A COMPUTATIONAL APPROACH TO ENGLISH CORE VOCABULARY BASED ON THE BRITISH NATIONAL CORPUS

LI HAN HONG

DOCTOR OF PHILOSOPHY
CITY UNIVERSITY OF HONG KONG
OCTOBER 2011
A Computational Approach to English Core Vocabulary Based on the British National Corpus
基於英國國家語料庫的英語核心詞彙計算研究

Submitted to
Department of Chinese, Translation and Linguistics
中文、翻譯及語言學系
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy
哲學博士學位

by

Li Han Hong
李漢宏

October 2011
二零一一年十月
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>CONTENTS</td>
<td>I</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>VII</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>VIII</td>
</tr>
<tr>
<td>ACKNOWLEDGMENTS</td>
<td>IX</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>XII</td>
</tr>
<tr>
<td>1. INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Definitions of Core Vocabulary</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Significance of Core Vocabulary</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Core Vocabulary vs. Mental Lexicon</td>
<td>4</td>
</tr>
<tr>
<td>1.3.1 Definition of Mental Lexicon</td>
<td>4</td>
</tr>
<tr>
<td>1.3.2 Mental Lexicon vs. Core Vocabulary</td>
<td>5</td>
</tr>
<tr>
<td>1.3.3 Mental Lexicon and Word Frequency</td>
<td>6</td>
</tr>
<tr>
<td>1.4 Historical and Theoretical Development of Core Vocabulary</td>
<td>7</td>
</tr>
<tr>
<td>1.5 Selection Criteria for Core Vocabulary</td>
<td>11</td>
</tr>
<tr>
<td>1.5.1 The Qualitative Approach</td>
<td>11</td>
</tr>
<tr>
<td>1.5.2 The Quantitative Approach</td>
<td>14</td>
</tr>
<tr>
<td>1.5.3 Combination of Quantitative and Qualitative Approach</td>
<td>16</td>
</tr>
<tr>
<td>1.6 Balance of Domain Selection</td>
<td>17</td>
</tr>
<tr>
<td>2. LITERATURE REVIEW</td>
<td>20</td>
</tr>
<tr>
<td>2.1 Core Vocabulary in Language Teaching</td>
<td>20</td>
</tr>
<tr>
<td>2.1.1 Basic English</td>
<td>20</td>
</tr>
<tr>
<td>2.1.2 Thousand-Word English</td>
<td>22</td>
</tr>
<tr>
<td>2.1.3 The Teacher’s Word Book</td>
<td>23</td>
</tr>
<tr>
<td>2.1.4 General Service List</td>
<td>24</td>
</tr>
<tr>
<td>2.2 Core Vocabulary in Lexicography</td>
<td>25</td>
</tr>
<tr>
<td>2.2.1 The Concept of Defining Vocabulary</td>
<td>25</td>
</tr>
</tbody>
</table>
2.2.2 Core Vocabulary in LDOCE 27
   2.2.2.1 Longman Defining Vocabulary 27
   2.2.2.2 Longman Communication 3000 29
2.2.3 Core Vocabulary in OALD 30
2.2.4 Core Vocabulary in MEDAL 32
2.2.5 Core Vocabulary in CCALD 34
2.2.6 Core Vocabulary in CALD 35
2.2.7 Summary 35

2.3 Core Vocabulary in Psycholinguistics 36
   2.3.1 Core Vocabulary in AAC 36
   2.3.2 MRC Psycholinguistic Database 37
      2.3.2.1 Age 38
         2.3.2.1.1 Age and Core Vocabulary 38
         2.3.2.1.2 Age and Word Frequency 39
         2.3.2.1.3 Age and Language Acquisition 39
      2.3.2.2 Imageability 40
      2.3.2.3 Familiarity 41

2.4 Core Vocabulary in Corpus Linguistics 44
   2.4.1 Genre Definition 44
   2.4.2 Core Vocabulary vs. Genre 46
   2.4.3 Dispersion Index 46

2.5 Quantitative Approaches 48
   2.5.1 F-based Approach 48
   2.5.2 D-based Approach 49
      2.5.2.1 Range 50
      2.5.2.2 Dispersion Index 51
      2.5.2.2.1 Juilland’s D 52
      2.5.2.2.2 Carroll’s D 54
   2.5.3 U-based Approach 55
      Juilland’s U 56
      Carroll’s U 57

2.6 Criticism 58
   2.6.1 Limitations of the Qualitative Approach 59
   2.6.2 Limitations of the Quantitative Approaches 59
      2.6.2.1 Disadvantages of F-based Approach 59
      2.6.2.2 Disadvantages of D-based Approach 60
      2.6.2.2.1 Deficiencies of Range 61
      2.6.2.2.2 Deficiencies of Dispersion index: Juilland’s D vs. Carroll’s D 62
   2.6.2.3 Juilland’s U vs. Carroll’s U 64

3. PROPOSED RESEARCH FOR CORE VOCABULARY 66
3.1 New Horizons for Core Vocabulary 66
3.2 Hypotheses for Core Vocabulary 67
3.3 Outline of the Present Research 68

4. METHODOLOGY 70
4.1 Research Purpose 70
4.2 Corpora 71
4.3 Process 72
4.4 Tools 74

5. CORE VOCABULARY IN DIFFERENT AGE GROUPS 76
5.1 Age Information in BNC XML 77
  5.1.1 Age Information in the Written BNC XML 77
  5.1.2 Age Information in the Spoken BNC XML 78

5.2 Modification of Age Tagging in Spoken BNC XML 81
  Procedure to Achieve Uniformed Age Tags 81

5.3 Set-up of Age-Group Corpus 85
  5.3.1 Set-up of the Spoken Age-Group Corpus 85
  5.3.2 Set-up of the Written Age-Group Corpus 86
  5.3.3 Sum-up of Headword Statistics in Age-Group Corpus 86
    5.3.3.1 Word Tagging in BNC XML 87
    5.3.3.2 Headword Statistics in BNC XML 89
  5.3.4 Balanced Sampling of Age-Group Corpus from BNC XML 90

5.4 Supplementary Corpora for the Balanced Age-Group Corpus 91
  5.4.1 Supplementary Corpora for Spoken Age0 92
    5.4.1.1 Polytechnic of Wales Corpus (POW) 92
      5.4.1.1.1 Introduction 92
      5.4.1.1.2 Extracting Pure Corpus Text from POW 94
    5.4.1.2 The Child Language Data Exchange System (CHILDES) 96
      5.4.1.2.1 Introduction 97
      5.4.1.2.2 Extract Pure Corpus Texts from CHILDES 98
        5.4.1.2.2.1 Modification of the XML Files in Wells 100
        5.4.1.2.2.2 Extracting the Modified CHILDES 103
  5.4.2 Supplementary Corpora for Written Age0 106
    5.4.2.1 Lancaster Corpus of Children’s Project Writing (LCCPW) 106
      5.4.2.1.1 Introduction 107
      5.4.2.1.2 Extract Pure Corpus Texts from LCCPW 107
    5.4.2.2 LUCY Corpus 109
      5.4.2.2.1 Introduction 109
5.4.2.2 Extract Pure Corpus Texts from LUCY

5.5 Lemmatization of Supplementary Corpora in BNC XML Style
5.5.1 Tagged with CLAWS4 C5 Tagset
5.5.2 Lemma List for BNC XML
5.5.2.1 Modification of BNC Exception-Word-List
5.5.2.2 Modification of BNC Suffix-List
5.5.2.3 Modification of BNC Tag-List
5.5.3 Lemmatization Procedure

5.6 Completion of the Balanced Age-Group Corpus

5.7 Testing Corpus

5.8 Core Vocabulary Lists

5.9 Computational Approach to Core Vocabulary Design

5.10 Comparing Core Vocabulary Lists

5.11 Cumulative Coverage

5.12 Contributed Coverage

5.13 Dispersion Index

5.14 Core Vocabulary Proportion in Different Age Groups

5.15 Sum-up

6. CORE VOCABULARY IN DIFFERENT GENRES

6.1 Building of Genre Corpus

6.2 A Statistical Method for Distributed Word Frequency in Different Genres

6.3 Comparison of Core Vocabulary Lists

6.4 Sum-up

7. CORE VOCABULARY VS. AGE OF ACQUISITION, FAMILIARITY AND IMAGEABILITY

7.1 Core Index and Age-Genre Corpus

7.2 Correlation between Core Index and Familiarity
Abstract

Core vocabulary has been a long-time concern for lexicographers, language teachers and other linguists. Many researchers have realized the existence of core vocabulary in English studies and defined core vocabulary by a variety of criteria. Ogden’s (1930) concept of Basic English, the vocabulary control movement in the 1920s and 1930s, and the proposal of culture-free Nuclear English for international use (Quirk 1982; Stein 1979a) all boosted the research into core vocabulary. Past research tends to focus on word frequency in written texts rather than both of the written and spoken texts. Moreover, the factor of age has been ignored in previous studies, but modern corpora with the tags of age information have made this kind of research possible.

In the light of previous theories and studies, core vocabulary is now defined as those words not only with high total frequency (raw frequency) but also with a wide distribution across different age groups and genres in both spoken and written modes of large corpora. The core words are now being selected and evaluated by the following parameters: 1) distributed word frequency in different age groups, 2) distributed word frequency in different text genres, and 3) word familiarity, imageability and age of acquisition.

Making use of the age-group and genre information tagged in the British National Corpus XML Edition (BNC XML 2007), our research adopts Carroll’s Usage Coefficient (U) to calculate a word’s distributed frequency in different age groups and text genres for evaluating the core degree of words. The findings are as follows:
First, higher cumulative coverage can be achieved when we select core vocabulary by the combined parameters of a word’s dispersion index and distributed frequency across age groups rather than raw frequency alone.

Second, young age group under 15 rely more on core vocabulary than adults. For age group over 15 years old, core vocabulary occupies a stable ratio in their mental lexicon despite age increase. Each age group tends to acquire more core words selected on the basis of distributed-frequency-cum-age than those selected on the basis of raw frequency or raw-frequency-cum-range (e.g., Longman Defining Vocabulary, Oxford Defining Vocabulary, i.e. Oxford 3,000).

Third, a core vocabulary selected according to distributed frequency across genres achieves higher cumulative coverage than a core vocabulary selected by raw frequency alone.

Fourth, a word’s core index and its familiarity, imageability and age of acquisition are significantly correlated. There is a significant positive correlation between a word’s core index and its familiarity, while the correlation between a word’s core index and imageability and age of acquisition is significantly negative. That is, familiarity is a good indicator of core vocabulary and core vocabulary should be familiar to people of all age groups and different backgrounds. However, highly imageable words are not necessarily core words and core words tend to be lower in imageability. Moreover, core words tend to be acquired earlier in language acquisition and language development.

Fifth, by applying our research result to the compilation of the electronic English Frequency Dictionary for English Learners, the distributed frequency across different genres is effective for learners to discriminate English synonyms in terms of genre, language formality, cultural indication, register, and spoken and written domains. With distributed
frequency across genres displayed in a histogram, EFL learners can gain a better insight into the difference between English synonyms.

Our study is thus significant in bringing into play the importance of balanced vocabulary selection based on distributed frequency across age groups and genres. The findings of the present research will not only contribute to English teaching and learning but also lexicography studies. Moreover, our research into correlation between a word’s core index, familiarity, imageability and age of acquisition deepens our understanding of the psycholinguistic features of the mental lexicon.