/applin/24.4.425
- Aitchison, J. (2000). *Language change: Progress or decay?*. Cambridge: Cambridge University Press.
- Albert, R., Jeong, H., & Barabási, A.-L. (1999). The diameter of the WWW. *Nature*, 401(6749), 130–131.
- Anthony, L. (2011). AntConc (3.2.4w) [Computer software]. Retrieved from <http://www.antlab.sci.waseda.ac.jp/software.html>
- Barabasi, A. L., & Bonabeau, E. (2003). Scale-free networks. *Scientific American*, 288(5), 50–59.
- Barlow, M., & Kemmer, S. (Eds.) (2000). *Usage-based models of language*. Stanford: CSLI Publications.
- Bauer, L. (1993). *Wellington corpus of written New Zealand English*. Wellington, New Zealand: Victoria University of Wellington.
- Bauer, L., & Nation, I. S. P. (1993) Word families. *International Journal of Lexicography*, 6(4), 253–279.
- Beier, E., Starkweather, J., & Miller, D. (1967). Analysis of word frequencies in spoken language of children. *Language and Speech*, 10, 217–227.
- Bell, H. (2012). Core vocabulary. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley-Blackwell.
- Biber, D. (1988). *Variation across speech and writing*. Cambridge: Cambridge University Press.

- Biber, D. (1989). A typology of English texts, *Linguistics*, 27, 3-43.
- Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, 8(4), 243–257.
- Biber, D. (1995). *Dimensions of cross-linguistic variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.
- Biber, D. (2006). *University language: A corpus-based study of spoken and written registers*. Amsterdam: John Benjamins.
- Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.
- Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English*. Harlow: Longman.
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.
- Biber, D., Conrad, S., Reppen, R., Byrd, P., Helt, M., Clark, V., Cortes, V., Csomay, E., & Urzua, A. (2004). *Representing language use in the university: Analysis of the TOEFL 2000 Spoken and Written Academic Language Corpus*. Princeton, New Jersey: Educational Testing Service. Available online at <http://www.ets.org/Media/Research/pdf/RM-04-03.pdf>
- Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow: Longman.
- Bloomfield, L. (1933). *Language*. New York: Holt.
- Boas, F. (Ed.). (1911/1969). *Smithsonian institution, bureau of American ethnology bulletin 40: Handbook of American Indian languages 1* (pp. 1–83). Oosterhout: Anthropological Publications. (Original work published 1911)
- Bongers, H. (1947). *The history and principles of vocabulary control*. Woerden: Wocopi.

- Bowen, J. D., Madson, H., Hilfery, A. (1985). *TESOL, techniques and procedures*. USA: Newbury House Publishers, Inc.
- Bright, J. A., & McGregor, G. P. (1970). *Teaching English as a second language*. London: Longman.
- Brown, D. (2010). An improper assumption? The treatment of proper nouns in text coverage counts. *Reading in a Foreign Language*, 22(2), 355–361.
- Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. doi:10.3758/BRM.41.4.977
- Buchanan, M. (2002). *Nexus: Small worlds and the groundbreaking science of networks*. New York: W. W. Norton & Company.
- Burgess, C., & Livesay, K. (1998). The effect of corpus size in predicting reaction time in a basic word recognition task: Moving on from Kucera and Francis. *Behavior Research Methods, Instruments, & Computers*, 30, 272–277.
- Burrows, J. F. (2002). ‘Delta’: A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing*, 17(3), 267–287.
- Burrows, J. F. (2007). All the way through: Testing for authorship in different frequency strata. *Literary and Linguistic Computing*, 22(1), 267–287.
- Bybee, J. (1985). *Morphology*. Amsterdam: John Benjamins.
- Bybee, J. (1995). Regular morphology and the lexicon. *Language and cognitive processes*, 10, 425–455.
- Bybee, J. (2006). *Frequency of use and the organization of language*. Oxford University Press, U.S.A.

- Bybee, J. (2010). *Language, usage, and cognition*. Cambridge, UK: Cambridge University Press.
- Bybee, J., & Hopper, P. (Eds.). (2001). *Frequency and the emergence of linguistic structure*. Amsterdam: John Benjamins
- Campion, M. E., & Elley, W. B. (1971). *An academic vocabulary list*. Wellington: NZCER.
- Carroll, J. B. (1972). A new word frequency book. *Elementary English*, 49, 1070–1074.
- Carroll, J. B., Davies, P., & Richman, B. (1971). *The American heritage word frequency book*. New York: Houghton Mifflin, Boston American Heritage.
- Carter, R. (1982). A note on core vocabulary. *Nottingham Linguistic Circular*, 11(2), 39–51.
- Carter, R. (1987a). *Vocabulary*. London: Allen & Unwin.
- Carter, R. (1987b). Is there a core vocabulary? *Applied Linguistics*, 8, 178–193.
- Carter, R. (2012). *Vocabulary: Applied linguistic perspectives*. New York: Routledge.
- Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide*. Cambridge: Cambridge University Press.
- Chen, T., & Kan, M.-Y. (2013). Creating a live, public short message service corpus: the NUS SMS corpus. *Language Resources and Evaluation*, 47(2), 299–335. doi: 10.1007/s10579-012-9197-9
- Chen, Y. (2012a). Zipf's law, 1/f noise, and fractal hierarchy. *Chaos, Solitons & Fractals*, 45, 63–73.
- Chen, Y. (2012b) The mathematical relationship between Zipf's law and the hierarchical scaling law. *Physica A*, 391, 3285–3299.

- Chou, C.-H., & Wang, C. (2006). Genetic relatedness between *Listeria monocytogenes* isolates from seafood and humans using PFGE and REP-PCR. *International Journal of Food Microbiology* 110(2), 135–148.
<http://dx.doi.org/10.1016/j.ijfoodmicro.2006.02.003>
- Chung, T. M. (2003). A corpus comparison approach for terminology extraction. *Terminology*, 9(2), 221–246.
- Clancy, S. (2008). Genetic Mutation. *Nature Education* 1(1). Retrieved from
<http://www.nature.com/scitable/topicpage/genetic-mutation-441>
- Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661-703. (arXiv:0706.1062) doi:10.1137/070710111
- Coady, J., & Huckin, T. (Eds.) (1997). *Second language vocabulary acquisition: A rationale for pedagogy*. Cambridge: Cambridge University Press.
- Collins WordBanks Online. (2009). Wordbanks online: English corpus. Retrieved from
http://wordbanks.harpercollins.co.uk/Docs/WBO/WordBanksOnline_English.html
- Cook, W. A., & O'Shea, M. V. (1914). *The child and his spelling: an investigation of the psychology of spelling, individual and sex differences in spelling abilities and needs, the character and range of the spelling vocabulary, and the practical problems of teaching spelling*. Indianapolis, Indiana: The Bobbs-Merrill company.
- Cowie, A. P. (1999). *English dictionaries for foreign language learners: A history*. Oxford: Clarendon Press.
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238.
- Crabbe, D., & Nation, I. S. P. (1991). A survival language learning syllabus for foreign travel. *System*, 19(3), 191–201.
- Croft, W. (2001). *Radical construction grammar*. Oxford: Oxford University Press.

- Crystal, D. (2003). *The Cambridge encyclopedia of the English language* (2nd ed.). Cambridge: Cambridge University Press.
- Daulton, F. (2003). The high-frequency vocabulary of English. *The Ryukoku Journal of Humanities and Science*, 24(2), 19–41.
- Davies, M. (2008-). *The corpus of contemporary American English: 450 million words, 1990-present*. Available online at <http://corpus.byu.edu/coca/>.
- Davies, M. (2010). The Corpus of Contemporary American English as the first reliable monitor corpus of English. *Literary & Linguistic Computing*, 25(4): 447–464. doi:10.1093/lc/fqq018
- Davies, M. (2013). *Corpus of Global Web-Based English: 1.9 billion words from speakers in 20 countries*. Available online at <http://corpus2.byu.edu/glowbe/>.
- Davies, M., & Gardner, D. (2010). A frequency dictionary of contemporary American English: Word sketches, collocates and thematic lists. Routledge. Free 5000-word list retrieved from: <http://www.wordfrequency.info/free.asp>
- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26 (3), 297–302.
- Du Bois, J. W., Chafe, W. L., Meyer, C., Thompson, S. A., Englebretson, R., & Martey, N. (2000-2005). *Santa Barbara corpus of spoken American English, Parts 1-4*. Philadelphia: Linguistic Data Consortium.
- EcoRover. (2010). Grid+Layout [Photograph]. Retrieved from http://ecorover.blogspot.co.nz/2010_08_01_archive.html
- EcoRover. (2010). Grid+Photo [Photograph]. Retrieved from http://ecorover.blogspot.co.nz/2010_08_01_archive.html
- Eldridge, R. C. (1911). *Six thousand common English words; Their comparative frequency and what can be done with them*. Buffalo: The Clement Press.

- Estoup, J.-B. (1916). *Les gammes stenographiques*. Paris: Institut Stenographique de France.
- Faloutsos, M., Faloutsos, P., & Faloutsos, C. (1999). On power-law relationships of the Internet topology. Retrieved from:
<http://www.cis.upenn.edu/~mkearns/teaching/NetworkedLife/power-internet.pdf>
- Ferraresi, A., Zanchetta, E., Baroni, M., & Bernardini, S. (2008). Introducing and evaluating ukWaC, a very large web-derived corpus of English. In Evert, S., Kilgarriff, A., & Sharoff, S. (Eds.), *Proceedings of the 4th Web as Corpus Workshop (WAC-4) Can we beat Google* (pp. 47–54). Marrakech, Morocco.
- Ferrer-i-Cancho, R., & Solé, R. V. (2001a). The small-world of human language. *Proceedings of The Royal Society of London. Series B, Biological Sciences*, 268(1482), 2261–2265.
- Ferrer-i-Cancho, R., & Solé, R. V. (2001b). Two regimes in the frequency of words and the origins of complex lexicons: Zipf's law revisited. *Journal of Quantitative Linguistics*, 8(3), 165–173.
- Fillmore, C. J., Ide, N., Jurafsky, D., & Macleod, C. (1998). An American national corpus: A proposal. Paper presented at the First International Language Resources and Evaluation Conference, Granada, Spain. Retrieved from
<http://www.cs.vassar.edu/~ide/papers/anc.final.pdf>.
- Fillmore, C. J., & Kay, P. (1993). *Construction Grammar*. Berkeley: University of California.
- Fillmore, C. J., Kay, P., & O'Connor, C. (1988). Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64, 501–538.
- Fludernik, M. (1996). *Towards a 'Natural' Narratology*. London: Routledge.
- Folch, H., Heiden, S., Habert, B., Fleury, S., Illouz, S., Lafon, P., Nioche, J., & Prévost, S. (2000). TypTex : Inductive typological text classification by multivariate statistical analysis for NLP systems tuning/evaluation. *Proceedings of the Second Language Resources and Evaluation Conference, LREC 2000* (pp. 141–148), Athens.

- Fries, C. C., & Traver, A. A. (1960). *English word lists*. Ann Arbor: George Wahr.
- Ghadessy, M. (1979). Frequency counts, word lists, and materials preparation: a new approach. *English Teaching Forum*, 17(1), 24–27.
- Giesen, K., Zimmermann, A., & Suedekum, J. (2010). The size distribution across all cities – double Pareto lognormal strikes, *Journal of Urban Economics*, 68, 129–137.
- Gnanadesikan, A. E. (2009). *The writing revolution: Cuneiform to the internet*. Chichester, UK: Wiley-Blackwell.
- Goldberg, A. E. (1995). *A construction grammar approach to argument structure*. Chicago: The University of Chicago Press.
- Gougenheim G., Michéa, R., Rivenc, P., & Sauvageot, A. (1964). *L'élaboration du français fondamental : 1er degré*. Paris: Didier.
- gramarye. (2012, November 23). Learn English vocabulary - which English words to learn? [web log]. Retrieved from <http://gramarye.hubpages.com/hub/How-to-learn-English-vocabulary-which-words>.
- Greenbaum, S. (Ed.). (1996). *Comparing English worldwide: The international corpus of English*. Oxford: Clarendon Press.
- Greenbaum, S., & Svartvik, J. (1990). The London corpus of spoken English: Description and research. In J. Svartvik (Ed.), *The London-Lund Corpus of Spoken English: Description and research* (pp. 11–59). Lund University Press.
- Greenspahn, F. E. (1986). Abraham Ibn Ezra and the origin of some medieval grammatical terms. *The Jewish Quarterly Review*, 76(3), 217–227.
- Gries, S. T. (2008e) Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, 13(4), 403–437.

- Gries, S. T. (2009d). *Statistics for linguistics with R: A practical introduction*. Berlin: De Gruyter Mouton.
- Gries, S. T. (2009e). *Quantitative corpus linguistics with R: A practical introduction*. New York: Routledge.
- Gries, S. T. (2010c). Dispersions and adjusted frequencies in corpora: Further explorations. In S. T. Gries, S. Wulff, & M. Davies (Eds.), *Corpus linguistic applications: Current studies, new directions* (pp. 197–212). Amsterdam: Rodopi.
- Halliday, M. A. K., McIntosh, A., & Stevens, P. (1964). *The linguistic sciences and language teaching*. London: Longmans.
- HarperCollins Publishers. (2004a). The bank of English [Electronic Version]. Retrieved from <http://www.collins.co.uk/books.aspx?group=153>.
- HarperCollins Publishers. (2004b). Collins word web [Electronic Version]. Retrieved from <http://www.collins.co.uk/books.aspx?group=180>.
- Hahsler, M., Grün, B., Hornik, K., & Buchta, C. (2008). Introduction to arules – A computational environment for mining association rules and frequent item sets. Retrieved from <http://cran.r-project.org/web/packages/arules/vignettes/arules.pdf>
- Herdağdelen, A. (2013). Twitter n-gram corpus with demographic metadata. *Language Resources and Evaluation*, 1–21. doi: 10.1007/s10579-013-9227-2
- Hindmarsh, R. (1980). *Cambridge English Lexicon*. Cambridge: Cambridge University Press.
- Hoffmann, S., Evert, S., Smith, N., Lee, D. Berglund Prytz, Y. (2008). *Corpus linguistics with BNCweb - A practical guide*. Frankfurt am Main: Peter Lang International Academic Publishers
- Hofland, K., & Johansson, S. (1982). *Word frequencies in British and American English*. London: Longman

- Holmes, J., Vine, B., & Johnson, G. (1998). *Wellington Corpus of Spoken New Zealand English*. Wellington: School of Linguistics and Applied Language Studies, Victoria University of Wellington.
- Hoover, D. L. (2004a). Testing Burrows' Delta. *Literary and Linguistic Computing*, 19(4), 453–475.
- Horn, E. (1926). A basic writing vocabulary. *Monographs in Education*, No. 4. University of Iowa Press.
- Howatt, A. P. R., & Widdowson, H. G. (2004). *A history of English language teaching*. Oxford : Oxford University Press.
- Howes, D. H. (1966). A word count of spoken English. *Journal of Verbal Learning and Verbal Behaviour*, 5, 572–606.
- Hsu, W. (2009). College English textbooks for general purposes: A corpus-based analysis of lexical coverage. *Electronic Journal of Foreign Language Teaching* 6(1), 42–62.
- Hu, M., & Nation, I. S. P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, 13(1), 403–430.
- Hundt, M., Sand, A., & Siemund, R. (Eds.). (1998). *Manual of information to accompany the Freiburg–LOB Corpus of British English*. Freiburg: Department of English, Freiburg University.
- Hundt, M., Sand, A., & Skandera, P. (Eds.). (1999). *Manual of information to accompany the Freiburg–Brown Corpus of American English*. Freiburg: Department of English, Freiburg University.
- Hunston, S. (2002). *Corpora in applied linguistics*. Cambridge: Cambridge University Press.
- Hutchby, I., & Wooffitt, R. (1998). *Conversation analysis. principles, practices, applications*. Cambridge: Polity.

- Hyland, K., & Tse, P. (2007). Is there an “academic vocabulary”? *TESOL Quarterly*, 41(2), 235–253.
- Ide, N., & Suderman, K. (2006). Integrating linguistic resources: The American national corpus model. Paper presented at the 5th annual meeting of the Language Resources and Evaluation Conference, Genoa, Italy. Retrieved from <http://www.cs.vassar.edu/~ide/papers/ANC-LREC06.pdf>
- Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., & Suchomel, V. (2013). The TenTen corpus family. International Conference on Corpus Linguistics, Lancaster. Retrieved from http://trac.sketchengine.co.uk/raw-attachment/wiki/AK/Papers/tentens_14may2013.docx
- Jang, S.-Ch. (1998). *Dimensions of spoken and written Taiwanese: A corpus-based register study* (Ph.D. dissertation, University of Hawaii).
- Jaworski, A., & Coupland, N. (Eds.). (1999). *The discourse reader*. London: Routledge.
- Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers*: Department of English, University of Oslo.
- Josselson, H. H. (1953). *The Russian word count and frequency analysis of grammatical categories of standard literary Russian*. Detroit: Wayne University Press.
- Joyal, M., McDougall, I., & Yardley, J. C. (2009). *Greek and Roman education: A sourcebook*. Oxon: Routledge.
- Julien, M. (2006). Word. In K. Brown (Ed.), *Encyclopedia of language and linguistics* (2nd ed) (2nd ed., Vol. 13, pp. 617-624). Oxford: Elsevier.
- Kachru, B. B. (1992). Meaning in deviation: Toward understanding non-native English texts. In B. B. Kachru (Ed.) *The other tongue: English across cultures* (2nd ed.) (pp. 301–326). Urbana: University of Illinois Press.

- Kachru, B. B. (1985). Standards, codification and sociolinguistic realism: The English language in the outer circle. In R. Quirk & H. G. Widdowson (Eds.) *English in the World: teaching and learning the language and literatures* (pp. 11–36). Cambridge: Cambridge University Press.
- Kennedy, G. (1998). *An introduction to corpus linguistics*. London: Longman.
- Kilgariff, A. (1997). Putting frequencies in the dictionary. *International Journal of Lexicography*, 10(2), 135–155.
- Kilgariff, A. (2001). Comparing corpora. *International Journal of Corpus Linguistics*, 6(1), 97–133.
- Kilgariff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics* 29(3), 333–347.
- Kirkpatrick, A. (2007). *World Englishes: Implications for International Communication and English Language Teaching*. Cambridge: Cambridge University Press.
- Klaus, A., Yu, S., & Plenz, D. (2011). Statistical analyses support power law distributions found in neuronal avalanches. *PLoS ONE* 6(5), e19779. doi:10.1371/journal.pone.0019779
- Knowles, G., & Don, Z. M. (2004). The notion of a “lemma”: Headwords, roots and lexical sets. *International Journal of Corpus Linguistics*, 9(1), 69–81.
- Kobeleva, P. P. (2012). Second language listening and unfamiliar proper names: Comprehension barrier? *RELC Journal*, 43(1), 83–98. doi:10.1177/0033688212440637
- Kučera, H., & Francis, W. N. (1967). *A computational analysis of present-day American English*. Providence, R.I.: Brown University Press.
- Kumar, R., Raghavan, P., Rajagopalan, S., & Tomkins, A. (1999). Trawling the Web for emerging cyber-communities, *Computer Networks*, 31, 1481–1493.

- Labov, W. (1972a). *Language in the inner city*. Philadelphia: University of Pennsylvania Press.
- Labov, W. (1972b). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Labov, W., & Waletzky, J. (1967). Narrative analysis: oral versions of personal experience. In J. Helm (Ed.), *Essays on the verbal and visual arts* (pp. 12–44). Seattle: University of Washington Press.
- Langacker, R. W. (1982). Space grammar, analysability, and the English passive. *Language*, 58(1) 22–80.
- Langacker, R. W. (1987). *Foundations of cognitive grammar* (Vol. 1), *Theoretical prerequisites*. Stanford: Stanford University Press.
- Langacker, R. W. (1988). A usage-based model. In B. Rudzka-Ostyn (Ed.), *Topics in cognitive linguistics* (pp. 127–161). Amsterdam: John Benjamins.
- Langacker, R. W. (1990). *Concept, image, and symbol: The cognitive basis of grammar*. (*Cognitive Linguistics Research 1.*) Berlin/New York: Mouton de Gruyter.
- Langacker, R. W. (1991). *Foundations of cognitive grammar* (Vol. 2), *Descriptive application*. Stanford: Stanford University Press.
- Langacker, R. W. (2000). A dynamic usage-based model, in M. Barlow & S. Kemmer (Eds.), 24–63.
- Langacker, R. W. (2008). *Cognitive grammar: A basic introduction*. New York: Oxford University Press.
- Larsen-Freeman, D., & Cameron, L. (2008). *Complex Systems and Applied Linguistics*. Oxford, UK: Oxford University Press.

- Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316–323). Clevedon: Multilingual Matters.
- Laufer, B. (1992). How much lexis is necessary for reading comprehension? In H. Bejoint & P. Arnaud (Eds.), *Vocabulary and applied linguistics* (pp.126–132). Basingstoke & London: Macmillan.
- Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, 22(1), 15–30.
- Lee, Y. W. D. (2001). Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British national corpus. *Journal of English Linguistics*, 29(3), 250–278.
- Leech, G., & Fallon, R. (1992). Computer corpora-What do they tell us about culture? *ICAME Journal*, 16. 29–50.
- Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal*, 16(1), 33–42.
- Longman Communication 3000. (n.d.) Retrieved from http://www.lex tutor.ca/freq/lists_download/longman_3000_list.pdf
- Lucas, A. (2012). Package 'amap'. Retrieved from cran.r-project.org/web/packages/amap/amap.pdf
- Lutosławski, W. (1890). *Principes de stylométrie*.
- Lyne, A. A. (1985). *The vocabulary of French business correspondence: Word frequencies, collocations, and problems of lexicometric method*. Genève: Slatkine.
- Lynn, R. E. (1973). Preparing word lists: a suggested method. *RELC Journal*, 4(1), 25–32.

- Maas, U., & Wunderlich, D. (1972). *Pragmatik und sprachliches handeln*. Frankfurt: Athenaeum.
- Mandelbrot, B. (1953). An informational theory of the statistical structure of language. In W. J. Betterworth (Ed.), *Symposium on Applications of Communications Theory*, London. (pp. 486–500).
- Mandelbrot, B. (1982). *The Fractal Geometry of Nature*. New York, W. H. Freeman and Company.
- Manning, C., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.
- Matsushita, T. (2012). *In what order should learners learn Japanese vocabulary? A corpus-based approach* (Doctoral thesis, Victoria University of Wellington, Wellington, New Zealand). <http://victoria.lconz.ac.nz/vwebv/holdingsInfo?bibId=1520372>
- McArthur, T. (1998). *Living words: Language, lexicography and the knowledge revolution*. Exeter: University of Exeter Press.
- McCarthy, M. J. (1998). *Spoken language and applied linguistics*. Cambridge: Cambridge University Press.
- McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. New York: Routledge.
- Meara, P. (1980). Vocabulary Acquisition: A Neglected Aspect of Language Learning. *Language Teaching*, 13(3-4), 221–246. doi:10.1017/S0261444800008879
- Meara, P., & Jones, G. (1987). Tests of vocabulary size in English as a foreign language. *Polyglot*, 8(1), 1–40.
- Meara, P., & Jones, G. (1990). Eurocentres vocabulary size test (10KA). Zurich: Eurocentres Learning Service.

- Mitzenmacher, M. (2003a). A brief history of generative models for power law and lognormal distributions. *Internet Mathematics* 1(2), 226–251
- Montemurro, M. A. (2001). Beyond the Zipf-Mandelbrot law in quantitative linguistics. *Physica A* 300, 567–578.
- Moon, R. (1998). *Fixed expressions and idioms in English: A corpus-based approach*. Alderley: Clarendon Press.
- Mosteller, F., & Wallace, D. (1964). *Inference and disputed authorship: The federalist*. Reading, MA: Addison-Wesley.
- Murison-Bowie, S. (1993). *MicroConcord: Manual*. Oxford: Oxford University Press.
- Nagy, W. E. (1997). On the role of context in first- and second-language vocabulary learning. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6–19). Cambridge: Cambridge University Press
- Nagy, W. E., Anderson, R. C., Schommer, M., Scott, J. A., & Stallman, A. C. (1989). Morphological families in the internal lexicon. *Reading Research Quarterly*, 24(3), 262–282.
- Nagy, W. E., Herman, P. A., & Andersen, R. C. (1985). Learning words from context. *Reading Research Quarterly*, 20(2), 233–253.
- Nattinger, J. R., & Decarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.
- Nation, I.S.P. (1983). Learning vocabulary. *New Zealand Language Teacher*, 9(1), 10–11.
- Nation, I.S.P. (1990). *Teaching and Learning Vocabulary*. New York: Newbury House.
- Nation, I.S.P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

- Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–13). Amsterdam: John Benjamins.
- Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, 63(1), 59–82.
- Nation, I. S. P. (to appear). Which words do you need? In J. R. Taylor (Ed.), *The Oxford Handbook of the Word*. Oxford: Oxford University Press.
- Nation, I. S. P. & Beglar, D. (2007) A vocabulary size test. *The Language Teacher*, 31(7), 9–13.
- Nation, I. S. P., & Heatly, A. (2002). Range: A program for the analysis of vocabulary in texts [Computer software]. Retrieved from <http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx>
- Nation, I. S. P., & Meara, P. (2002) Vocabulary. In N. Schmitt (ed.) *An Introduction to Applied Linguistics*. Edward Arnold: 35–54.
- Nation, I. S. P., & Newton, J. (2009) *Teaching ESL/EFL Listening and Speaking*. New York: Routledge.
- Nation, I. S. P., & Webb, S. (2011) *Researching and analyzing vocabulary*. Boston: Heinle Cengage Learning.
- Nelson, G. (2006a). The core and periphery of world Englishes: a corpus-based exploration. *World Englishes*, 25(1), 115–129.
- Nelson, G. (2006b). World Englishes and corpora studies. In B. Kachru, Y. Kachru & C. L. Nelson (Eds.), *The handbook of world Englishes* (pp. 733–750). Oxford: Blackwell Publishing.
- Nisbet, J. D. (1960). Frequency counts and their uses. *Educational Research*, 3, 51–64.

- Ogden, C. K. (1930). *Basic English: A general introduction with rules and grammar*. London: Kegan Paul
- Ogden, C. K. (1935). *Counter-offensive: An exposure of certain misrepresentations of Basic English*. London: The Orthological Institute.
- Ornstein, J. (1954). Review: The Russian word count and frequency analysis of grammatical categories of standard literary Russian by Josselson, Harry H. *The Modern Language Journal*, 38(8), 428–429.
- Palmer, H. E. (1917). The scientific study & teaching of languages. In IRET (Ed.), *The selected writings of Harold E. Palmer* (2nd ed., Vol. 1, pp. 5–330). Tokyo: IRET.
- Palmer, H. E. (1934). The grading and simplifying of literary material: A memorandum. In IRET (Ed.), *The selected writings of Harold E. Palmer* (2nd ed., Vol. 2, pp. 661–772). Tokyo: IRET.
- Palmer, H. E. (1936). The history and present state of the movement towards vocabulary control. In R. C. Smith (Ed.), *Teaching English as a foreign language, 1912-1936: Pioneers of ELT* (Vol. V, pp. 369–380): Routledge.
- Palmer, H. E., & Hornby, A. S. H. (1937). *Thousand-word English*. London: George C. Harrap and Co. Ltd.
- Palmer, H. H. (1938). *A Grammar of English Words*. London: Longmans, Green.
- Parent, K. (2012). The most frequent English homonyms. *RELC Journal*, 43(1), 69–81.
- Peters, P. (1987). Towards a corpus of Australian English. *International Computer Archive of Modern English (ICAME) JOURNAL*, 11 (1987), 27–38.
- Peyawary, A. S. (1999). *The Core Vocabulary of international english: A corpus approach*. (Vol. 2). Bergen.

- Pomikálek, J., Rychlý, P., & Kilgariff, A. (2009). Scaling to billion-plus word corpora. *Advances in Computational Linguistics [Special issue]. Research in Computing Science 41*, 3–14.
- Popescu, M., & Dinu, L. P. (2008). Rank distance as a stylistic similarity. In *Coling 2008: Companion volume – Posters and Demonstrations* (pp. 91–94).
- Praninskas, J. (1972). *American university word list*. London: Longman.
- Proctor, P. (Ed.). (1987). *Longman dictionary of contemporary English*. London: Longman.
- R Core Team. (2012). R: A Language and Environment for Statistical Computing [Computer software]. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Rayson, P., Leech, G., & Hodges, M. (1997). Social differentiation in the use of English vocabulary: some analyses of the conversational component of the British National Corpus. *International Journal of Corpus Linguistics*, 2(1), 133–152.
- Richards, J. C. (1974). Word lists: Problems and prospects. *RELC Journal*, 5(2), 69–84. doi: 10.1177/003368827400500207
- Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, 8(3), 382–439.
- Rosenblatt, L. (1978). *The reader, the text, the poem: The transactional theory of the literary work*. Southern Illinois University Press.
- Sacks, H., Schegloff, E. A., & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50, 696–735.
- Sadock, J. M. (1974). *Towards a linguistic theory of speech acts*. New York: Academic Press.
- Saenger, P. (1997). *Space between words: The origins of silent reading*. Stanford: Stanford University Press.

- Sánchez, S., García-Sánchez, A., Martínez, R., Blanco, J., Blanco, J. E., Blanco, M., Dahbi, G., Mora, A., Hermoso de Mendoza, J., Alonso, J. M., Rey, J. (2009). Detection and characterisation of Shiga toxin-producing *Escherichia coli* other than *Escherichia coli* O157:H7 in wild ruminants. *The Veterinary Journal* 180(3), 384–388.
<http://dx.doi.org/10.1016/j.tvjl.2008.01.011>
- Scannell, J. W. (1997). Determining cortical landscapes. *Nature*, 386, 452.
- Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge: Cambridge University Press.
- Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1), 26–43. doi: 10.1111/j.1540-4781.2011.01146.x
- Schmitt, N., & Schmitt, D. (2012). A reassessment of frequency and vocabulary size in L2 vocabulary. *Language Teaching*, 1–20. doi:10.1017/S0261444812000018
- Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, 18(1), 55–88.
- Scholes, R., Phelan, J., & Kellogg, R. (2006). *The Nature of Narrative: Revised and Expanded*. Oxford: Oxford University Press.
- Schonell, F. J., Meddleton, I. G., & Shaw, B. A. (1956). *A study of the oral vocabulary of adults*. Brisbane: University of Queensland Press.
- Shastri, S. V. (1986). *Manual of information to accompany the Kolhapur corpus of Indian English, for use with digital computers*. Kolhapur: Department of English, Shivaji University.
- Shew, W. L., Yang, H., Yu, S., Roy, R., & Plenz, D. (2011). Information capacity and transmission are maximized in balanced cortical networks with neuronal avalanches. *The Journal of Neuroscience*, 31(1), 55–63.

- Smith, L. (1976) English as an international auxiliary language. *RELC Journal*, 7(2), 38–42.
- Smith, R. C. (Ed.). (2003). *Teaching English as a foreign language, 1912-1936: Pioneers of ELT* (Vol. V) Routledge.
- Sorell, C. J. (2012). Zipf's law and vocabulary. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics*. Oxford: Wiley-Blackwell.
- Sorell, C. J. (to appear). Word Frequencies. In J. R. Taylor (Ed.), *The Oxford handbook of the word*. Oxford: Oxford University Press.
- Sorell, C. J., Chen, S.-R., & Lin, L.-C. (2013). *A corpus of childhood English language teaching*. Manuscript in preparation.
- Stern, H. H. (1983). *Fundamental concepts of language teaching*. Oxford: Oxford University Press.
- Stubbs, M. (1986). Language development, lexical competence and nuclear vocabulary. In K. Durkin (Ed.), *Language Development in the School Years*. London: Croom Helm.
- Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, 25(2), 34–50.
- Swadesh, M. (1971). *The origin and diversification of language*. Edited post mortem by Joel Sherzer. Chicago: Aldine.
- Svartvik, J. (Ed.). (1990). *The London-Lund Corpus of spoken English: Description and research*. Lund: Lund University Press.
- Svartvik, J., & Quirk, R. (Eds). (1980). *A corpus of English conversation*. Lund: C. W. K. Gleerup.
- Taylor, J. R. (2002) *Cognitive grammar*. Oxford Textbooks in Linguistics. Oxford: Oxford University Press.

- Thornbury, S. (2002). *How to teach vocabulary*. Harlow: Longman.
- Thorndike, E. L. (1921). *The Teacher's Word*. New York: Teachers' College, Columbia University.
- Thorndike, E. L. (1932). *The Teacher's Word Book of 20,000 Words*. New York: Teachers' College Columbia University.
- Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers' College Columbia University.
- Tickoo, M. L. (2008). *Harold E. Palmer: from learner-teacher to legend*. Hyderabad: Orient Longman.
- Tolpin, J. G. (1967). Review: A word count of spoken Russian: The Soviet usage by N. P. Vakar. *The Journal of Higher Education*, 38(1), 54–56.
- Turnbull, J. (Ed). (2010). *Oxford advanced learner's dictionary* (8th ed.). Oxford: Oxford University Press.
- Vakar, N. P. (1966). A word count of spoken Russian: The Soviet usage. Columbus: Ohio State University Press.
- van Geert, P. (2009). A comprehensive dynamic systems theory of language development. in K. De Bot and R.W. Schrauf (Eds.), *Language development over the life span*. (pp. 60–104). New York/London: Routledge.
- VOICE. (2011). The Vienna-Oxford International Corpus of English (version 1.1 XML). Retrieved from <http://voice.univie.ac.at>
- Voice of America word book: A list of words used in special English programs on radio, television and the Internet*. (2009). Washington, D.C.: VOA Special English. Retrieved from http://www.speakenglisheasilytoday.com/uploads/4/2/0/4/4204084/2009edition_wordbook.pdf

- Wang, M.-T. K., & Nation, I. S. P. (2004). Word meaning in academic English: Homography in the Academic Word List. *Applied Linguistics*, 25(3), 291–314.
- Warnes, G. R. (2012). Package “gtools”. Retrieved from <http://cran.r-project.org/web/packages/gtools/gtools.pdf>
- West, M. (1953). *A general service list of English words*. London: Longman, Green & Co.
- West, M. (1960). *Teaching English in difficult circumstances. Teaching English as a foreign language, with notes on the techniques of textbook construction*. London: Longmans, Green.
- West, M. P., Swenson, E., et al. (1934). *A Critical Examination of Basic English*. Toronto: The University of Toronto Press. (Reprinted in G. W. Terrence (Ed.), *C.K. Ogden and linguistics* (Vol. 4). London: Routledge / Thoemmes Press.
- Werner, D. H., Haupt, R. L. & Werner, P. L. (1999). Fractal antenna engineering: The theory and design of fractal antenna arrays. *IEEE Antennas and Propagation Magazine*, 41(5), 37.
- Wickham, H. (2010). Stringr: Modern, consistent string processing. *The R Journal*, 2(2), 38–40.
- Wilkins, D. A. (1972). *Linguistics and language teaching*. London: Edward Arnold.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.
- Xu, J. & Liang, M. (2012). *Crown: A 2009 Brown family corpus of present-day American English*. National Research Centre for Foreign Language Education, Beijing Foreign Studies University.

- Xu, J. & Liang, M. (2012). *CLOB: A 2009 Brown family corpus of present-day British English*. National Research Centre for Foreign Language Education, Beijing Foreign Studies University.
- Xue, G. and Nation, I.S.P. (1984) A university word list. *Language Learning and Communication*, 3(2), 215-229.
- Zimmerman, C. B. (1997). Historical trends in second language vocabulary instruction. In J. Coady, & T. Huckin (Eds.), *Second language vocabulary acquisition* (pp. 5–19). New York: Cambridge University Press.
- Zipf, G. K. (1935). *The Psycho-Biology of Language*. Cambridge, Mass.: M.I.T. Press.
- Zipf, G. K. (1949). *Human behavior and the principle of least effort: An introduction to human ecology*. New York: Hafner.

Appendix A.1

Corpus segments used to represent the *interactive* or *conversation* text type

A total of 7,050,000 tokens were collected to represent the *interactive* text type, of which 100,000 were compiled from text that remained after the files from each corpus was cut into 10,000-token blocks.

British National Corpus (4,190,000 tokens)

Spoken demographic sections AB, C1, C2, DE

International Corpus of English (1,890,000 tokens)

Files S1A 001 to 090 were used from each of the following ICE national corpora, except for ICE-East Africa. The spoken portion of the ICE corpora are designed to be comparable. However, due to differences between what was counted as a word by the corpus designers and the definition of an orthographic word used in this thesis, the total number of tokens from each of the corpora sometimes differ substantially.

ICE-Canada	180,000
ICE-Great Britain	180,000
ICE-Hong Kong	370,000
ICE-India*	190,000
ICE-Ireland	170,000
ICE-Jamaica	190,000
ICE-New Zealand	200,000
ICE-Philippines	190,000
ICE-Singapore	190,000
ICE-East Africa*	30,000

*Only the first part of the Kenyan spoken data is used since the Tanzanian and the second section of the Kenyan conversation files appear to be mostly formal interviews. Many of the ICE India files are also suspect since they seem to be from a walk-in interview setting rather than authentic casual conversation. The Indian files are included, however since no other data was available from the subcontinent.

Wellington Spoken Corpus (300,000 tokens)

Many files from the Wellington Spoken Corpus are also included in ICE-NZ. The conversational files from the WSC not already included in ICE-NZ have been added to the collection representing the *interactive* text type. These are DPC 002-004, 007-010, 012-014, 022-025, 028, 030, 032-033, 036-037, 039-043, 045, 049, 057, 059-064, 066-071, 073-075, 077-080, 082-083, 089-091, 093, 096-098, 111-112, 115, 118-121, 123-133, 136, 138-141, 151, 153-154, 156-164, 166, 169-172, 174-177, 179-182, 185, 187-193, 201, 206, 207, 212-213, 218, 224, 235, 237, 248, 265, 269, 277, 290, 301, 308, 310, 311, 318, 322-323, 336, 344-346.

London Lund Corpus (280,000 tokens)

All of the files containing spontaneous, face-to-face dialogue have been included. Those are S.1-S.4 and S.5.8-11.

Santa Barbara Corpus of Spoken American English (200,000 tokens)

Files sbc001-013, 015, 017, 019-021, 024-025, 027, 030-039, 041-051, 054-056, 058-060 are included since they contain casual face-to-face conversations.

Vienna-Oxford International Corpus of English (80,000 tokens)

Only the “Leisure” section appears to be authentic casual conversation. The VOICE files LEcon 8, 227, 228-229, 329, 351-353, 405, 417-418, 420, 545, 547, 548, 560, 562, 565-566, 573 and 575 are included in the collection for the interactive text type.

Appendix A.2

Corpus segments used to represent the *narrative writing* text type

A total of 18,420,000 tokens were collected to represent the *narrative writing* text type, of which 80,000 were compiled from text that remained after the files from each corpus was cut into 10,000-token blocks.

British National Corpus (16,430,000 tokens)

The BNC section labeled “fiction and drama” was selected to represent narrative writing.

International Corpus of English (470,000 tokens)

The category of “novels and short stories” were used from each of the following ICE national corpora. Though the parallel portions of the ICE corpora are designed to be comparable, they are not always exactly balanced since identical genre were not always available in each region. Due to differences between what was counted as a word by the corpus designers and the definition of an orthographic word used in this thesis, the total number of tokens from each of the corpora sometimes differ substantially.

ICE-Canada	40,000
ICE-Great Britain	40,000
ICE-Hong Kong	40,000
ICE-India	40,000
ICE-Ireland	40,000
ICE-Jamaica	40,000
ICE-New Zealand	70,000
ICE-Philippines	40,000
ICE-Singapore	40,000
ICE-East Africa	80,000

The Brown family of corpora (1,670,000 tokens)

Sections K (General fiction), L (Mystery and detective fiction), M (Science fiction), N (Western and adventure fiction), and P (Romantic fiction) were collected from each of the Brown family of corpora.

Brown University Corpus of American English	240,000
Freiburg-LOB Corpus of British English	240,000
Lancaster-Oslo/Bergen Corpus of British English	240,000
Freiburg-Brown Corpus of American English	240,000
Kolhapur Corpus of Indian English	230,000
Wellington Corpus of Written New Zealand English	250,000

Appendix A.3

Corpus segments used to represent the *general writing* text type

A total of 47,650,000 tokens were collected to represent the *general writing* text type, of which 170,000 were compiled from text that remained after the files from each corpus was cut into 10,000-token blocks.

British National Corpus (42,380,000 tokens)

Four text categories from the BNC were selected to represent the general writing text type:

Instructional	440,000
Persuasive	1,780,000
Reportage	9,250,000
Popular Writing	30,910,000

International Corpus of English (1,790,000 tokens)

Most of the following ICE national corpora contained around 80,000 tokens of the category of “popular writing,” 40,000 tokens of “reportage,” 20,000 tokens of “persuasive writing” and 20,000 tokens of “hobbies and skills.” There are some discrepancies due to what genres were available in each region, as well as differences between how words were counted by the corpus designers and the definition of an orthographic word used in this thesis.

ICE-Canada	160,000
ICE-Great Britain	190,000
ICE-Hong Kong	170,000
ICE-India	150,000
ICE-Ireland	160,000
ICE-Jamaica	160,000
ICE-New Zealand	230,000
ICE-Philippines	170,000
ICE-Singapore	160,000
ICE-East Africa	280,000

The Brown family of corpora (4,110,000 tokens)

From the Brown family of corpora, sections A (Press reportage), B (Press editorials), C (Press reviews), D (Religion), F (Popular lore), G (Biographies and essays), and R (Humor) have been included in this collection.

Brown University Corpus of American English	550,000
Freiburg-LOB Corpus of British English	550,000
Lancaster-Oslo/Bergen Corpus of British English	550,000
Freiburg-Brown Corpus of American English	550,000
Kolhapur Corpus of Indian English	530,000
Wellington Corpus of Written New Zealand English	540,000

Appendix A.4

Corpus segments used to represent the *academic writing* text type

A total of 17,120,000 tokens were collected to represent the *academic writing* text type, of which 130,000 were compiled from text that remained after the files from each corpus was cut into 10,000-token blocks.

British National Corpus (15,910,000 tokens)

Six text categories from the BNC were selected to represent the academic writing text type:

Academic humanities and arts	3,340,000
Medicine	1,400,000
Academic natural sciences	1,100,000
Academic politics and law	4,650,000
Social sciences	4,740,000
Academic technology and engineering	680,000

International Corpus of English (940,000 tokens)

Most of the following ICE national corpora contained around 40,000 tokens from the humanities and social sciences and 40,000 tokens from the natural sciences and technology. Due to differences between what was counted as a word by the corpus designers and the definition of an orthographic word used in this thesis, the total number of tokens from each of the corpora sometimes differs substantially.

ICE-Canada	80,000
ICE-Great Britain	90,000
ICE-Hong Kong	100,000
ICE-India	80,000
ICE-Ireland	70,000
ICE-Jamaica	80,000
ICE-New Zealand	130,000
ICE-Philippines	80,000
ICE-Singapore	80,000
ICE-East Africa	150,000

The Brown family of corpora (9,300,000 tokens)

Section J (Science / Academic prose) was collected from each of the Brown family of corpora.

Brown University Corpus of American English	160,000
Freiburg-LOB Corpus of British English	150,000
Lancaster-Oslo/Bergen Corpus of British English	160,000
Freiburg-Brown Corpus of American English	150,000
Kolhapur Corpus of Indian English	160,000
Wellington Corpus of Written New Zealand English	150,000

Appendix B.1

Sample R-scripts for “cleaning” corpus data

B.1.1 London Lund Corpus

The text files of the London Lund Corpus (Svartvik & Quirk, 1980) were annotated in 1975 before personal computers were available for linguistic research. Therefore, the markup and annotations were not devised with automated processing in mind, and they are sometimes a challenge to isolate and modify.

Each line of the text begins with a header which identifies the file and line in the text, as well as the speaker. The upper and lowercase letters at the end of the line header mark speakers who were aware (lowercase) or unaware (uppercase) that their conversation was being recorded. The original mark up for prosody and stress has been rendered in the text file with punctuation marks inserted in and around words. Some partial or non-lexical utterances, such as minimal responses ("mhm") are enclosed in square brackets. These comments occasionally break across lines, so that another line header intervenes in the middle of the comment. Other responses, such as laughter, as well as, comments about voice quality ("whispered speech"), environmental noises ("siren"), and comments on the recording procedure ("fault in tape"), are all enclosed in single parentheses. Where speech was incomprehensible, a comment on how many syllables could not be transcribed is inserted. Unfortunately, these comments are not consistent in format or content. They often break across lines, as well. In LLC_01, for example, the comment ((7 Greek sylls Greek . . . !s\yll#)) breaks across a line. This example had to be corrected manually. In file LLC_05, the utterances from an interviewer are also marked as a comment. These interviewer turns are often several lines long. The brackets marking off their turns were manually replaced with asterisks. Apparently, the researchers planned to exclude these turns as they did with contributions in the “surreptitious” files from speakers who were aware they were being recorded. All other comments have been reformatted using a series of substitutions and regular expressions. Below is a sample of one original text:

```

4 1 14 1950 1 1 a 11 and e^v\entually# /
4 1 14 1960 1 1 a 11 there`d be ^s\o {^m=uch#}# /
4 1 14 1970 1 1 a 11 you`d ^find !great d/\ifficulty# /
4 1 14 1980 1 1 a 11 in ^saying 'what it !w=as# /
4 1 15 1990 1 1 a 11 ( . laughs) d`you ^know what I m\ean# - - /

```

```

4 1 15 2000 1 1 b 11 +^how d`you m\ean##+ /
4 1 15 2010 1 2 a 11 +((3 to 4 sylls))+ . well ^sort of 'union acc\ount /
4 1 15 2010 1 1 a 11 {of ^what`s h\appened##}# /
4 1 15 2020 1 1 b 11 ( . clears throat) ^y\es but# . /
4 1 15 2030 1 2 b 11 there`s a ^tendency 'then for her to 'go and /
4 1 15 2030 1 1 b 11 re'peat !\every'thing# /
4 1 15 2040 2 1 b 21 [tu@s] - - - *an^n\oyance* of /
4 1 15 2050 1 1 a 11 *^wh\y d`you 'say (('that))#* /
4 1 15 2040 1 1(b 11 _every_body# - - - /
4 1 15 2060 1 1 a 11 [u] it was for "^that !r\eason# /
4 1 15 2070 1 1 a 11 that I ^thought of !d\oing 'this# /
4 1 15 2080 1 1 a 11 *be^cause if there was* - "!one 'issue# /
4 1 15 2090 2 1 a 21 she`d /

```

The R function described below modified these texts so that the line headers, comments on incomprehensible speech, voice quality, environmental noises and recording problems are all enclosed in angled brackets. This preserves the structure of the corpus text for future reference, while allowing the splitting function (Appendix B.3) to remove these from the data used to calculate Dice distance. Prosody and stress markup and the contents of square brackets have been deleted. This produces a text like the following:

```

< 4 1 14 1950 1 1 a 11 > and eventually
< 4 1 14 1960 1 1 a 11 > there'd be so much
< 4 1 14 1970 1 1 a 11 > you'd find great difficulty
< 4 1 14 1980 1 1 a 11 > in saying what it was
< 4 1 15 1990 1 1 a 11 > < laughs> d'you know what I mean
< 4 1 15 2000 1 1 b 11 > how d'you mean
< 4 1 15 2010 1 2 a 11 > <3 to 4 sylls unclear> well sort of union account
< 4 1 15 2010 1 1 a 11 > of what's happened
< 4 1 15 2020 1 1 b 11 > < clears throat> yes but
< 4 1 15 2030 1 2 b 11 > there's a tendency then for her to go and
< 4 1 15 2030 1 1 b 11 > repeat everything
< 4 1 15 2040 2 1 b 21 > annoyance of
< 4 1 15 2050 1 1 a 11 > why d'you say that
< 4 1 15 2040 1 1~b 11 > everybody
< 4 1 15 2060 1 1 a 11 > it was for that reason
< 4 1 15 2070 1 1 a 11 > that I thought of doing this
< 4 1 15 2080 1 1 a 11 > because if there was one issue
< 4 1 15 2090 2 1 a 21 > she'd

```

Several successive passes were made through each file to remove the various types of comments and mark up. Each line passes a temporary file to the next culminating in a final stripped file for which a dialog box is opened to save it to an already existing file. The target text file was created earlier. (A hash mark # tells R that what follows the hash mark is a comment and should not be processed as computer code.)

1	<code>clean.LLC <- function(file=choose.files()) { #Creates stripping function for London Lund Corpus.</code>
2	<code>LLCfile <- scan(file, what="char", sep="\n") #loads a single file</code>

The regular expression "`\\[.*\\]`" matches an open square bracket "`\\[`" followed by any number of characters "`.*`" followed by a closed square bracket "`\\]`". This eliminates comments like `[@ dh@:]` which do not break across lines.

3	<code>LLCfile.strip1<-gsub("\\[.*\\]", "", LLCfile, perl=T) #Strips certain comments and unfortunately also minimal responses, such as <i>mh</i>m.</code>
---	--

The regular expression "`^`" matches the beginning of each line in the file and places an open angled bracket in that position. Nothing is replaced since "`^`" simply marks the position in the line.

4	<code>LLCfile.strip2 <- gsub("^", "<", LLCfile.strip1, perl=T) #Places an open angled bracket at the beginning of the line header of each line.</code>
---	--

The positive lookbehind expression "`(?<=^{28})`" looks for the beginning of a line "`^`" followed by 28 characters of any type, including white space. Each line header fortunately contains exactly 27 characters plus the open angled bracket inserted in the previous step. This pattern of 28 characters is not replaced since the expression "`?<=`" instructs R to look back for, but not replace the pattern. A closed angled bracket is then inserted after this pattern.

5	<code>LLCfile.strip3 <- gsub("(?<=^{28})", "> ", LLCfile.strip2, perl=T) #Places a closed angled bracket at the end of the line header of each line.</code>
---	--

The LLC contains comments enclosed in single parentheses while double parentheses are part of the prosody and stress markup. These double parentheses have to be eliminated first before the comments can be placed in angled brackets.

6	<code>LLCfile.strip4 <- gsub("\\(\\(\\ \\)\\)", "", LLCfile.strip3, perl=T) #Strips double parentheses.</code>
---	--

The line headers also contain open parentheses in front of some of the speaker identifications apparently marking overlapping speech. These are replaced with "`~`" by

using positive lookbehind "`?<=`" again to search for a digit "`\\d`" followed by a literal open parenthesis "`\\(`". Since only the parenthesis is outside the lookbehind expression, only it is replaced.

7	<pre>LLCfile.strip5 <- gsub("(?<=\\d)\\(", "~", LLCfile.strip4, perl=T) #Replaces the open parenthesis in some of the line headers with an open curly bracket since they might interfere with the next replacement</pre>
---	--

The contents of the comments are extremely diverse and may break across lines in the corpus, so it is safer to simply replace the open and closed parentheses individually rather than try to write an expression to cover all the possible contents.

8	<pre>LLCfile.strip6 <- gsub("\\(", "<", LLCfile.strip5, perl=T) #Replaces open parenthesis for contextual comments with an open angled bracket.</pre>
9	<pre>LLCfile.strip7 <- gsub("\\)", ">", LLCfile.strip6, perl=T) #Replaces closed parenthesis for contextual comments with a closed angled bracket.</pre>

The long regular expression "`#|%|@|=|!|&|\\B-\\B|_|;|\\:|\\.|\\'|\\{\\|\\}\\|\\*|\\^|\\+|\\\\\\\\|\\\\/`" matches almost all the prosody and stress markers.

The expression "`\\B-\\B`" matches only hyphens that do not border a word character, since "`\\B`" indicates a non-word boundary. Simply deleting hyphens would collapse hyphenated compounds. Notice that four backslashes are required to match a single literal backslash.

10	<pre>LLCfile.strip8 <- gsub("# % @ = ! & \\B-\\B _ ; \\: \\. \\' \\{\\ \\}\\ * \\^ \\+ \\\\\\\\ \\\\/", "", LLCfile.strip7, perl=T) #Strips intonation markings.</pre>
----	---

Single and double quotation marks are interchangeable in R, as long as one is consistent within a single expression. The expressions "`a`" and '`a`' are equivalent. However, if one wishes to match a literal quotation mark, the quotation marks enclosing it must be of the other type. Otherwise, R has no way of distinguishing them from the end of the expression. Therefore "`""`" and '`''`' are impossible. For that reason, only the single quotation marks "`\\'`" were eliminated in line 10. A separate operation to replace the double quotation marks '`\\'`' is required in line 11.

11	<pre>LLCfile.strip9 <- gsub('\\\\', "", LLCfile.strip8, perl=T) #Strips double quotation marks.</pre>
----	--

Because the apostrophe or single quotation was used for markup, actual apostrophes were thankfully encoded as ```. In order to make the cleaned text more similar to other texts, these are replaced with normal apostrophes at this point.

12	<pre>LLCfile.strip10 <- gsub("`", "\\'", LLCfile.strip9, perl=T) #Replaces ` with a normal apostrophe.</pre>
----	---

As mentioned above, the annotation regarding indecipherable syllables is highly inconsistent. Since the expressions sometimes break across lines there are several possible patterns that will need to be matched. In order make the expression easier to understand, one can break the expression into its constituent parts:

- ① `(\\d{1,2} *to *){0,1}`
- ② `((?<=>) *to *\\d{1,2} *){0,1}`
- ③ `(\\d{1,2} *or *){0,1}`
- ④ `((?<=>) *or *\\d{1,2} *){0,1}`
- ⑤ `((?<= )c *\\d{1,2} ){0,1}`
- ⑥ `(more than \\d{1,2}|more |several |some |\\d{1,2} *){0,1}`
- ⑦ `(sylls{0,1}\\b)`

The first six parts of this regular expression end with the expression "`{0,1}`" since they look for zero or one occurrence of a pattern. These are all the possible ways (that have been observed) that the "unclear syllable" comment might begin. The first and third parts begin with the pattern of one or two digits "`\\d{1,2}`" followed by any number of spaces "`*`" followed by the word "`to`" or the word "`or`" respectively, continuing with any number of spaces "`*`".

However, the first digit in a comment, such as, "`3 to 4`" or "`3 or 4`" may be on the previous line, so patterns two and four search for the words "`to`" or "`or`" followed by any number of spaces, subsequently followed by one or two digits. The positive lookbehind expression "`((?<=>))`" at the beginning of parts two and four first checks that this pattern is preceded by a closed angled bracket. This would only happen if the pattern immediately follows a line header. Without the positive lookbehind, this expression would match the pattern "`to`" or "`or`" followed by digits inside those comments that had already been found by patterns one or three.

The fifth part begins with a positive lookbehind "`?<=`" that checks that there is a space before a letter "`c`" (short for circa) followed by spaces "`*`" and one or two digits "`\\d{1,2}`".

Part six contains five optional elements. The first is "`more than`" followed by one or more digits, as in "`more than 15`". Next the expression tries to match one of the words "`more`", "`several`" or "`some`". If no match is found, it looks for the final option, which is one or two digits.

The only mandatory part of this expression is that there must be an occurrence of "`syll`" or "`sylls`". The `{0,1}` indicates that the previous character "`s`" may occur once or not at all. Part seven ends with a word boundary "`\\b`" so that it does not match words like "`syllable`" or "`syllables`" which do occur several times in these files. This complicated regular expression has not captured the initial elements of such comments when they break across lines. These are dealt with in the next expression.

13	<pre> LLCfile.strip11 <- gsub("(\\d{1,2} *to *){0,1}((?<=>) *to *\\d{1,2} *){0,1}(\\d{1,2} *or *){0,1}((?<=>) *or *\\d{1,2} *){0,1}((?<=)c *\\d{1,2}){0,1}(more than \\d{1,2} more several some \\d{1,2} *){0,1}(sylls{0,1}\\b)", " <\\1\\2\\3\\4\\5\\6\\7 unclear> ", LLCfile.strip10, perl=T) #Moves a complete or the second half of fragmentary "unclear syllables" comments inside angled brackets. </pre>
----	--

The previous regular expression allowed for multiple combinations of the elements that might appear in the comment. In this case, only one of four possible options is possible with the two expressions below. These two expression are enclosed by an additional set of parentheses and are separated by a bar "`|`" which functions as the logical operator "`OR`".

① `((?<= )\\d{1,2} *(to|or) \\d{1,2} (?= ))`

② `((?<= )\\d{1,2} *(to|or) (?= ))`

Both patterns begin with a positive lookbehind expression "`(?<= )`". In each case, the expression first checks that a space precedes the first digit in the pattern. (Originally, a negative lookbehind "`?<!`" was considered for checking that the digit was not preceded by an open angled bracket having already been identified as a comment. However, this fails when the pattern begins with two digits, as in "`<10 to 12 sylls unclear`". The first digit "`1`"

is preceded by a bracket, but the "0" is not, so the expression matches the pattern "0 to 12" placing this it in brackets inside the already bracketed comment.) Pattern one checks for comment fragments like "3 to 4" and "3 or 4". The digit or digits should be followed by one or more spaces and either "to" or "or" and then subsequently followed by one or two more digits. A final positive lookahead expression checks that this pattern is followed by three spaces. This is to protect against retagging the unfragmented comments matched in the previous step – those in which the word "syll" follows the second digit(s) . The second pattern looks for fragments, such as "3 to" and "3 or", so the second set of digits "\\d{1,2}" is omitted.

14	<pre> LLCfile.strip12 <- gsub("(((?<=)\\d{1,2} *(to or) \\d{1,2}(?=)) ((?<=)\\d{1,2} *(to or)(?=)))", " <\\1> ", LLCfile.strip11, perl=T) #Moves the first part of "unclear syllables" comments inside angled brackets, except for free-standing digits. </pre>
----	--

One further possibility will be dealt with in the next substitution. A slightly different method will need to be used if the comment has broken across lines leaving only the digits on the first line. In this case, instead of looking to see what precedes the digit, we need to look forward to be sure nothing else follows this digit besides spaces. The expression "(\\d{1,2}) *(?=\$)" looks for one or two digits "(\\d{1,2})" followed by any number of spaces "*", which is then followed by a positive lookahead "(?=\$)" for the end of a line "\$". The spaces cannot be part of the lookahead expression, since lookarounds must be fixed length, whereas "*" is variable in length. This expression contains three elements: the digits, the spaces and the lookahead for the line end. In the previous replacements each matched element was replicated inside the angled brackets. This pattern has been treated separately since it would be inelegant to include the long string of spaces at the end of the line in the brackets. Instead this substitution only replicates the first element, the digits, inside brackets, "<\\1>".

15	<pre> LLCfile.strip13 <- gsub("(\\d{1,2}) *(?=\$)", " <\\1> ", LLCfile.strip12, perl=T) #Moves free-standing digits inside angled brackets that are left when "unclear syllables" comments break across lines. </pre>
----	--

At this point it is safe to eliminate the comments that were enclosed in square brackets, but split across lines. The expression "`\\[.*$`" matches a literal open square bracket "`\\[`" followed by any number of characters "`.*`" and ending at the end of the line "`$`".

16	<code>LLCfile.strip14 <- gsub("\\[.*\$", "", LLCfile.strip13, perl=T)</code> #Strips the beginnings of comments that broke across lines.
----	--

Deleting the second half of the broken comment needs to avoid removing the line header, so using positive lookbehind "`?<=`" R searches for, but does not replace instances of closed angled brackets "`>`". These would mark the end the of the line headers. It then replaces all subsequent characters "`.*`" until it reaches a closed square bracket "`]`". Closed angled brackets also mark the end of comments, but at this point, there should be no closed square brackets except where they have broken across lines.

17	<code>LLCfile.stripped <- gsub("(?<=).*\\]", "", LLCfile.strip14,</code> <code>perl=T)</code> #Strips the end of comments that broke across lines.
18	<code>write(LLCfile.stripped, file=choose.files())</code> #Opens dialog box for saving the stripped file.
19	<code>}</code> #Closes the function.

B.1.2 Wellington Corpus of Spoken New Zealand English

As mention in section 3.2.5, the indigenous vocabulary (mostly Maori) have not been deleted as they are an natural part of NZ English and are not usually complete code-switches. Indigenous words are only deleted in cases where three or more are contiguous within the tags <indig> or <indig=Maori> (line 9). The tags themselves are removed with the other tags.

1	<code>clean.WSC <- function(file=choose.files()) { #Creates stripping function for the Wellington Corpus of Spoken New Zealand English.</code>
2	<code>WSCfile <- scan(file, what="char", sep="\n") #Reads a file into memory.</code>
3	<code>WSCfile.pasted <- paste(WSCfile, collapse=" ") #Removes hard returns from the file.</code>
4	<code>WSCfile.by_turn <- unlist(strsplit(WSCfile.pasted, " (?U)<WSC\\#\\w{6}\\:\\d{4}\\:\\. *>")) #Splits the file at turn breaks and removes turn tags.</code>
5	<code>WSCfile.strip1 <- gsub("\\<\\&\\>.*\\<\\//\\&\\>", " ", WSCfile.by_turn, perl=T) #Strips environmental sound descriptions and their tags.</code>
6	<code>WSCfile.strip2 <- gsub("\\<\\O\\>.*\\<\\//\\O\\>", " ", WSCfile.strip1, perl=T) #Strips voice quality comments and their tags.</code>
7	<code>WSCfile.strip3 <- gsub("\\<unclear\\>.*\\<\\//unclear\\>", " ", WSCfile.strip2, perl=T) #Strips unclear utterances and their tags.</code>
8	<code>WSCfile.strip4 <- gsub("\\<(\\//)?(\\, \\. \\ \\[\\ \\{ \\ \\?+)\\>", "", WSCfile.strip3, perl=T) #Strips other interrupting tags.</code>
9	<code>WSCfile.strip5 <- gsub("\\<indig=Maori\\>\\W*\\w+\\W+\\w+\\W+\\w+.*\\<\\//indig=M aori\\>", " ", WSCfile.strip4, perl=T) #Strips code-switches of three or more words. This is different from the ICE-HK code. Tags in ICE-WSC are not followed by a space so the tags are replaced with a space to avoid fusing words together.</code>
10	<code>WSCfile.strip6 <- gsub("\\<foreign.*\\>.*\\<\\//foreign.*\\>", " ", WSCfile.strip5, perl=T) #Strips foreign utterances and their tags.</code>
11	<code>WSCfile.strip7 <- gsub("(?U)\\<.*\\>", " ", WSCfile.strip6, perl=T) #Strips all remaining tags.</code>

12	<code>WSCfile.stripped <- gsub("\\\\? \\\\.", " ", WSCfile.strip7, perl=T)</code> #Replaces various remaining punctuation with a space.
13	<code>write(WSCfile.stripped, file=choose.files())</code> #Opens dialog box for saving the stripped file.
14	<code>}</code> #Closes the function.

B.1.3 International Corpus of English (written)

This function script is designed to remove tags and annotations from the written sections of ICE-Canada, ICE-India, ICE-Jamaica, ICE-Philippines, ICE-New Zealand, ICE-Singapore, and ICE-Hong Kong. These corpora sometimes contain HTML codes for non-ASCII characters that need to be replaced with the appropriate Unicode character. These are replaced, and the file passed to the next line in bucket-brigade fashion. This is admittedly inelegant. A substitution table of some sort would be better.

1	<code>clean.ICEW_Gen <- function(file=choose.files()) { #Creates stripping function for written sections of many of the ICE corpora.</code>
2	<code>ICEWfile <- scan(file, what="char", sep="\n") #Loads a single file.</code>
3	<code>ICEWfile.pasted <- paste(ICEWfile, collapse=" ") #Removes hard returns from file, since tags break across lines in most of the ICE corpora.</code>
4	<code>ICEWfile.by_line <- unlist(strsplit(ICEWfile.pasted, "(?U)<ICE-\\w{2,3}\\:\\w{3}\\-\\d*\\#.*>")) #Splits the file at true line breaks and removes line tags. The length of the suffix is variable for each part of ICE, e.g. ICE-JA is two, not three.</code>
5	<code>ICEWfile.no_breaks <- gsub("(\\<1\\> \\<1\\^\\>)", "", ICEWfile.by_line, perl=T) #Strips line breaks that have spaces on either side or the NZ variant with a carat after the "I".</code>
6	<code>ICEWfile.strip1 <- gsub("\\<\\&\\>.*\\<\\/\\&\\>", " ", ICEWfile.no_breaks, perl=T) #Strips formatting descriptions and their tags.</code>
7	<code>ICEWfile.strip2 <- gsub("\\<\\O\\>.*\\<\\/O\\>", " ", ICEWfile.strip1, perl=T) #Strips comments on photos, diagrams, maps, etc. and their tags.</code>
8	<code>ICEWfile.strip3 <- gsub("\\<*\\>.*\\<\\/*\\>", " ", ICEWfile.strip2, perl=T) #Strips alternate formatting comments and their tags.</code>
9	<code>ICEWfile.strip4 <- gsub("\\<\\+\\>.*\\<\\/\\+\\>", " ", ICEWfile.strip3, perl=T) #Strips words that have been inserted to correct misspelled or misused words along with their tag.</code>
10	<code>ICEWfile.strip5 <- gsub("(?U)\\<.*\\>", " ", ICEWfile.strip4, perl=T) #Strips all remaining tags.</code>

11	ICEWfile.strip6 <- gsub("á", "á", ICEWfile.strip5, perl=T) #Replaces the tag for "a" with an acute accent.
12	ICEWfile.strip7 <- gsub("é", "é", ICEWfile.strip6, perl=T) #Replaces the tag for "e" with an acute accent.
13	ICEWfile.strip8 <- gsub("í", "í", ICEWfile.strip7, perl=T) #Replaces the tag for "i" with an acute accent.
14	ICEWfile.strip9 <- gsub("ó", "ó", ICEWfile.strip8, perl=T) #Replaces the tag for "o" with an acute accent.
15	ICEWfile.strip10 <- gsub("ú", "ú", ICEWfile.strip9, perl=T) #Replaces the tag for "u" with an acute accent.
16	ICEWfile.strip11 <- gsub("âumflex;", "â", ICEWfile.strip10, perl=T) #Replaces the tag for "a" with an circumflex.
17	ICEWfile.strip12 <- gsub("êumflex;", "ê", ICEWfile.strip11, perl=T) #Replaces the tag for "e" with an circumflex.
18	ICEWfile.strip13 <- gsub("îumflex;", "î", ICEWfile.strip12, perl=T) #Replaces the tag for "i" with an circumflex.
19	ICEWfile.strip14 <- gsub("ôumflex;", "ô", ICEWfile.strip13, perl=T) #Replaces the tag for "o" with an circumflex.
20	ICEWfile.strip15 <- gsub("ûumflex;", "û", ICEWfile.strip14, perl=T) #Replaces the tag for "u" with an circumflex.
21	ICEWfile.strip16 <- gsub("(&A-ACUTE; Á)", "Á", ICEWfile.strip15, perl=T) #Replaces the tag for "A" with an acute accent.
22	ICEWfile.strip17 <- gsub("(&E-ACUTE; É)", "É", ICEWfile.strip16, perl=T) #Replaces the tag for "E" with an acute accent.
23	ICEWfile.strip18 <- gsub("(&I-ACUTE; Í)", "Í", ICEWfile.strip17, perl=T) #Replaces the tag for "I" with an acute accent.
24	ICEWfile.strip19 <- gsub("(&O-ACUTE; Ó)", "Ó", ICEWfile.strip18, perl=T) #Replaces the tag for "O" with an acute accent.
25	ICEWfile.strip20 <- gsub("(&U-ACUTE; Ú)", "Ú", ICEWfile.strip19, perl=T) #Replaces the tag for "U" with an acute accent.
26	ICEWfile.strip21 <- gsub("(ä äaut;)", "ä", ICEWfile.strip20, perl=T) #Replaces the tag for "a" with an umlaut.
27	ICEWfile.strip22 <- gsub("(ë ëaut;)", "ë", ICEWfile.strip21, perl=T) #Replaces the tag for "e" with an umlaut.

28	ICEWfile.strip23 <- gsub("ï ïaut;)", "i ", ICEWfile.strip22, perl=T) #Replaces the tag for "i" with an umlaut.
29	ICEWfile.strip24 <- gsub("ö öaut;)", "ö", ICEWfile.strip23, perl=T) #Replaces the tag for "o" with an umlaut.
30	ICEWfile.strip25 <- gsub("ü üaut;)", "ü" , ICEWfile.strip24, perl=T) #Replaces the tag for "u" with an umlaut.
31	ICEWfile.strip26 <- gsub("à", "à", ICEWfile.strip25, perl=T) #Replaces the tag for "a" with a grave accent.
32	ICEWfile.strip27 <- gsub("è", "è", ICEWfile.strip26, perl=T) #Replaces the tag for "e" with a grave accent.
33	ICEWfile.strip28 <- gsub("ì", "ì ", ICEWfile.strip27, perl=T) #Replaces the tag for "i" with a grave accent.
34	ICEWfile.strip29 <- gsub("ò", "ò", ICEWfile.strip28, perl=T) #Replaces the tag for "o" with a grave accent.
35	ICEWfile.strip30 <- gsub("ù", "ù" , ICEWfile.strip29, perl=T) #Replaces the tag for "u" with a grave accent.
36	ICEWfile.strip31 <- gsub("æature;", "æ", ICEWfile.strip30, perl=T) #Replaces the tag for " an "a-e" ligature.
37	ICEWfile.strip32 <- gsub("&oeligature;", "œ", ICEWfile.strip31, perl=T) #Replaces the tag for an "o-e" ligature.
38	ICEWfile.strip33 <- gsub("çle;", "ç", ICEWfile.strip32, perl=T) #Replaces the tag for a "c" cedille.
39	ICEWfile.strip34 <- gsub("ñ &ntidle;)", "ñ", ICEWfile.strip33, perl=T) #Replaces the tag for an "n" with a tilde (including a common metathesis).
40	ICEWfile.strip35 <- gsub("Ω", "Ω", ICEWfile.strip34, perl=T) #Replaces the tag for an Omega.

41	<pre>ICEWfile.stripped <- gsub("‘ ’ “ ” &obrack; &cbrack; &lsqbrack; &rsqbrack; ; & &ampersand; ° &degree; &degree-sign; ⁁ &percent; &scol; > < &plus-or-minus; ‐ &curved-dash; &very- long-dash; &long-dash; &dotted-line; &dotted- line; ˙ • &arrowhead; &right-arrow; □ &black-square; ⋄ &club; &heart; &spade; $ &peso; &centavo; £ &pound-sign; \\{\\#\\})"," ", ICEWfile.strip35, perl=T) #Replaces all non-alphabetic HTML codes seen in the text and one formatting tag found in ICE-India with a space.</pre>
42	<pre>write(ICEWfile.stripped, file=choose.files()) #Opens dialog box for saving the stripped file. An empty text file should be prepared in advance.</pre>
43	<pre>} #Closes the function.</pre>

Appendix B.2

R-script for combining corpus files

This script creates a function `combine.files()` that will save all the selected files into a text file called `combined.file.txt`, which should be renamed appropriately after the function has run.

1	<code>combine.files <- function(file=choose.files()) {</code> # The first line defines the function and opens a dialogue box in which to select the files to be combined.
2	<code> n <- length(file)</code> # The variable n is set at the number of files that were selected.
3	<code> for (i in 1:n) {</code> # The for loop iterates through from the first to the last of the files selected.
4	<code> textfile <- scan(file[i], what="char", sep="\n", quote="", comment.char="")</code> # Each file is first scanned into memory and stored in the temporary string textfile.
5	<code> cat((paste(textfile, sep=" ")), file="combined.file.txt", sep="\n", append=T)</code> # It is then pasted into the file <code>combined.file.txt</code> appending each new file at the end of the file. By appending each new file to the text file, only one file needs to be held in memory at a time.
6	<code> }</code>
7	<code>}</code> # At the end of the loop the file is complete and the function is closed.

Appendix B.3

R-script for splitting corpus data into equal-sized blocks

This script creates a function `file.split()` that will divide the selected text file into equal-sized numbered text files each containing 10,000 tokens.

1	<code>file.split <- function(file=choose.files(), sec.size=10000)</code> <code>{</code> <code>#Defines the function and sets the default section size to 10,000 tokens. A file</code> <code>selection window is also opened by default.</code>
2	<code> textfile <- scan(file, what="char", sep="\n", quote="",</code> <code> comment.char="")</code> <code>#The function scans the selected file into memory.</code>
3	<code> stripped.textfile <- gsub("<.*?>", "", textfile, perl=T)</code> <code>#</code> <code> Tags are removed from the file.</code>
4	<code> text.no.free.digits <- gsub("\\b\\d+\\b", "",</code> <code> stripped.textfile, perl=T)</code> <code>#Free standing digits are removed, but not</code> <code> those that are in strings with alphabetic characters.</code>
5	<code> word.series <- unlist(strsplit(text.no.free.digits, "\\W+"))</code> <code> #A vector of word strings is created. The function strsplit outputs a list. A list is</code> <code> a collection of vectors, in this case, one vector per line in the corpus. The function</code> <code> unlist smoothes the list into a single vector that the words can be counted as one long</code> <code> series. See note on strsplit below*</code>
6	<code> word.series <- word.series[nchar(word.series)>0]</code> <code>#Any</code> <code> remaining empty strings are removed.</code>
7	<code> n <- length(word.series)</code> <code>#The function length counts the number of</code> <code> elements contained in a vector. In the vector word.series, each element is a one-</code> <code> word string.</code>
8	<code> sec.number <- 1</code> <code>#Initializes the sec.number variable that will be used for</code> <code> numbering the file sections.</code>
9	<code> sec <- vector()</code> <code>#Creates a vector to contain the words for each section</code> <code> while it is being created.</code>
10	<code> i <- 1</code> <code>#Creates a variable that will act as a word counter and initialize it as 1.</code>
11	<code> while (i < n) {</code> <code>#Initiates a while loop that will continue processing the file</code> <code> to be split until it reaches the last word of the file, that is until the variable i (the</code> <code> counter) is no longer less than the variable n (the number of words in the file).</code> <code> }</code>

12	<pre>if (i > sec.number*sec.size) {</pre> <p>#Until the variable <code>i</code> equals <code>n</code>, the function needs to count off sections of the desired size. If the value of the variable <code>i</code> is greater than the section number multiplied by the desired section size, that section is full and should be saved. The section number is set to 1 at the beginning, so initially, the function is counting word by word until it goes over $1 * 10,000$.</p>
14	<pre>sec.name <- paste("sec", sec.number, ".txt", sep="_")</pre> <p>#The file name for that section is created by pasting together the letters "sec", the number of that section and a ".txt" suffix so that the computer's operating systems recognizes it as a text file. An underline separates the elements of the file name.</p>
14	<pre>cat(sec, file=sec.name[1], append=T)</pre> <p>#The words stored in the vector <code>sec</code> are saved in a file that has the name of the first string contained in the vector <code>sec.name</code>. The vector <code>sec</code> only ever contains one file name since the name is overwritten at each iteration.</p>
15	<pre>cat(paste(sec.name, length(sec), sep=" = "), file="report.txt", sep="\n", append=T)</pre> <p>#A report.txt file is created and saved in the default directory. The name of each section and the number of words it contains is added to a new line ("\n") in the report after each section file is created. The argument <code>append</code> needs to be "TRUE", otherwise each new report would overwrite the contents of the file.</p>
16	<pre>sec.number <- sec.number+1</pre> <p>#The section number is iterated, so that the script can begin filling the next section. After the first section has been filled, the section number will be 2 and so the function will count <code>i</code> from 10,001 until it is greater than $2 * 10,000$.</p>
17	<pre>rm(sec)</pre> <p>#The vector <code>sec</code> is removed and a new one is created in line 18 so that it is ready for the words from the next section. The variable <code>i</code> continues until it reaches the number of words in the file, but the <code>sec</code> variable needs to be refreshed after each section is filled – a little bit like the water buckets in the sculpture on the Cuba mall.</p>
18	<pre>sec <- vector()</pre>
19	<pre>} else {</pre> <p>#This is the end of the if loop. The function checks again to see, if the section vector <code>sec</code> is full, otherwise the function goes to the <code>else</code> portion of the loop again.</p>

20	<code>sec <- append(sec, word.series[i])</code> #As seen in lines 6 and 14, the brackets [] are like sliding doors that allow the script to access the items stored in a vector. In line 6, the function retrieved items based on a characteristic (those strings that contained more than 0 characters). One can also select items in a vector based on their position in the vector, as seen in line 14. The vector <code>sec.name</code> only ever has one element, but the vector <code>word.series</code> likely has many thousand. The variable <code>i</code> was initially set to 1. So, during the first pass through the loop, <code>word.series[i]</code> will select the word in position 1 in the vector <code>word.series</code> and this word will be appended to the already existing (but so far empty) <code>sec</code> vector.
21	<code>i <- i+1</code> #The function iterates the variable <code>i</code> to 2, so that it can select the second word in the vector <code>word.series</code> on its next pass through the loop.
22	<code>}</code> #The if-else loop is closed, and the script goes back to check the <code>while</code> loop again.
23	<code>}</code> #Once <code>i</code> reaches the end of the file, i.e. it equals <code>n</code> , the <code>while</code> loop is also closed and the function proceeds to the last section of the script.
24	<code>sec.name <- paste("sec", sec.number, ".txt", sep="_")</code> #The script creates a file name for the last file section it has been building even though it will likely be smaller than desired size.
25	<code>sec.name <- cat("sec_", sec.number, sep="")</code>
26	<code>file.create(sec.name)</code>
27	<code>cat(sec, file=sec.name[1], append=T)</code> #The final section is saved.
28	<code>paste(sec.name, length(sec), sep=" = ")</code> #Prepares a file entry for the report.
29	<code>cat((paste(sec.name, length(sec), sep=" = ")), file="report.txt", sep="\n", append=T)</code> #An entry is recorded in the report.
30	<code>rm(i, n, sec, sec.name, sec.size, sec.number, textfile, stripped.textfile, text.no.free.digits, word.series)</code> #The temporary files are removed so that they do not interfere with subsequent processes.
31	<code>}</code> #End of function script.

*The function `strsplit` splits the text into individual strings based on the pattern indicated. The pattern "`\\W+`" tells `strsplit` to split the text at each occurrence of one or more non-word characters. These whitespace characters and punctuation (shown in gray below) are “eaten” in the process. What is left is a series of words, or to be precise, a series of alphanumeric strings. (It is important to use the regular expression "`\\W`" and not "`\\w`" that I once used by mistake. Double backslash, lower case “w” indicates all alphanumeric characters, in which case only the non-word characters will be left in the file!) The plus + symbol in the expression indicates one or more instances of the preceding expression. For example, in this line from the Brown corpus one sees 10 words since the possessive 's is counted as a separate word:

Without the plus + symbol, R would create a string from the space before a word, e.g. produced to the space after the word. That creates the string "produced" which is fine. But then, R would also create a between the space after the word and the next non-word character, which is the quotation mark. Of course there's nothing between the space and the quotation mark. Nonetheless, R would dutifully create this empty string, unless the regular expression stipulates that there may be more than one non-word character between the strings.

The function `strsplit` outputs a list made up of one vector for each line of the corpus. Each vector is the sequence of words in that line. Double brackets indicate the number or name of a list element. The first vector contains nine individual words or strings.

```
[[1]]
[1] "the"          "fulton"       "county"       "grand"
[5] "jury"         "said"         "friday"       "an"
[9] "investigation"
```

```
[[2]]
[1] "of"          "atlanta"     "s"           "recent"      "primary"
"election"
[7] "produced"    "no"          "evidence"    "that"
```

Since the function `strsplit` outputs a list containing many vectors, this output needs to be streamlined back into a vector by the function `unlist` so that subsequent functions can work on it as a single series.

```
[1] "the"      "fulton" "county" "grand"  "jury"    "said"
```

Appendix C

R-scripts for calculating Dice distance

Rather than showing each step in the evolution of the Dice distance scripts, the final automated script is presented here since it combines all the previous elements. The automated version was written to select files for each sample without regard to which were selected for other samples, i.e. they were selected with replacement. The final Dice distance study demonstrated that sampling with replacement inevitably reduces the distance between samples. Therefore, an alternative version of the sampling section is given in Appendix C.2. If lines 24-28 are replaced with the lines from B.2, the script will randomly select files for each sample without reusing any of them for subsequent samples. Of course, there need to be enough files in the source directory to allow for that number of independent samples.

C.1 Automated script for calculating Dice distance with replacement.

1	<code>time.started <- format(Sys.time(), "%a %b %d %X %Y")</code> #Records the start time of the script run.
2	<code>#</code>
3	<code>#SETTINGS AND LIBRARIES</code>
4	<code>#</code>
5	<code>library(arules)</code> #Opens the arules library (Hahsler, Grün, Hornik, & Buchta, 2008) that contains the function for calculating Dice distance.
6	<code>options("scipen"=10)</code> #Prevents scientific notation of large number as exponents. E-notation can interfere with loops.
7	<code>text_type <- "Gen_Writ"</code> #WRITE NAME OF TEXT CATEGORY HERE. This will be the first part of all file names produced by the script.
8	<code>#</code>
9	<code>#FILE LIST</code>
10	<code>#</code>
11	<code>text.files <- list.files(path="C:/Users/Joseph/Documents/All_Gen_Writ_split_files", pattern="*.txt")</code> #Captures the names of the files in the source directory.
12	<code>list.length <- length(text.files)</code> #Records the number of files in the directory.
13	<code>file.list <- c(1:list.length)</code> #Creates a vector with the same length as the number of files in the directory.

14	<code>names(file.list) <- list.files(path="C:/Users/Joseph/Documents/All_Gen_Writ_split_files", pattern="*.txt")</code> #Assigns the names of the files in source directory to one of the numbers on the list.
15	<code>#</code>
16	<code>#SETTINGS FOR MULTI-LEVELED ANALYSIS</code>
17	<code>#</code>
18	<code>p <- 1</code> #Initiates the variable for multi-level automated processing.
19	<code>for (p in 1:25) {</code> #Loop for iterating through all the desired sample sizes, e.g. from 1 to 25 million.
20	<code> x <- 1</code> #Resets the variable for the random sample number at 1 for each level.
21	<code>#</code>
22	<code>#RANDOM SAMPLE with replacement</code>
23	<code>#</code>
24	<code> for (x in 1:10) {</code> #Creates 10 random samples
25	<code> q <- 100*p</code> #Number of files to be randomly selected at base level -- usually 100 (= 1 million tokens) times the level being created, e.g. 3 x 1 million = 3 million.
26	<code> s <- sample(list.length, q, replace = FALSE)</code> #Creates a vector of the reference numbers of the files to be selected. Replace is set to FALSE which means a single file will not be selected more than once in the same sample. In section 3.3 this was changed to <code>replace = TRUE</code> in order to bootstrap the conversation to a total sample size greater than the quantity of data available.
27	<code> s <- names(file.list[s])</code> #Assigns the file names to their reference numbers.
28	<code> cat((paste("Random list", p, x, s, sep="\t")), file="Dice_report.r", sep="\n", append=T)</code> #Creates a report of which files were assigned to each sample.
29	<code>#</code>
30	<code>#FREQUENCY LIST</code>
31	<code>#</code>
32	<code> combined.list = NULL</code> #Creates an empty vector for the combined list of words.

33	<code>length(combined.list) <-60000000</code> #Pre-allocates memory space for this vector, otherwise R needs to find a new space in memory each time the vector increases in size.
34	<code>for (i in 1:q) {</code>
35	<code>text.file<-scan(s[[i]], what="char", sep="\n", quote="", comment.char="")</code> #Inputs one of the selected text files.
36	<code>text.file<-tolower(text.file)</code> #Changes all alphabetic characters to lower case.
37	<code>text.file<-gsub("<.*?>", "", text.file, perl=T)</code> #Removes tags in angled brackets.
38	<code>word.list<-strsplit(text.file, "\\W+")</code> #Extracts words from the file.
39	<code>word.vector<-unlist(word.list)</code> #Changes the output from the function strsplit back into a vector.
40	<code>word.vector<-word.vector[nchar(word.vector)>0]</code> #Removes any empty strings.
41	<code>freq.list<-table(word.vector)</code> #Creates a table of named integers (word types and their frequencies).
42	<code>assign(paste("freq.list", i, sep="."), freq.list)</code> #Creates a unique file name for each iteration of this loop.
43	<code>combined.list<-c(combined.list, freq.list)</code> #Adds each frequency list to combined list. They are simply tacked onto each other like a series of train carriages.
44	<code>}</code>
45	<code>combined.freq <- as.table(tapply(combined.list, names(combined.list), sum))</code> #Transforms the series of individual frequency lists into a unified frequency list.
46	<code>combined.sorted.freq.list <- sort(combined.freq, decreasing=T)</code> #Sorts the frequency list in descending order of frequency.
47	<code>combined.table <- paste(names(combined.sorted.freq.list), combined.sorted.freq.list, sep="\t")</code> #Creates a table with columns for words and total frequency.
48	<code>L <- floor(length(combined.table)/1000)</code> #Records the length of the current samples vocabulary rounded down to the last full 1000.

49	<code>assign(paste("L", x, sep="."), L) #Assigns a unique vector name to each limit value for each of the samples.</code>
50	<code>Leng <- length(combined.table) #Records the length of the current samples vocabulary.</code>
51	<code>assign(paste("Leng", x, sep="."), Leng) #Assigns a unique vector name to each limit value.</code>
52	<code>cat("Word\tTotal", combined.table, file=paste(text_type, p, "M", x, "freq_list.csv", sep="_"), sep="\n") #Saves a frequency list for the current sample.</code>
53	<code>#</code>
54	<code># 100-LEVEL COMPARISON LISTS</code>
55	<code>#This loop saves 100 text files of the most frequent word types in the current sample from the top 100 to top 10,000 words in increments of 100.</code>
56	<code>#</code>
57	<code>for (j in 1:100) {</code>
58	<code> k <- j*100</code>
59	<code> cat((paste(names(combined.sorted.freq.list[1:k]), sep="")), file=paste(text_type, p, k, x, "h.txt", sep="_"), sep=" ")</code>
60	<code> j <- j+1</code>
61	<code>}</code>
62	<code>}</code>
63	<code>#</code>
64	<code># 1000-LEVEL COMPARISON LISTS</code>
65	<code>#This loop saves text files of the most frequent word types in the current sample from the top 1000 words to the length of the sample with the smallest vocabulary in increments of 1000.</code>
66	<code>#</code>
67	<code>k_range <- c(Leng.1, Leng.2, Leng.3, Leng.4, Leng.5, Leng.6, Leng.7, Leng.8, Leng.9, Leng.10) #Creates a vector containing the vocabulary size of each of the samples.</code>
68	<code>cat((paste(text_type, p, "M", "sample length", k_range, sep="\t")), file="sample_length_report.r", sep="\n", append=T) #Records a report on the vocabulary sizes for the ten samples.</code>

69	<code>k_limit <- min(L.1, L.2, L.3, L.4, L.5, L.6, L.7, L.8, L.9, L.10) #Determines the minimum vocabulary size limit among the 10 samples. Comparison files will not be made beyond this limit.</code>
70	<code>for (K in 1:10) {</code>
71	<code> sample.freq.list <- read.table(file=paste(text_type, p, "M", K, "freq_list.csv", sep="_"), stringsAsFactors=FALSE, comment.char="", header=T, sep="\t") #Scan in frequency file n.b.: without stringsAsFactors=FALSE, R will not save the actual words. Instead it will save files containing the numbered positions of each of the words if they were sorted alphabetically!</code>
72	<code> for (n in 1:k_limit) { #Creates the 1000-level comparison files from 1,000 to the limit for this level established in line 69.</code>
73	<code> m <- n*1000 #Records how many thousand words are compared in this list. This will be part of the file name recorded in line 74.</code>
74	<code> cat(sample.freq.list[[1]][1:m], file=paste(text_type, p, m, K, "k.txt", sep="_"), sep=" ") #Extract the names for the appropriate comparison size.</code>
75	<code> n <- n+1</code>
76	<code> }</code>
77	<code>}</code>
78	<code>rm(list=ls(pattern=("freq.list\\.\\d* combined.* selected.files text.file word.list word.vector j L\\.//d*"))) #Removes vectors for creating a frequency list.</code>
79	<code>#</code>
80	<code>#DICE BINARY 100 LEVEL</code>
81	<code>#</code>
82	<code> for (X in 1:100) { #Iterates through all 100 of the 100-levels of the Dice distance comparison.</code>
83	<code> y <- X*100 #Calculates the level for each file name.</code>
84	<code> z <- paste(text_type, p, y, "\\d{1,2}", "h.txt", sep="_") #Adds the level to the rest of the file name search pattern.</code>
85	<code> selected.files <- list.files(path=getwd(), pattern=z) #Selects all the files from the working directory that match pattern z, i.e. all the files for that comparison level.</code>
86	<code> combined.list = NULL #Creates an empty initial list.</code>

87	<code>length(combined.list) <- 6000000</code> #Preallocates memory space for this vector.
88	<code>for (I in 1:length(selected.files)) {</code>
89	<code>text.file <- scan(selected.files[[I]], what="char", sep="\n", quote="", comment.char="")</code> #Inputs the text file containing the appropriate list of most frequent words from a sample.
90	<code>word.list <- strsplit(text.file, "\\W+")</code> #Extracts words from the file.
91	<code>word.vector <- unlist(word.list)</code> #Changes the output from the function strsplit back into a vector.
92	<code>word.vector <- word.vector[nchar(word.vector)>0]</code> #Removes any empty strings.
93	<code>freq.list <- table(word.vector)</code> #Creates a table of named integers (word types and their frequencies).
94	<code>assign(paste("freq.list", I, sep="."), freq.list)</code> #Creates a unique file name for each iteration of this loop.
95	<code>combined.list <- c(combined.list, freq.list)</code> #Adds each frequency list to combined list.
96	<code>}</code>
97	<code>combined.freq <- as.table(tapply(combined.list, names(combined.list), sum))</code> #Transforms the series of individual frequency lists into a unified frequency list.
98	<code>combined.sorted.freq.list <- sort(combined.freq, decreasing=T)</code> #Sorts the frequency list in descending order of frequency.
99	<code>all.freq.lists <- ls(pattern="freq.list\\.\\d{1,2}")</code> #Lists all frequency files in memory with 1 or 2 digits as a suffix in their names.
100	<code>for (J in 1:length(all.freq.lists)) {</code> #Loop sorts each individual frequency list into alphabetical order according to the combined vocabulary of all the samples.
101	<code>freq.list.J <- get(all.freq.lists[[J]])</code>
102	<code>J.in.combined <- freq.list.J[freq.list.J=names(combined.sorted.freq.list)]</code> #Pastes the names of the frequency vector, i.e. the words, alongside the frequency integers in a matrix.

103	<code>J.in.combined[is.na(J.in.combined)] <- 0</code> #Changes "NA" to a numeric 0.
104	<code>assign(paste("combined", J, sep="."), J.in.combined)</code> #Writes the current frequency list to a file.
105	<code>}</code>
106	<code>combined.table <- paste(names(combined.sorted.freq.list), combined.sorted.freq.list, combined.1, combined.2, combined.3, combined.4, combined.5, combined.6, combined.7, combined.8, combined.9, combined.10, sep="\t")</code> #Creates a table with columns for the total and all of the component lists.
107	<code>cat("Word\tTotal\t1\t2\t3\t4\t5\t6\t7\t8\t9\t10", combined.table, file=paste(text_type, p, "words_h", y, ".csv", sep="_"), sep="\n")</code> #Saves binary matrices with words for reference.
108	<code>combined.matrix <- paste(combined.1, combined.2, combined.3, combined.4, combined.5, combined.6, combined.7, combined.8, combined.9, combined.10, sep="\t")</code> #Creates a table with columns for only the component lists in the combined order. The word types are not recorded -- only whether their frequency is 0 or 1, i.e. whether they were present in the sample or not. These will be used to calculate the Dice distance between the samples.
109	<code>cat("1\t2\t3\t4\t5\t6\t7\t8\t9\t10", combined.matrix, file=paste(text_type, p, y, "h.csv", sep="_"), sep="\n")</code> #Saves the binary matrices.
110	<code>rm(list=ls(pattern=("freq.list.* combined.* selected.files text.file word.list word.vector I J")))</code> #Removes the vectors for a 10-way Dice comparison.
111	<code>X <- X+1</code>
112	<code>}</code>
113	<code>rm(X, y, z)</code>
114	<code>#</code>
115	<code>#DICE BINARY 1000 LEVEL</code>
116	<code>#</code>

117	<code>for (X in 1:k_limit) { #Iterates through all 1000-levels of the Dice distance comparison.</code>
118	<code>y <- X*1000 #Calculates the value for each 1000-level to be used for the file names.</code>
119	<code>z <- paste(text_type, p, y, "\\d{1,2}", "k.txt", sep="_") #Adds the level value to the rest of the file name search pattern.</code>
120	<code>selected.files <- list.files(path=getwd(), pattern=z) #Selects all the files from the working directory that match pattern z, i.e. all the files for that level.</code>
121	<code>combined.list = NULL #Creates an empty initial list.</code>
122	<code>length(combined.list) <-6000000 #Preallocates memory space for this vector.</code>
123	<code>for (I in 1:length(selected.files)) {</code>
124	<code>text.file <- scan(selected.files[[I]], what="char", sep="\n", quote="", comment.char="") #Inputs one of the text files for this comparison.</code>
125	<code>word.list <- strsplit(text.file, "\\W+") #Extracts words from the file.</code>
126	<code>word.vector <- unlist(word.list) #Changes the output from the function strsplit back into a vector.</code>
127	<code>word.vector <- word.vector[nchar(word.vector)>0] #Removes any empty strings.</code>
128	<code>freq.list<-table(word.vector) #Creates a table of named integers (word types and their frequencies).</code>
129	<code>assign(paste("freq.list", I, sep="."),freq.list) #Creates a unique file name for each iteration of this loop.</code>
130	<code>combined.list <- c(combined.list, freq.list) #Adds each frequency list to combined list.</code>
131	<code>}</code>
132	<code>combined.freq <- as.table(tapply(combined.list, names(combined.list), sum)) #Transforms the series of individual frequency lists into a unified frequency list.</code>

133	<code>combined.sorted.freq.list <- sort(combined.freq, decreasing=T) #Sorts the frequency list in descending order of frequency.</code>
134	<code>all.freq.lists <- ls(pattern="freq.list\\.\\d{1,2}") #Lists all frequency files in memory with 1 or 2 digits as a suffix.</code>
135	<code>for (J in 1:length(all.freq.lists)) { #Loop sorts each individual frequency list into the order of the combined vocabulary of all the samples.</code>
136	<code>freq.list.J <- get(all.freq.lists[[J]]) #Collects the appropriate frequency list.</code>
137	<code>J.in.combined <- freq.list.J[freq.list.J=names(combined.sorted.freq.lis t)] #Pastes the names (the words) of the frequency vector alongside the frequency integers in a matrix.</code>
138	<code>J.in.combined[is.na(J.in.combined)] <- 0 #Changes "NA" to numeric 0.</code>
139	<code>assign(paste("combined", J, sep="."), J.in.combined) #Writes the current frequency list to file.</code>
140	<code>}</code>
141	<code>combined.table <- paste(names(combined.sorted.freq.list), combined.sorted.freq.list, combined.1, combined.2, combined.3, combined.4, combined.5, combined.6, combined.7, combined.8, combined.9, combined.10, sep="\t") #Creates a table with columns for the combined list and all of the component lists.</code>
142	<code>cat("Word\tTotal\t1\t2\t3\t4\t5\t6\t7\t8\t9\t10", combined.table, file=paste(text_type, p, "words_k", y, ".csv", sep="_"), sep="\n") #Saves the binary matrices with words for reference for the 10-way comparison.</code>

143	<code>combined.matrix <- paste(combined.1, combined.2, combined.3, combined.4, combined.5, combined.6, combined.7, combined.8, combined.9, combined.10, sep="\t") #Creates a table with columns for component lists only in the combined order. These will be used to calculate the Dice distance between them.</code>
144	<code>cat("1\t2\t3\t4\t5\t6\t7\t8\t9\t10", combined.matrix, file=paste(text_type, p, y, "k.csv", sep="_"), sep="\n") #Saves the binary matrices.</code>
145	<code>rm(list=ls(pattern=("freq.list.\d* combined.* selected. files text.file word.list word.vector I J")) #Removes vectors for a 10-way Dice comparison.</code>
146	<code>X <- X+1</code>
147	<code>}</code>
148	<code>rm(X, y, z)</code>
149	<code>#</code>
150	<code>#CALCULATE DICE DISTANCE 100 LEVEL</code>
151	<code>#</code>
152	<code>i <- 1</code>
153	<code>for (i in 1:100) {</code>
154	<code>j <- i*100</code>
155	<code>Dice_table <- read.table(file=paste(text_type, p, j, "h.csv", sep="_"), header=T, sep="\t", quote="")</code>
156	<code>Dice_table_T <- t(Dice_table) #Transposes the table to be in the order the dissimilarity function expects.</code>
157	<code>Dice.dist <- dissimilarity(Dice_table_T, y=NULL, method = "dice") #Calculates the Dice distance between the samples.</code>
158	<code>cat((paste(p, j[1], "mean", mean(Dice.dist), "sd", sd(Dice.dist), sep="\t")), file="Dice_report.r", sep="\n", append=T) #Calculates and records the mean and standard deviation for each matrix of Dice distances.</code>
159	<code>echo=T #Opens a connection to the results that will be output. They would be displayed to screen except that line 160 will “sink” this connection to a file instead.</code>
160	<code>sink("Dice_cent_matrix_report.r", append=T) #Connects to the file that will receive the Dice distance matrix for the current level.</code>

161	<code>options(width=10000) #Make sure the lines are wide enough.</code>
162	<code>print(j[1]) #Prints a label of the current level.</code>
163	<code>print(Dice.dist) #Prints the current matrix.</code>
164	<code>sink() #Sends the open print connection to the file specified previously.</code>
165	<code>i <- i+1</code>
166	<code>}</code>
167	<code>#</code>
168	<code>#CALCULATE DICE DISTANCE 1000 LEVEL</code>
169	<code>#</code>
170	<code>i <- 1</code>
171	<code>for (i in 1:k_limit) { #Runs one loop up to the last full 1000 types in the vocabulary of the shortest of the 10 comparison samples.</code>
172	<code>j <- i*1000</code>
173	<code>Dice_table <- read.table(file=paste(text_type, p, j, "k.csv", sep="_"), header=T, sep="\t", quote="")</code>
174	<code>Dice_table_T <- t(Dice_table)</code>
175	<code>Dice.dist <- dissimilarity(Dice_table_T, y=NULL, method = "dice")</code>
176	<code>cat((paste(p, j[1], "mean", mean(Dice.dist), "sd", sd(Dice.dist), sep="\t")), file="Dice_report.r", sep="\n", append=T)</code>
177	<code>echo=T #Opens a connection to the results that will be output. They would be displayed to screen except that line 160 will “sink” this connection to a file instead.</code>
178	<code>sink("Dice_kilo_matrix_report.r", append=T) #Connects to the file that will receive the Dice distance matrix for the current level.</code>
179	<code>options(width=10000) # Make sure the lines are wide enough.</code>
180	<code>print(j[1]) #Prints a label for the current level.</code>
181	<code>print(Dice.dist) #Prints the current matrix.</code>
182	<code>sink() #Sends the open print connection to the file specified previously.</code>
183	<code>i <- i+1</code>
184	<code>rm(i, j, k, K, k_limit, k_range, L, m, n, q, s, x) #These specific variables are removed here. Including them as patterns in the regular expression below will also match and remove other variables that are still needed.</code>

185	<code>rm(list=ls(pattern=("Dice.* freq.list L.//d{1,2} Leng.//d{1,2}")))</code>
186	<code>}</code>
187	<code>}</code>
188	<code>cat((paste("time started", time.started, sep=" ")), file="Dice_report.r", sep="\n", append=T) #Creates a report of the time the script started.</code>
189	<code>time.finished <- format(Sys.time(), "%a %b %d %X %Y")</code>
190	<code>cat((paste("time finished", time.finished, sep=" ")), file="Dice_report.r", sep="\n", append=T) #Adds a line giving the time the script finished.</code>
191	<code>echo=T</code>
192	<code>alarm() #Audibly signals the script has finished running.</code>
193	<code>rm(list=ls(all=T)) #Clears the memory.</code>

C.2 Automated script for calculating Dice distance without replacement.

To modify the script in C.1 so that files are selected without replacement, line 18b should be added following line 18 of the script:

18b	<code>already.selected = NULL #Creates an empty vector for the list of files that have already been selected for samples.</code>
-----	--

The lines below should replace lines 22-28 in the C.1 script:

22	<code>#RANDOM SAMPLE without replacement</code>
23	<code>#</code>
24	<code>for (x in 1:10) { #Creates 10 random samples</code>
25	<code>q <- 100*p #Number of files to be randomly selected at base level -- usually 100 (= 1 million tokens) times the level being created, e.g. 3 x 1 million = 3 million.</code>
26	<code>s <- sample(setdiff(list.length, already.selected), q, replace = FALSE) #Creates a vector of the reference numbers of the files to be selected. The function sample only selects files from the subset of files not already selected for previous samples. Replace is set to FALSE which means a file will not be selected more than once for a sample.</code>
27	<code>s <- names(file.list[s]) #Assigns the file names to their reference numbers.</code>
28	<code>cat((paste("Random list", p, x, s, sep="\t")), file="Dice_report.r", sep="\n", append=T) #Creates a report of which files were assigned to each sample.</code>
28b	<code>already.selected <- union(already.selected, s) #Adds the current list of samples to the list of already selected files.</code>

Appendix D.1

R-script for adding head words to the BNC/COCA lists

1	<code>library(stringr) #loads the stringr library (Wickham, 2010) that contains the command str_extract that is used in line 10.</code>
2	<code>family.list <- scan(file = choose.files(), what="char", sep="\n", quote="", comment.char="") #Inputs the family list created by combining all of the BNC/COCA lists used with the Range program.</code>
3	<code>family.list.no.digits <-gsub(" \\d{1,2}", "\t_*", family.list, perl=T) #Replaces space and digits with a marker for later headword insertion.</code>
4	<code>family.list.hw.marked <-gsub("(^\\b)", "`", family.list.no.digits, perl=T) #Adds marker to headwords.</code>
5	<code>family.list.types.aligned <-gsub("^\\t", "", family.list.hw.marked, perl=T) #Removes all the indenting tabs.</code>
6	<code>family.list.pasted <- paste(family.list.types.aligned, sep = "") #Pastes the individual text strings together as one continuous text.</code>
7	<code>cat(family.list.pasted, file="family_list_pasted.txt", sep="") # Saves the list as a text file. This was the easiest way to ensure the list is one continuous file.</code>
8	<code>family.list.inline <- scan(file = "family_list_pasted.txt", what="char", sep="`", quote="", comment.char="") #Inputs the combined list treating each headword marker as the beginning of a new line.</code>
9	<code>for (i in 1:length(family.list.inline)) { #In a loop, processing one line at a time from the first line to the length of the file.</code>
10	<code>head_word <- str_extract(family.list.inline[i], "(?U) (^.+\\b)") #Reads the first word, i.e. the headword from each line.</code>
11	<code>line_hw_added <-gsub("_", head_word, family.list.inline[i], perl=T) #Replaces the underline character that was inserted after each word in the family in line 3 with a duplicate of the headword.</code>
12	<code>types_divided <-gsub("*", "\n", line_hw_added, perl=T) #Re-divides the file with each word type on an individual line</code>
13	<code>cat(types_divided, file="file_list_hw_added.txt", sep="\n", append=T) #Saves the list of word types with headword labels as a text file.</code>
14	<code>}</code>

Appendix D.2

R-scripts for merging text type frequency lists and the BNC/COCA type-head word list

D.2.1 Script for merging words common to both the BNC/COCA lists and the text type sample list

1	<code>ACAD_10M <- read.csv ("ACAD_10_M_freq_list_FOR_TYPE_FAMILY_STUDY_uppercase.csv", header=T, sep="\t") #inputs the frequency file with the sample of one of the text types, e.g. Academic</code>
2	<code>BNC_COCA_FAMILIES <- read.csv ("BNC_COCA_family_list_hw_added.csv", header=T, sep="\t") #inputs a csv file made from the list created with the script in Appendix C.1</code>
3	<code>ACAD_FAMILIES_merged <- merge(BNC_COCA_FAMILIES, ACAD_10M, by="Word_Type", all=F) #merges the two lists by joining all the words common to both lists; all=F or headwords will be changed to NA</code>
4	<code>ACAD_FAMILIES_merged[is.na(ACAD_FAMILIES_merged)] <- 0 #changes "NA" to numeric 0 #replaces the NA with a 0 as the frequency for BNC/COCA words that did not occur in the sample of that text type</code>
5	<code>write.csv(ACAD_FAMILIES_merged, "ACAD_FAMILIES_merged.csv") #saves the merged list as a csv file</code>

D.2.2 Script for listing words only found in the text type sample list

The resulting list of words that were not found on the BNC/COCA list must be appended manually to the table created in D.2.1.

1	<code>ACAD_not_BNC_COCA <- setdiff(ACAD_10M\$Word_Type, BNC_COCA_FAMILIES\$Word_Type) #finds all word types in the text type sample which do not occur on the BNC/COCA lists</code>
2	<code>cat("Word_Type", ACAD_not_BNC_COCA, file="ACAD_not_BNC_COCA.csv", sep="\n") #saves this list as a csv file</code>
3	<code>ACAD_not_BNC_COCA <- read.csv("ACAD_not_BNC_COCA.csv", header=T, sep="\t") #re-inputs the list of non-BNC/COCA words</code>
4	<code>ACAD_not_BNC_COCA_w_freq <- merge(ACAD_not_BNC_COCA, ACAD_10M, by="Word_Type", all=F) #merges the non-BNC/COCA words with the text type frequency list;all=F or headwords will be changed to NA</code>
5	<code>write.csv(ACAD_not_BNC_COCA_w_freq, "ACAD_not_BNC_COCA_w_freq.csv") #saves the frequency list of headwords, word types and frequencies</code>

Appendix E.1

Multi-file frequency script

This script scans text files and returns a word type frequency list showing the total frequency for each word and its frequency in each of the individual files. The file "word_frequencies.csv" is saved as a comma-separated-value file that can be opened using any spreadsheet software. For OpenOffice or LibreOffice, simply open the file using the "Open" command on the "File" ribbon or Ctrl+O. For MS-Excel, select the "Data" ribbon and click on "From Text" and then click "Next" until "Finish." There is no theoretical limit to the number of files that can be processed at one time (See line 18). This script is set to remove tags enclosed in angled brackets < >. If one does not wish to remove tags, place a hash mark # in front of line 8.

Using this script requires only a minimal knowledge of R. Once the R program has been installed. The command `getwd()` will return the current work directory. The command `setwd()` can be used to change the work directory to the folder containing the texts one wishes to include in the frequency table. The right-hand column of the table below can be copied and saved as a text file by copying and pasting it into a program like NotePad. The file name should end in the suffix `.r` rather than `.txt`, e.g. `MultiFile_frequency.r` so that the frequency script does not include it in the frequency table. This file should also be saved in the work directory. At the command prompt one can then enter the command `source("MultiFile_frequency.r")`. The file will then run automatically.

1	<code>library(gtools) #The mixedsort function that is used repeatedly in this script is from the Warnes' (2012) gtools library.</code>
2	<code>selected.files <- list.files(path=getwd(), pattern="*.txt") #Selects all the files from the working directory. If one wishes to only select particular files within the directory, simply replace everything to the right of the arrow with choose.files()</code>
3	<code>selected.files <- mixedsort(selected.files) #Reorders the files in their proper numeric order.</code>
4	<code>combined.list = NULL #Initiates the vector to receive the combined list.</code>
5	<code>for (i in 1:length(selected.files)) {</code>
6	<code> text.file<-scan(selected.files[[i]], what="char", sep="\n", quote="", comment.char="") #Inputs the text file.</code>

7	<code>text.file<-tolower(text.file) #Changes all alphabetic characters to lower case.</code>
8	<code>text.file<-gsub("<.*?>", "", text.file, perl=T) #Removes tags.</code>
9	<code>word.list<-strsplit(text.file, "\\W+") #Extracts words from the file.</code>
10	<code>word.vector<-unlist(word.list) #Changes output from strsplit back into a vector.</code>
11	<code>word.vector<-word.vector[nchar(word.vector)>0] #Removes any remaining empty strings.</code>
12	<code>freq.list<-table(word.vector) #Creates a table of named integers (word types and their frequencies).</code>
13	<code>assign(paste("freq.list",i,sep="."),freq.list) #Creates a unique file name for each iteration of this loop.</code>
14	<code>combined.list<-c(combined.list, freq.list) #Adds each frequency list to combined list.</code>
15	<code>}</code>
16	<code>combined.freq <- as.table(tapply(combined.list, names(combined.list), sum)) #Transforms the series of individual frequency lists into a unified frequency list.</code>
17	<code>combined.sorted.freq.list<-sort(combined.freq, decreasing=T) #Sorts the frequency list in descending order of frequency.</code>
18	<code>all.freq.lists <- ls(pattern="freq.list\\.\\d{1,3}") #Lists all frequency files in memory with 1-3 digits. This allows a frequency list for up to 999 files. To increase that, simply change the 3 to a 4 in this line and line 26. That will allow for up to 9,999 files.</code>
19	<code>all.freq.lists <- mixedsort(all.freq.lists) #Reorders lists numerically.</code>
20	<code>for (j in 1:length(all.freq.lists)) { #Loop sorts each individual frequency list into the combined order.</code>
21	<code>freq.list.j<-get(all.freq.lists[[j]])</code>
22	<code>j.in.combined<-freq.list.j[freq.list.j=names(combined.sorted.freq.list)] #Pastes the names of the frequency vector alongside the frequency integers in a matrix.</code>

23	<code>j.in.combined[is.na(j.in.combined)] <- 0</code> #Changes "NA" to numeric 0.
24	<code>assign(paste("combined", j, sep="."), j.in.combined)</code> #Writes the current frequency list to file.
25	<code>}</code>
26	<code>all.j.in.combined<-ls(pattern="combined.\\d{1,3}")</code> #Lists all combined files in memory with 1-3 digits.
27	<code>all.j.in.combined <- mixedsort(all.j.in.combined)</code>
28	<code>combined.file.count<-1:length(selected.files)</code> #Counts number of files originally selected.
29	<code>combined.file.list<-paste("combined", combined.file.count, sep=".")</code> #Creates the file names for the combined lists by catenating "combined" with each file number separated by a period.
30	<code>combined.table<-paste(names(combined.sorted.freq.list), combined.sorted.freq.list, sep="\t")</code> #Creates a table with columns for the words of the combined vocabulary and the total frequency.
31	<code>for (x in 1:length(combined.file.list)) {</code>
32	<code>add.to.table<-get(combined.file.list[[x]])</code>
33	<code>combined.table<-paste(combined.table, add.to.table, sep="\t")</code> #Adds each of the component lists to the combined table.
34	<code>}</code>
35	<code>header<-paste(selected.files, sep="\t")</code> #Creates header labels for the component frequency lists.
36	<code>table.header <- c("Word_type", "Total_Frequency", header, "\n")</code> #Adds column labels for the Word_type and total columns. The hard return "\n" at the end will force the first row of the frequency table onto the second line of the file.
37	<code>cat(table.header, file="word_frequencies.csv", sep="\t")</code> #Saves the header row to a spreadsheet file.
38	<code>cat(combined.table, file="word_frequencies.csv", sep="\n", append=TRUE)</code> #Adds the table to the spreadsheet file.
39	<code>rm(list=ls(all=T))</code> #Clears the memory so files do not interfere with subsequent processes.

Appendix E.2

Multi-file dispersion and range script

This script can be added to the script in Appendix E.2 and the combined script can be saved as “MultiFile_frequency_range_DP.r” Alternately, it can be used independently if it is saved and run in the same folder in which the results from the frequency script `word_frequencies.csv` have been saved. The script can be executed in the same way as the script in Appendix E.1 by writing the name of the script in quotation marks inside the parentheses of the command `source()`. The script will save separate and combined files for frequency, range and DP. This is designed as a “sign of life” when running very large data sets and as insurance in case of power failures or other accidents.

40	#Vocabulary Range
41	#
42	<code>vocab.table <- read.csv("word_frequencies.csv", header=TRUE, sep="\t")</code> #re-inputs the frequency table from the script in Appendix E.1
43	<code>vocab.range = NULL</code> #Initiates the vector to receive the range scores. Range, here, means the number of files in which a word type occurs.
44	<code>length(vocab.range) <- 2500</code> #Preallocates memory space to the vector so that a new space does not need to be found as it is enlarged with each iteration of the loop below. This number should be equal to the number of files being scanned.
45	<code>v <- length(vocab.table[,1])</code> #Calculates the number of word types in the combined list
46	<code>f <- length(vocab.table)-3</code> #Calculates the number of columns and then subtracts one for the word_type, total and closing columns
47	<code>x <- f+2</code> #Calculates the column number of the final file, i.e. the number of files plus the word_type and total columns.
48	<code>for (i in 1:v) {</code>
49	<code> range.count <- sum(vocab.table[i,3:x] >0)</code> #Counts number of files in which each type occurs, i.e. frequency is >0.
50	<code> vocab.range <- paste(vocab.range, range.count, sep="\n")</code>
51	<code>}</code>
52	<code>cat("Range", file="vocab_range.csv")</code>
53	<code>cat(vocab.range, file="vocab_range.csv", sep="\n", append=TRUE)</code>
54	#

55	<code>range.table <-unlist(read.csv("vocab_range.csv", header=TRUE, sep="\t"))</code>
56	<code>range_freq.table <- paste(vocab.table[,1], vocab.table[,2], range.table, sep="\t")</code>
57	<code>range_freq.table.header <- c("Word_type", "Total_Frequency", "Range", "\n")</code> #Adds column labels for the Word_type and total columns. The hard return "\n" at the end will force the first row of the frequency table onto the second line of the file.
58	<code>cat(range_freq.table.header, file="word_freq_range.csv", sep="\t")</code>
59	<code>cat(range_freq.table, file="word_freq_range.csv", sep="\n", append=TRUE)</code>
60	<code>#</code>
61	<code>#DP (Gries, 2008, 2010)</code>
62	<code>#</code>
63	<code>vocab.DP = NULL</code> #Initiates the vector to receive the range scores.
64	<code>length(vocab.DP) <- 25000</code> #Preallocates memory space to the vector so that a new space does not need to be found as it is enlarged with each iteration of the loop below. Should equal number of files being scanned.
65	<code>for (j in 1:v) {</code>
66	<code> DP.score <- sum(abs((1/f) - vocab.table[j,3:x]/vocab.table[j,2]))/2</code>
67	<code> vocab.DP <- paste(vocab.DP, DP.score, sep="\n")</code>
68	<code>}</code>
69	<code>cat("DP", file="vocab_DP.csv")</code>
70	<code>cat(vocab.DP, file="vocab_DP.csv", sep="\n", append=TRUE)</code>
71	<code>#</code>
72	<code>DP.table <-unlist(read.csv("vocab_DP.csv", header=TRUE, sep="\t"))</code>
73	<code>range_freq.DP.table <- paste(vocab.table[,1], vocab.table[,2], range.table, DP.table, sep="\t")</code> #Creates a file with the combined list of word types present in the scanned files, each word's total frequency, its range and DP score.

74	<code>range.freq.DP.table.header <- c("Word_type", "Total_Frequency", "Range", "DP", "\n") #Adds column labels for the Word_type and total columns.</code> The hard return "\n" at the end will force the first row of the frequency table onto the second line of the file.
75	<code>cat(range.freq.DP.table.header, file="word_freq_range_DP.csv", sep="\t")</code>
76	<code>cat(range_freq.DP.table, file="word_freq_range_DP.csv", sep="\n", append=TRUE)</code>
77	<code>#rm(list=ls(all=T)) #Clears the memory so files do not interfere with subsequent processes.</code>

The resulting `word_freq_range_DP.csv` file will appear similar to the sample table below. Notice the words at the top of the list have a very low DP score since they were very evenly dispersed. Those at the bottom of the list are close to 1 since they are only present in one sample and are therefore very poorly dispersed.

	A	B	C	D
1	Word_type	Total_Frequency	Range	DP
2	the	17414	25	0.0520500747
3	of	8335	25	0.0921895621
4	to	6915	25	0.0620390456
5	and	6503	25	0.0584837767
6	a	5612	25	0.0677548111
7	in	5458	25	0.0714254306
20452	zion	1	1	0.96
20453	zones	1	1	0.96
20454	zoological	1	1	0.96
20455	zoomed	1	1	0.96
20456	zoos	1	1	0.96
20457	zurich	1	1	0.96

Appendix F

Frequency lists from two general writing samples of 23.8 million tokens each

(Frequencies are given per million tokens.)

Sample A			Sample B	
Rank	word type	frequency	word type	frequency
1	the	65,816.3	the	66,205.9
2	of	33,453.3	of	33,255.6
3	to	27,310.8	to	27,289.4
4	and	27,154.9	and	26,845.9
5	a	23,179.5	a	23,071.1
6	in	21,501.5	in	21,382.2
7	is	10,349.5	is	10,473.0
8	for	9,881.8	for	9,909.4
9	that	9,627.7	that	9,555.5
10	was	8,414.7	it	8,248.7
11	it	8,291.5	was	8,236.1
12	on	7,937.5	on	7,942.4
13	s	7,142.9	s	7,202.3
14	with	7,039.5	with	7,057.1
15	as	6,872.8	as	6,800.0
16	be	6,514.0	be	6,531.4
17	by	6,251.3	by	6,232.4
18	at	5,552.1	at	5,489.6
19	he	5,121.5	he	5,054.3
20	are	4,838.9	are	4,893.4
21	from	4,807.0	from	4,805.6
22	i	4,560.5	i	4,455.6
23	this	4,375.7	this	4,316.4
24	have	4,226.2	have	4,257.2
25	but	4,216.6	but	4,186.4
26	his	4,046.3	his	4,056.7
27	not	4,008.4	not	4,018.1
28	which	3,992.2	which	4,005.3
29	had	3,911.4	an	3,778.1
30	an	3,798.9	had	3,757.2
31	or	3,667.1	or	3,633.3
32	they	3,606.1	they	3,583.2
33	were	3,240.7	were	3,240.6
34	has	3,066.0	has	3,155.9
35	their	3,015.8	their	3,020.9
36	you	2,941.1	will	2,931.1
37	one	2,842.3	you	2,869.3
38	will	2,823.2	one	2,791.3
39	been	2,634.0	been	2,608.6
40	we	2,574.6	we	2,567.3
41	all	2,542.9	all	2,529.4

42	there	2,519.8	there	2,498.3
43	who	2,359.1	who	2,320.3
44	more	2,241.8	would	2,315.8
45	would	2,238.0	more	2,249.2
46	can	2,130.1	can	2,109.2
47	when	2,102.4	when	2,104.9
48	its	2,078.3	its	2,095.9
49	if	2,033.1	if	2,059.6
50	up	2,017.4	up	2,029.5
51	so	1,871.3	so	1,849.9
52	out	1,809.0	out	1,829.1
53	new	1,759.9	new	1,757.1
54	no	1,723.8	no	1,690.3
55	some	1,663.5	about	1,670.0
56	about	1,653.6	said	1,637.9
57	time	1,637.1	two	1,634.1
58	two	1,623.3	some	1,629.0
59	said	1,612.3	than	1,614.6
60	other	1,592.0	time	1,600.1
61	than	1,577.3	other	1,574.5
62	into	1,568.7	also	1,552.2
63	also	1,539.3	into	1,549.4
64	only	1,529.5	only	1,541.0
65	her	1,499.2	first	1,454.1
66	first	1,456.6	t	1,444.6
67	what	1,455.0	what	1,417.5
68	she	1,430.3	her	1,414.0
69	may	1,411.6	them	1,392.5
70	them	1,408.9	after	1,377.4
71	t	1,390.1	over	1,354.1
72	after	1,383.0	she	1,351.2
73	over	1,362.0	may	1,350.7
74	people	1,325.1	people	1,308.4
75	could	1,283.7	could	1,259.5
76	these	1,217.2	year	1,249.7
77	most	1,209.8	these	1,207.6
78	year	1,202.7	most	1,197.8
79	now	1,139.6	such	1,138.8
80	such	1,138.9	do	1,132.9
81	do	1,138.2	now	1,124.7
82	my	1,118.0	years	1,112.3
83	any	1,108.8	any	1,104.1
84	years	1,106.2	my	1,081.0
85	then	1,081.0	like	1,078.0
86	like	1,079.9	then	1,073.8
87	your	1,068.0	should	1,059.7
88	should	1,047.0	many	1,038.6
89	many	1,045.3	your	1,018.4
90	him	1,022.7	very	999.6
91	very	1,021.6	him	992.9

92	our	989.5	made	976.0
93	made	985.7	being	968.8
94	being	974.8	last	952.4
95	work	949.2	work	943.0
96	between	947.9	between	941.9
97	last	939.5	our	936.8
98	even	923.1	well	928.8
99	us	923.0	even	922.9
100	well	919.9	government	921.8
101	because	903.0	because	920.3
102	see	900.7	where	914.5
103	where	899.7	us	897.6
104	way	884.4	see	897.5
105	just	880.7	much	883.7
106	three	879.8	just	883.2
107	much	876.4	mr	872.7
108	government	865.4	three	872.0
109	mr	852.0	way	860.5
110	those	843.9	before	843.9
111	before	827.6	those	825.3
112	world	807.3	make	795.4
113	make	798.4	world	795.4
114	through	788.2	through	794.5
115	how	765.3	how	774.5
116	me	763.2	good	730.8
117	own	754.2	back	729.7
118	back	746.4	both	729.2
119	both	738.7	me	726.4
120	per	736.6	per	722.7
121	good	732.4	own	721.9
122	day	724.2	day	711.3
123	down	719.4	down	709.6
124	still	711.3	under	697.9
125	life	700.8	still	697.4
126	off	691.8	off	693.7
127	however	674.9	however	688.0
128	under	673.2	old	683.2
129	use	667.6	use	666.4
130	old	667.4	take	661.3
131	take	661.4	against	654.3
132	did	656.9	life	651.3
133	against	653.5	did	649.1
134	long	644.6	while	645.3
135	get	637.7	long	642.6
136	while	634.5	get	630.8
137	used	630.0	each	626.2
138	each	628.3	used	624.2
139	another	619.2	too	620.3
140	too	615.4	same	618.9
141	same	613.1	must	617.5

142	part	605.8	another	597.5
143	must	587.6	part	596.0
144	home	583.4	home	583.5
145	party	578.5	company	581.2
146	end	576.4	end	579.5
147	during	574.8	since	571.0
148	man	564.8	during	564.3
149	go	559.0	party	557.2
150	since	553.2	national	554.4
151	cent	550.1	system	552.1
152	great	547.6	go	547.6
153	system	539.2	set	537.7
154	company	537.8	high	535.7
155	right	533.4	man	532.6
156	little	530.5	right	530.7
157	high	523.7	great	526.8
158	set	522.3	cent	524.4
159	national	521.3	little	519.2
160	four	513.3	four	516.3
161	second	504.7	public	514.8
162	group	503.1	next	507.0
163	need	502.9	place	506.6
164	place	500.9	second	502.4
165	british	499.7	local	501.5
166	number	493.6	need	500.4
167	left	490.9	market	497.6
168	next	490.0	number	492.5
169	might	485.5	house	491.8
170	found	485.4	might	489.2
171	house	484.3	british	481.5
172	public	480.8	although	480.3
173	state	476.6	without	472.9
174	without	474.6	group	472.3
175	market	472.6	until	471.8
176	children	472.2	general	471.8
177	local	472.1	within	470.5
178	within	471.8	state	470.5
179	few	471.1	says	466.4
180	although	470.5	left	465.0
181	around	470.3	children	462.4
182	later	468.7	around	461.2
183	come	464.4	few	458.5
184	london	464.1	country	454.5
185	until	461.6	later	452.0
186	never	460.6	small	451.3
187	women	460.4	found	448.6
188	small	452.5	london	448.6
189	general	450.4	come	447.7
190	country	445.3	never	446.4
191	different	439.7	different	440.7

192	john	436.7	best	439.0
193	says	436.1	john	436.9
194	less	434.9	put	436.6
195	best	433.4	school	432.3
196	put	432.3	know	432.0
197	five	432.0	less	430.2
198	know	429.9	does	429.6
199	given	426.9	five	429.2
200	again	425.5	water	426.9
201	does	421.9	money	425.6
202	say	421.1	given	422.4
203	early	419.5	again	419.7
204	men	415.5	say	419.6
205	war	414.8	though	418.7
206	family	413.6	d	414.5
207	always	413.4	far	414.5
208	often	412.0	case	412.9
209	far	410.4	political	410.7
210	every	408.5	often	409.3
211	rather	407.9	here	408.9
212	case	407.4	rather	407.9
213	large	406.9	always	407.7
214	called	406.3	every	405.2
215	came	405.7	week	405.0
216	help	405.3	large	404.6
217	p	404.4	don	404.5
217	though	404.4	power	404.4
219	school	399.1	early	403.7
220	important	398.9	business	401.1
221	business	395.1	war	399.9
222	find	394.5	council	399.5
223	political	393.8	came	398.8
224	social	393.7	called	397.6
225	information	392.3	minister	395.0
226	power	391.1	going	392.3
227	council	390.0	women	390.3
228	m	389.5	men	389.9
229	d	388.7	important	387.9
230	think	388.3	p	387.6
231	minister	387.4	re	387.5
232	million	386.1	million	387.5
233	don	385.9	think	387.4
234	course	385.6	major	386.7
235	week	384.2	help	386.3
236	give	382.6	course	385.1
237	water	382.3	family	383.7
238	money	381.8	m	381.7
239	become	380.9	give	381.3
240	here	380.1	find	374.7
241	service	379.4	social	374.5

242	going	377.8	support	372.5
243	away	376.7	days	371.2
244	re	372.9	further	369.8
245	support	371.9	point	367.3
246	days	371.3	become	366.1
247	major	370.0	members	365.2
248	once	369.2	development	363.2
249	south	366.9	information	362.7
250	further	366.3	form	361.6
251	members	365.9	away	361.0
252	labour	365.6	once	360.7
253	half	363.2	half	359.7
254	fact	362.4	president	358.6
255	took	361.8	service	358.0
256	taken	361.4	interest	357.1
257	night	360.9	yesterday	356.6
258	point	360.4	fact	356.4
259	development	358.0	taken	355.7
260	full	357.1	south	355.5
261	having	356.2	full	354.7
262	form	355.7	night	353.4
263	example	353.6	based	352.6
264	based	352.4	having	352.2
265	young	351.4	already	352.1
266	yesterday	349.5	example	350.9
267	yet	348.8	labour	350.0
268	possible	347.6	took	348.2
269	look	346.6	area	346.3
270	already	346.2	yet	345.3
271	went	346.0	north	344.3
272	president	343.4	want	344.1
273	area	343.1	better	339.1
274	order	338.0	side	338.6
275	want	334.5	order	336.8
276	top	333.4	went	336.6
277	international	332.5	top	336.3
278	told	330.5	possible	334.8
279	making	329.7	making	334.1
280	working	328.2	young	331.8
281	including	328.1	look	331.2
282	side	327.3	change	331.1
283	change	325.8	got	330.7
284	britain	325.4	control	330.0
285	better	325.2	told	328.5
286	interest	323.9	international	327.8
287	west	323.6	times	327.2
288	least	323.3	britain	326.8
289	got	322.4	working	326.1
290	six	321.9	six	325.8
291	almost	320.8	west	325.3

292	others	320.4	economic	324.9
293	whether	319.5	cost	322.5
294	times	318.5	including	321.8
295	police	317.5	available	320.4
296	seen	317.0	open	319.7
297	available	316.4	held	318.1
298	control	316.2	months	317.1
299	thought	315.1	almost	315.6
300	open	314.8	whether	315.4
301	economic	312.2	least	315.3
302	show	311.8	city	315.2
303	months	311.2	following	314.3
304	north	311.1	something	313.0
305	things	309.5	things	311.8
306	following	308.3	seen	310.9
307	line	307.5	level	310.4
308	known	307.3	known	310.0
309	century	307.3	line	309.6
310	england	307.1	show	309.6
311	level	306.7	england	309.2
312	problems	305.1	thought	309.2
313	something	304.5	trade	306.8
314	city	304.2	problem	306.4
315	held	303.9	together	305.5
316	why	303.0	enough	302.8
317	among	300.4	problems	302.4
318	able	299.5	really	302.0
319	road	298.9	free	300.5
320	problem	298.2	among	300.3
321	became	297.5	century	299.9
322	european	297.0	others	299.6
323	quite	296.2	why	298.1
324	free	296.1	centre	297.6
325	today	295.5	police	296.6
326	together	295.1	price	295.8
327	whole	292.6	union	294.5
328	court	292.1	able	293.4
329	really	291.9	today	292.2
330	east	291.1	road	292.0
331	centre	290.2	court	291.6
332	services	288.0	bank	290.6
333	enough	287.7	head	290.3
334	office	287.6	law	289.3
335	several	287.1	policy	288.8
336	cost	286.4	services	288.7
337	church	285.4	whole	288.3
337	run	285.4	main	285.5
339	head	285.2	quite	285.3
340	means	285.0	became	285.1
341	future	284.9	east	284.8

342	hand	283.1	european	284.0
343	policy	282.7	means	283.4
344	price	282.7	office	282.1
345	main	281.6	run	281.0
346	health	281.0	c	280.3
347	book	280.6	several	278.1
347	c	280.6	former	277.6
349	period	279.9	hand	276.8
350	trade	277.1	church	276.7
351	himself	275.8	play	276.3
352	present	275.4	management	275.8
353	former	275.2	using	275.8
354	play	273.9	book	275.6
355	third	272.6	industry	275.4
356	meeting	270.9	meeting	274.7
357	short	270.5	future	274.7
358	perhaps	269.6	period	274.6
359	american	269.0	rate	274.3
360	ever	268.3	team	273.6
361	law	268.0	himself	273.4
362	using	267.3	car	272.6
363	themselves	266.9	real	270.9
364	community	266.8	community	269.8
365	black	266.7	short	269.5
366	education	265.8	education	267.8
367	companies	265.8	big	267.3
368	real	265.0	ve	266.9
369	staff	264.4	themselves	265.5
370	industry	262.2	march	265.5
371	union	262.1	society	265.4
372	ve	262.1	foreign	265.3
373	club	262.1	american	264.9
374	english	262.0	perhaps	264.6
375	special	261.4	third	263.9
376	past	261.2	companies	263.1
377	europe	260.8	health	262.5
378	clear	260.8	tax	261.7
379	society	260.6	present	260.2
380	report	260.4	month	260.1
381	research	259.7	hard	259.8
382	united	259.7	either	259.6
383	team	259.6	act	259.4
384	won	259.1	europe	258.8
385	result	257.1	report	258.6
386	itself	255.5	ever	257.8
387	done	255.3	special	256.7
388	body	255.1	done	255.1
389	god	254.8	itself	255.0
390	foreign	254.7	united	253.9
391	month	254.7	clear	252.5

392	white	254.6	english	252.5
393	hard	252.6	keep	252.2
394	particular	252.4	won	251.7
395	either	251.2	result	251.7
396	action	251.0	action	251.3
397	view	249.7	club	251.2
398	range	249.5	co	251.2
399	bank	249.5	start	251.1
400	rate	249.5	god	249.7
401	process	249.3	view	249.7
402	above	248.4	particular	248.4
403	art	248.4	past	248.2
404	act	248.4	uk	247.9
405	likely	248.2	white	247.2
406	whose	247.2	provide	247.0
407	big	246.7	likely	246.4
408	keep	245.5	above	246.3
409	provide	245.5	whose	245.8
410	committee	245.0	staff	245.4
411	death	244.1	pay	245.3
411	human	244.1	building	244.7
413	care	244.0	committee	244.0
414	march	243.5	states	243.2
415	position	243.1	black	243.1
416	pay	240.4	french	242.8
417	experience	240.4	countries	241.8
418	start	240.1	position	241.5
419	management	237.7	research	239.4
420	person	237.4	terms	239.2
421	terms	237.0	range	239.0
422	tax	236.6	art	238.1
423	name	236.4	body	237.0
424	age	236.1	round	236.9
425	states	236.0	nothing	236.4
426	round	235.5	areas	236.1
427	close	234.4	april	235.3
428	car	234.3	close	235.0
429	uk	234.2	process	234.0
430	street	234.1	financial	233.4
430	towards	234.1	street	233.2
432	nothing	233.3	june	233.1
433	co	232.6	final	232.4
434	building	232.2	cannot	232.2
435	asked	231.9	name	231.9
436	april	230.8	particularly	231.8
437	countries	230.5	effect	231.4
438	particularly	230.4	upon	231.2
439	areas	229.7	single	231.1
440	effect	228.6	probably	230.3
441	light	227.6	value	230.0

442	feel	227.0	land	229.4
443	face	226.6	light	229.4
444	non	226.5	towards	229.3
445	french	226.4	ago	228.9
446	needs	226.3	food	228.8
447	single	226.3	member	228.7
448	cannot	226.2	b	228.4
448	financial	226.2	central	228.0
450	late	225.5	cut	227.5
451	ago	224.1	return	227.0
452	june	223.8	music	226.9
453	land	223.4	human	226.4
454	town	223.3	job	225.8
455	systems	223.1	experience	225.5
456	upon	223.0	town	225.4
457	difficult	222.3	along	225.4
458	along	221.9	common	225.2
459	self	221.8	taking	225.0
460	final	220.8	difficult	224.9
461	near	220.7	late	224.6
461	return	220.7	board	224.5
463	common	220.2	private	224.0
464	personal	220.0	asked	223.9
465	taking	219.8	systems	222.6
466	central	219.7	class	222.3
467	question	219.6	total	222.3
468	job	219.4	death	220.9
469	mother	219.2	care	220.5
470	probably	218.5	age	220.1
471	child	218.5	non	220.0
472	began	218.3	expected	219.7
472	sense	218.3	therefore	219.5
474	therefore	217.9	history	219.2
475	front	217.8	low	219.1
476	total	217.8	decision	219.0
477	am	217.7	term	217.9
478	food	217.4	question	217.1
479	involved	217.0	according	217.0
480	seems	216.2	face	216.8
481	room	215.2	personal	216.7
482	changes	214.9	feel	216.3
483	royal	214.6	workers	216.1
484	term	214.4	involved	215.5
485	class	214.2	thus	214.5
486	cut	213.9	changes	214.4
487	soviet	213.7	person	213.8
488	decision	213.5	front	213.7
489	lot	212.5	thing	213.7
490	kind	212.1	room	213.2
491	lost	211.9	seems	213.0

491	member	211.9	de	212.5
491	music	211.9	capital	212.2
494	move	211.7	lot	209.7
495	game	211.4	lost	209.1
496	low	211.2	child	209.0
497	private	210.6	near	208.4
498	according	210.5	david	208.2
499	role	210.4	secretary	207.7
500	certain	210.3	self	207.6
501	hours	210.2	am	207.6
502	programme	209.3	kind	207.0
502	turn	209.3	move	206.5
504	history	208.9	offer	206.2
505	expected	208.7	certain	206.1
506	offer	207.9	turn	205.5
507	father	207.4	soon	205.2
508	gave	206.9	works	205.0
509	air	206.8	costs	205.0
510	love	206.5	looking	204.8
511	board	205.4	sense	204.7
512	individual	205.4	game	204.6
513	outside	204.9	training	204.2
514	soon	204.4	began	203.6
514	thing	204.4	increase	203.5
516	groups	204.0	issue	203.1
517	idea	203.9	idea	202.9
518	value	203.9	role	202.7
519	brought	203.3	outside	202.4
519	force	203.3	production	202.1
521	thus	203.1	needs	202.0
522	training	202.8	brought	201.6
523	recent	202.5	hours	201.6
524	secretary	201.6	air	201.3
525	king	201.2	programme	199.9
526	issue	200.3	let	199.7
527	computer	199.8	royal	199.5
528	st	199.4	gave	199.0
529	led	199.3	quality	198.8
529	looking	199.3	matter	198.3
531	b	199.3	july	198.2
532	david	199.2	led	198.0
533	deal	199.1	recent	197.8
534	try	198.3	series	197.7
535	mrs	197.1	especially	197.6
536	strong	196.6	added	196.5
537	costs	195.8	love	196.4
538	across	195.3	v	195.8
539	especially	195.0	security	195.5
540	ll	194.7	usually	195.5
541	security	194.6	similar	195.3

542	mind	194.5	prime	195.1
542	record	194.5	e	195.0
544	behind	193.9	election	194.8
545	works	193.8	strong	193.7
546	increase	193.6	behind	193.5
547	let	193.6	force	192.8
548	evidence	192.6	groups	191.9
549	sometimes	192.6	deal	191.7
550	de	192.4	despite	191.3
551	similar	191.9	schools	191.2
552	usually	191.6	mind	191.2
553	added	191.2	individual	190.8
554	production	191.1	military	190.7
555	provided	191.1	st	190.7
556	live	191.1	makes	190.6
557	stage	191.0	stage	190.3
557	v	191.0	post	190.3
559	lead	190.7	share	189.3
560	workers	190.3	king	189.3
561	post	190.0	across	188.9
562	earlier	189.4	father	188.8
562	woman	189.4	season	188.5
564	e	189.1	try	188.5
565	quality	189.0	plan	188.4
566	series	188.7	include	188.2
567	believe	188.1	authority	187.8
568	saw	187.3	earlier	187.4
569	include	187.2	due	186.8
570	university	187.1	forces	186.8
571	matter	187.1	necessary	186.7
572	due	186.9	director	186.6
573	military	185.7	university	186.4
574	wife	185.6	sometimes	186.3
575	makes	185.5	doing	185.8
576	hospital	185.1	ll	185.5
577	plan	184.7	saw	185.5
578	doing	184.7	provided	185.5
579	indeed	184.4	record	185.3
580	sir	184.1	mother	184.4
581	director	184.0	mrs	183.8
582	nature	183.4	anything	183.2
583	despite	183.0	current	182.4
584	prime	182.6	living	182.1
585	felt	182.5	department	182.0
586	necessary	182.4	lead	181.9
587	living	182.1	believe	181.8
588	july	182.1	nature	181.7
589	anything	182.0	manager	181.6
590	manager	181.7	simply	181.5
591	capital	181.6	live	181.2

592	words	181.3	built	180.6
593	needed	180.9	agreement	180.3
594	german	180.9	language	179.6
595	season	179.8	evidence	179.4
596	simply	179.7	soviet	179.3
597	authority	179.6	announced	179.1
598	parents	179.2	indeed	179.1
599	operation	179.1	needed	179.0
600	forces	178.9	computer	178.9
601	current	178.7	woman	178.6
601	longer	178.7	france	177.9
603	minutes	177.2	paid	177.7
604	subject	176.7	sir	177.1
605	department	175.7	press	177.0
606	share	175.5	german	176.5
607	performance	175.1	operation	176.0
608	wanted	175.0	didn	175.9
609	france	174.6	call	175.7
610	call	174.0	style	174.7
611	ground	173.8	higher	174.3
611	success	173.8	project	174.2
613	didn	173.6	account	174.0
614	situation	173.5	seven	173.8
615	natural	173.3	felt	173.5
616	announced	173.2	success	173.4
617	required	172.8	minutes	173.4
617	seven	172.8	prices	173.4
619	rights	172.8	paper	173.1
620	eight	172.4	chairman	172.9
621	style	172.3	wanted	172.7
622	agreement	172.0	words	172.7
623	died	171.8	firm	172.6
623	true	171.8	news	172.6
625	middle	171.5	chief	172.5
626	election	171.4	technology	172.5
627	data	171.0	subject	172.4
628	reason	171.0	longer	171.6
629	defence	170.9	ground	171.5
630	built	170.0	growth	171.4
631	reported	169.7	hospital	170.9
632	weeks	169.6	students	170.8
633	higher	169.2	performance	170.7
634	chairman	168.9	agreed	170.1
635	friends	168.8	eight	169.9
636	certainly	168.6	parents	169.8
637	study	168.3	wife	169.7
638	lord	168.2	reason	169.5
639	chief	168.0	conditions	168.2
640	paid	167.3	visit	168.2
641	army	167.1	true	168.2

642	account	166.9	modern	168.1
643	win	166.8	cup	167.8
644	comes	165.8	weeks	167.8
645	ten	165.6	easy	167.4
646	project	165.3	middle	166.9
646	software	165.3	various	166.6
648	peter	165.2	reported	166.6
649	agreed	165.1	fish	166.1
649	various	165.1	greater	166.0
651	paper	165.0	defence	165.8
652	movement	164.5	natural	165.7
653	story	164.4	leader	165.4
654	type	163.8	plans	165.3
655	easy	163.6	died	165.3
656	visit	163.5	story	165.3
657	prices	163.5	planning	165.3
658	son	163.4	demand	165.0
659	described	163.0	data	164.7
660	modern	163.0	type	164.7
661	leave	162.9	study	164.3
662	planning	162.6	certainly	163.9
663	received	162.5	leave	163.5
664	seem	162.5	turned	163.1
665	environment	162.2	win	163.0
666	press	161.9	energy	162.9
666	technology	161.9	germany	162.9
668	started	161.7	points	162.6
669	cup	161.2	lord	162.3
669	decided	161.2	economy	162.2
671	turned	161.0	red	161.5
672	running	160.8	decided	161.3
673	approach	160.8	oil	161.3
674	bill	160.5	army	161.3
675	growth	160.3	comes	161.3
676	word	160.2	getting	161.1
677	film	159.9	recently	160.5
678	germany	159.8	required	160.2
678	produced	159.8	green	160.0
680	news	159.7	standard	160.0
681	instead	159.6	described	160.0
682	america	159.5	received	160.0
683	colour	159.3	commission	159.9
684	seemed	159.1	industrial	159.7
685	green	158.9	started	159.5
686	sea	158.8	instead	159.4
687	conditions	158.4	pressure	159.2
688	products	158.3	income	159.0
689	field	158.2	ten	158.9
690	plans	158.1	situation	158.7
691	points	158.0	rights	158.7

692	greater	157.7	sales	158.4
693	results	157.4	hope	158.2
694	test	157.1	son	158.1
695	throughout	157.0	bill	158.0
696	figure	156.6	design	158.0
697	energy	156.4	rates	157.9
698	association	156.3	association	157.9
699	ways	156.1	cases	157.8
700	schools	155.9	products	157.3
701	practice	155.6	results	157.2
702	firm	155.5	park	157.2
703	recently	155.4	america	157.1
704	cases	155.4	test	156.7
705	mark	155.3	product	156.3
706	included	155.0	included	156.2
707	red	154.6	seem	155.6
708	rest	154.3	concerned	155.3
709	poor	154.2	running	155.0
710	october	154.0	practice	155.0
711	read	153.9	buy	153.5
712	concerned	153.7	movement	153.0
713	getting	153.6	field	152.8
714	leading	153.5	friends	152.7
715	sure	153.4	rest	152.5
716	demand	153.3	established	152.4
716	parties	153.3	read	152.1
718	economy	153.2	approach	152.1
719	worth	153.2	conference	151.9
720	dr	153.1	scotland	151.9
720	followed	153.1	environment	151.8
720	hope	153.1	january	151.8
723	whom	151.9	competition	151.8
724	space	151.8	software	151.7
725	investment	151.6	sure	151.7
726	bring	151.3	simple	151.1
727	played	150.9	paul	150.9
728	design	150.8	followed	150.2
728	park	150.8	parties	150.1
730	met	150.6	worth	150.0
731	section	150.5	cause	149.8
732	scotland	150.5	bring	149.7
733	leader	150.4	investment	149.5
734	simple	150.3	produced	149.5
735	sent	150.2	written	149.4
736	conference	149.4	match	149.2
737	sales	149.3	space	149.1
738	written	149.1	poor	149.0
739	model	148.8	peter	148.9
740	trying	148.7	figure	148.7
741	language	148.4	lower	148.7

742	january	148.1	mean	148.7
743	basis	147.6	throughout	148.3
744	complete	147.2	october	148.3
745	wide	147.1	seemed	148.3
746	western	147.0	meet	147.9
747	december	146.9	leading	147.7
748	clearly	146.8	played	147.6
749	bit	146.7	mark	147.1
750	york	146.2	sent	147.1
751	september	146.1	basis	146.8
752	cause	145.8	dr	146.7
753	commission	145.7	sound	146.3
754	established	145.7	northern	146.2
755	paul	145.5	wide	146.0
756	continued	145.5	popular	145.9
757	standard	145.4	below	145.6
758	sun	145.3	relations	145.5
759	tell	144.6	bit	145.5
760	date	144.4	authorities	145.1
760	product	144.4	word	144.9
762	popular	144.3	books	144.8
763	actually	144.0	official	144.7
763	serious	144.0	film	144.6
765	published	143.9	amount	144.5
766	amount	143.9	increased	144.4
767	pressure	143.7	met	144.4
768	key	143.5	size	144.4
769	match	143.2	campaign	144.3
769	size	143.2	whom	144.2
771	mean	143.2	ways	144.1
772	chance	143.2	actually	144.0
773	buy	143.0	cash	143.6
774	opposition	142.9	players	143.6
775	continue	142.9	published	143.6
776	campaign	142.6	attempt	143.5
776	meet	142.6	supply	143.4
778	someone	142.3	chance	143.3
779	november	142.2	september	142.9
780	below	141.9	section	142.7
781	oil	141.8	developed	142.5
782	anyone	141.2	york	142.3
783	league	141.1	complete	142.1
784	industrial	141.0	rise	142.1
785	material	140.1	western	142.1
785	peace	140.1	george	142.1
787	developed	139.9	model	141.7
788	summer	139.5	trying	141.6
789	increased	139.4	material	141.2
790	produce	139.3	continued	140.9
791	hall	138.9	bad	140.9

792	letter	138.8	december	140.3
793	rates	138.7	shown	140.1
794	shown	138.7	choice	140.0
795	northern	138.3	coming	140.0
796	sort	137.9	opposition	140.0
797	inc	137.8	exchange	140.0
798	fish	137.7	sea	140.0
799	heard	137.5	fine	139.7
799	issues	137.5	everything	139.7
801	basic	137.1	serious	139.5
802	coming	136.9	summer	139.5
803	loss	136.7	table	139.5
804	relationship	136.6	shares	139.1
805	george	136.6	finally	138.9
806	james	136.6	loss	138.8
807	everything	136.3	fire	138.6
808	page	136.3	charge	138.5
809	finally	136.3	previous	138.1
809	looked	136.3	original	137.6
811	population	136.1	key	137.2
812	official	135.7	allowed	137.1
813	forward	135.6	playing	137.1
814	figures	135.5	continue	137.0
814	previous	135.5	colour	136.7
816	activities	135.4	carried	136.7
816	books	135.4	anyone	136.4
818	carried	135.4	ireland	136.4
819	else	135.2	clearly	136.2
820	players	135.0	letter	136.1
821	television	135.0	figures	136.1
822	direct	135.0	legal	136.1
823	knew	134.9	population	136.1
824	original	134.7	knew	135.8
825	allowed	134.4	scheme	135.5
826	authorities	134.3	produce	135.0
827	independent	133.9	tell	134.9
828	hold	133.7	date	134.8
829	charge	133.2	hall	134.8
829	playing	133.2	kept	134.7
831	designed	133.2	designed	134.7
832	rise	132.9	direct	134.5
833	spent	132.9	someone	134.5
834	choice	132.7	november	134.4
835	income	132.6	looked	133.1
835	lower	132.6	village	132.9
837	attention	132.6	worked	132.9
838	legal	132.6	hit	132.8
839	giving	132.5	else	132.7
839	heart	132.5	structure	132.7
841	considered	132.4	takes	132.7

842	sound	132.4	shows	132.6
843	attempt	132.1	forward	132.5
843	morning	132.1	league	132.2
845	fall	131.6	heard	132.1
846	worked	131.4	allow	132.1
847	shows	131.3	property	132.1
848	professional	131.1	details	131.9
849	risk	130.9	o	131.9
850	richard	130.5	page	131.9
851	michael	130.3	executive	131.7
852	kept	130.2	independent	131.6
853	details	130.1	spent	131.4
854	offered	130.0	sort	130.7
855	o	129.9	banks	130.5
856	list	129.7	hold	130.5
857	fire	129.7	average	130.4
857	successful	129.7	morning	130.4
859	event	129.6	goods	130.3
859	parts	129.6	inc	130.2
861	chapter	129.6	television	130.2
862	bad	129.4	leaders	130.1
863	fine	129.3	heart	130.0
863	parliament	129.3	issues	129.8
865	lives	129.2	parts	129.6
866	r	129.1	college	129.5
867	takes	129.0	sun	129.5
868	g	128.9	list	129.4
869	budget	128.9	cover	129.3
870	structure	128.6	fall	129.3
871	moment	128.4	considered	129.2
871	nor	128.4	everyone	129.2
873	majority	128.4	g	129.1
874	scheme	128.3	majority	129.1
875	hands	128.0	budget	129.0
876	africa	127.9	activities	128.8
876	scottish	127.9	hands	128.8
878	saying	127.6	talk	128.7
879	cash	127.5	risk	128.6
880	alone	127.4	successful	128.5
881	station	127.1	giving	128.1
882	unit	127.1	saying	127.9
883	writing	126.9	attention	127.9
884	beginning	126.8	michael	127.9
885	relations	126.7	county	127.8
885	supply	126.7	scottish	127.7
885	table	126.7	writing	127.3
888	potential	126.5	wrong	127.2
889	allow	126.4	nor	127.1
889	born	126.4	professional	127.1
891	claimed	126.4	parliament	126.8

891	reports	126.4	james	126.3
893	leaders	126.3	potential	126.3
894	attack	126.3	firms	126.1
895	exchange	126.2	basic	126.0
896	everyone	126.0	claimed	125.9
897	machine	126.0	behaviour	125.8
898	students	125.9	peace	125.6
899	charles	125.8	africa	125.4
899	shall	125.8	beginning	125.4
899	stop	125.8	remember	125.3
902	windows	125.8	knowledge	125.1
903	hit	125.6	radio	125.0
904	moved	125.6	event	124.8
905	failed	125.5	generally	124.8
906	cross	125.2	stop	124.7
907	competition	125.1	february	124.5
908	february	125.0	alone	124.5
909	ireland	124.6	attack	124.2
910	doubt	124.5	moment	124.2
911	race	123.9	equipment	124.1
912	response	123.4	unit	124.1
913	tried	123.2	darlington	123.9
914	remember	122.7	doubt	123.7
915	site	122.6	traditional	123.7
916	wrote	122.6	science	123.7
917	anti	122.5	cross	123.6
917	growing	122.5	stock	123.3
919	wrong	122.4	born	123.3
920	hour	122.3	stand	123.1
920	traditional	122.3	r	123.0
922	behaviour	122.3	concern	122.8
923	smith	122.2	levels	122.7
924	knowledge	122.1	extra	122.1
925	ec	122.1	moved	122.1
926	levels	121.6	august	121.8
927	talk	121.4	reports	121.8
928	answer	121.3	houses	121.4
929	influence	121.2	failed	121.3
930	events	121.1	wrote	121.1
931	average	121.1	appear	121.0
932	county	121.1	talks	120.9
933	senior	121.0	anti	120.7
934	treatment	120.8	relationship	120.7
935	talks	120.7	tried	120.6
936	gone	120.6	sale	120.4
937	college	120.5	claim	120.3
938	sold	120.3	chapter	120.1
939	nine	120.3	answer	119.5
940	appear	120.1	aid	119.2
941	advice	120.0	sector	119.2

942	william	119.9	credit	119.0
943	property	119.8	gone	118.9
944	standards	119.8	cabinet	118.8
945	appeared	119.6	finance	118.7
945	questions	119.6	sold	118.7
947	cover	119.4	growing	118.7
948	break	119.3	offered	118.7
949	extra	119.0	senior	118.7
949	friend	119.0	events	118.3
949	nearly	119.0	goes	118.3
952	created	118.8	contract	118.1
953	aid	118.7	treatment	118.1
954	science	118.6	picture	118.0
955	radio	118.4	shall	117.9
956	august	118.4	teachers	117.9
957	sector	118.0	lives	117.7
958	sale	117.9	appeal	117.6
958	shares	117.9	existing	117.5
960	goes	117.8	hour	117.5
961	generally	117.8	base	117.4
962	equipment	117.7	division	117.4
963	village	117.6	charles	117.3
964	ask	117.5	ec	117.1
965	mid	117.3	response	116.9
966	stand	117.3	nearly	116.8
967	ideas	117.1	lack	116.7
968	division	116.8	resources	116.6
969	picture	116.7	effective	116.5
970	executive	116.6	balance	116.4
971	daughter	116.4	ideas	116.4
972	star	116.1	richard	116.2
973	finance	115.9	windows	116.2
974	hotel	115.8	advice	116.1
974	nuclear	115.8	claims	116.1
976	prepared	115.6	elections	116.1
977	statement	115.3	mid	116.1
978	miles	115.2	hotel	115.8
979	career	115.2	william	115.4
979	resources	115.2	access	115.3
979	theory	115.2	nuclear	115.3
982	access	114.9	appeared	115.0
983	beyond	114.7	jobs	115.0
984	significant	114.7	corp	114.8
985	j	114.6	ask	114.7
986	contact	114.2	statement	114.6
986	remain	114.2	consider	114.6
988	version	113.9	machine	114.6
989	annual	113.9	annual	114.5
989	lack	113.9	inside	114.3
989	species	113.9	add	114.3

992	scale	113.8	significant	114.3
993	ministers	113.6	prepared	114.2
994	firms	113.5	nine	114.1
995	garden	113.5	double	113.9
995	pre	113.5	station	113.9
995	quarter	113.5	influence	113.8
998	region	113.3	miles	113.7
999	addition	113.2	tv	113.7
999	gives	113.2	door	113.5
1001	employment	113.1	organisation	113.5
1002	queen	113.0	plant	113.4
1003	couple	112.9	benefit	113.3
1003	goods	112.9	highly	113.3
1005	houses	112.9	ministers	113.1
1006	appeal	112.5	officials	112.9
1007	commercial	112.3	couple	112.8
1008	claims	112.2	smith	112.8
1009	culture	112.2	lines	112.6
1009	sunday	112.2	region	112.6
1011	effective	111.9	sunday	112.6