LETTERED WORDS AND ROMAN LETTER CHARACTERS IN CHINESE WRITING: A STUDY OF ALPHABETIC WRITING IN CHINESE NEWSWIRES

DISSERTATION

Presented in Partial Fulfillment of the Requirements for

the Degree Doctor of Philosophy in the

Graduate School of the Ohio State University

By

Helena Riha, M.S., M.A.

* * * * *

The Ohio State University
2008

Dissertation Committee:

Professor Mary E. Beckman, Adviser

Professor Brian D. Joseph

Professor Marjorie K.M. Chan

Professor J. Marshall Unger

Approved by

Linguistics Graduate Program
ABSTRACT

As Chinese society continues to internationalize, Mandarin Chinese is exhibiting considerable influence from Chinese-English language contact. One effect of this contact is the increasing use of roman letters in Chinese writing. Whereas Chinese has traditionally been written almost exclusively with Chinese characters, roman letters are now commonly used as well.

The mixed use of Chinese and roman scripts presents an interesting problem from a linguistic standpoint. While English writing is morphophonemic, Chinese writing is morphosyllabic, creating a mismatch between the two writing systems with respect to the linguistic units they write. I explore the use of roman letters in Chinese writing to determine what linguistic units roman letters can write in Chinese and how the mismatch between the writing systems is resolved. I take the view that analogy is central to the comprehension of imported forms. The receiving group has to find “something understandable in terms of their own patterns” (Herskovits 1938) to accept foreign forms.

I examined the use of roman letters in the Chinese Gigaword Third Edition, analyzing letter strings in the corpus to determine what types of linguistic units roman letters write. My findings indicate that roman letters are used most frequently to write imported
abbreviations, as well as hybrid forms containing individual roman letters and Chinese characters. I conclude that Chinese speakers prefer to use roman letters as a new type of Chinese character that is usually pronounced with its letter name and that stands apart visually from the surrounding units by fitting into the equidimensional square of Chinese writing. The letter may also have an independent meaning as a morpheme.

I also measured the rate of growth in roman letter use in the People’s Republic of China (PRC) and Taiwan from 1991 to 2006. I found that roman letters are used more in Taiwan than in the PRC, which I suggest can be attributed to Taiwan’s longer period of exposure to English, as well as to political and social differences between the two societies. Spelled-out words are also more frequent in Taiwan, for the same reasons. Both Taiwan and the PRC show increasing use of roman letters over time, but the rate of increase in Taiwan is faster.

My study shows that alphabetic writing is adapted to the patterns of Chinese writing in both societies. As speakers gain familiarity with English, however, roman letters are also increasingly used in an English-like way to spell out words.
Dedicated to my family and to Stephen
ACKNOWLEDGMENTS

I would like to express my sincere appreciation to Mary Beckman, my adviser, for her support and guidance as I wrote my dissertation. I have benefited greatly from Prof. Beckman’s extensive knowledge of Asian languages and from her many years of experience in various subfields of linguistics. Prof. Beckman encouraged me to think through my research questions and methodology carefully, which I truly appreciate. I would also like to thank Brian Joseph for his insightful suggestions regarding my dissertation and for our many interesting discussions about language and linguistics. The courses I took with Prof. Joseph inspired me to take on this project. I am grateful to Marjorie Chan for her helpful advice about my dissertation and for the informative readings she introduced me to on lexical borrowing in Chinese. I benefited tremendously from the courses I took with Prof. Chan on Chinese linguistics, which laid the groundwork for my thesis and dissertation research. James Unger took the time to discuss critical issues with me regarding Asian writing systems and introduced me to many important readings. I appreciated Prof. Unger’s advice and comments very much as I designed the corpus study and analyzed my data.

I thank Don Winford for introducing me to the field of contact linguistics. I really enjoyed Prof. Winford’s courses on language contact and appreciated the opportunity to begin exploring issues that I went on to examine in my dissertation.
Thomas Bishop of the Department of Statistics offered valuable suggestions for designing the corpus study. I thank Prof. Bishop for taking the time to discuss methods of statistical analysis with me and for his willingness to answer my many questions.

I especially want to thank my colleague Kirk Baker. I am very grateful to Kirk for his assistance in conducting the corpus search and obtaining counts of lettered words. My study would have been conducted on a much smaller scale without his help.

I also want to thank Clint Roberts and Zhen Wang, my consultants at the OSU Statistical Consulting Service. Clint and Zhen helped me work through various issues to analyze the data in a meaningful way.

My colleagues Xiaofei Lu, Jianguo Li, Lei Xu, Na’im Tyson, and Andrew Plummer provided assistance at different stages of this project. Xiaofei, Lei, and Na’im helped me in conceptualizing the study, and Jianguo and Andy assisted me in the final stages of dissertation writing. I also thank my colleague Laurie Maynell for her friendship and advice as I planned and wrote my dissertation.

I thank the Department of Linguistics for generously providing travel funds that allowed me to attend professional conferences to present my dissertation work. I also benefited from the help of Jim Harmon and Jon Dehdari, who provided excellent technical support in using the Chinese Gigaword corpus. Thanks also go to Claudia Morettini, Jane Harper, and Joanna Anderson for the administrative support they kindly provided over the years.

Finally, I thank my family and my husband, Stephen, for their love and encouragement – and patience.
VITA

1991………………………………… B.A. in Chinese Language and Literature
Indiana University

1994………………………………… M.S. in Foreign Service
Georgetown University

1999-2001………………………… Graduate Teaching Associate, Department of
East Asian Languages and Literatures
The Ohio State University

2001-2002………………………… Fulbright Fellow, Beijing Broadcasting
Institute, Beijing, China

2003-2005………………………… Graduate Teaching Associate, Department of
Linguistics
The Ohio State University

2006………………………………… M.A. in Chinese Linguistics
The Ohio State University

2005-2008………………………… Editorial Assistant, Language

FIELDS OF STUDY

Major Field: Linguistics
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Dedication</th>
<th>iv</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vii</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiv</td>
</tr>
<tr>
<td><strong>Chapters:</strong></td>
<td></td>
</tr>
<tr>
<td>1. What are ‘lettered words’?</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1.2 The special nature of roman letter characters</td>
<td>3</td>
</tr>
<tr>
<td>1.3 Comparison with orthographic borrowing in English</td>
<td>4</td>
</tr>
<tr>
<td>1.4 The significance of lettered words</td>
<td>5</td>
</tr>
<tr>
<td>1.5 Lettered words in Chinese linguistics</td>
<td>6</td>
</tr>
<tr>
<td>1.6 Research questions</td>
<td>7</td>
</tr>
<tr>
<td>1.7 Corpus study</td>
<td>8</td>
</tr>
<tr>
<td>1.8 Understanding native speaker intuitions about lettered words</td>
<td>9</td>
</tr>
<tr>
<td>1.9 Organization of the dissertation</td>
<td>9</td>
</tr>
<tr>
<td>2. Structure of the Mandarin lexicon</td>
<td>11</td>
</tr>
<tr>
<td>2.1 Nomenclature</td>
<td>11</td>
</tr>
<tr>
<td>2.2 Mandarin lexicon or Chinese lexicon?</td>
<td>12</td>
</tr>
<tr>
<td>2.3 Mandarin syllables</td>
<td>14</td>
</tr>
<tr>
<td>2.3.1 The importance of syllables in Mandarin</td>
<td>14</td>
</tr>
<tr>
<td>2.3.2 Mandarin syllable structure</td>
<td>15</td>
</tr>
<tr>
<td>2.4 Morphemes in Mandarin</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Zi in Chinese</td>
<td>21</td>
</tr>
<tr>
<td>2.6 Defining word in Chinese</td>
<td>22</td>
</tr>
</tbody>
</table>
5.1.1 Selecting a corpus ................................................. 92
5.1.2 Focus on Standard Mandarin ................................. 93
5.2 Corpus .................................................................. 94
5.2.1 Chinese Gigaword Third Edition ............................ 94
5.2.2 Data file markup .................................................. 95
5.2.3 Time frame and data quantity ................................. 97
5.3 Research questions and hypotheses ........................... 97
5.3.1 Introduction ......................................................... 97
5.3.2 Use of roman letter strings in the PRC vs. Taiwan ...... 98
5.3.3 Use of roman letter strings over time ........................ 99
5.3.4 Characteristics of roman letter strings ..................... 100
5.3.5 Case combinations used in writing RL strings .......... 101
5.3.6 Characteristics of hybrid strings ............................. 101
5.3.7 Contextual support in newswires ............................ 103
5.4 Finding lettered words ............................................ 104
5.5 Sampling procedures .............................................. 106
5.5.1 Sample for 1991, 1998, 2005 ................................. 106
5.5.2 RL strings in XIN and CNA from 1991 to 2006 ......... 108
5.5.3 Using whole articles ............................................. 110
5.6 Variables that were coded ....................................... 111
5.7 Sample coded data .................................................. 117
5.8 Data analysis .......................................................... 118

6. Corpus study: Results .............................................. 119
6.1 Introduction ........................................................... 119
6.2 Increase in use of lettered strings from 1991 to 2006 .... 119
6.3 Prevalence of roman letter zi strings ........................ 122
6.4 Comparison of fully spelled strings in XIN and CAN ...... 124
6.5 Most RL zi strings are initialisms ......................... 125
6.6 Initialisms vs. spelled words in CNA and XIN .......... 125
6.6.1 Comparing initialisms and spelled words ............... 125
6.6.2 Initialisms and spelled words in CNA ................... 127
6.6.3 Initialisms and spelled words in XIN ................. 128
6.7 Proportion of initialisms in CNA and XIN ................. 129
6.8 Length of roman letter strings ................................ 132
6.8.1 FS strings compared with RL zi strings ............... 132
6.8.2 Preference for “short” RL strings .......................... 135
6.9 Preferred case of roman letters ................................. 137
6.9.1 Upper case predominates .................................. 137
6.9.2 Upper case writes RL zi strings .......................... 138
6.9.3 Upper case writes initialisms .............................. 139
6.10 Characterization of [M N] nominals ......................... 139
6.10.1 Roman letter modifiers are preferred ................. 139
6.10.2 Modifiers are composed of RL zi ...................... 141
6.10.3 Heads are composed of RL zì ........................................... 142
6.11 Prevalence of RL zì in hybrid strings ........................................ 143
6.12 Contextual support for initialisms .......................................... 144
   6.12.1 Translations for initialisms ......................................... 144
   6.12.2 Transcriptions for initialisms ..................................... 145
6.13 Contextual support for fully spelled strings ................................ 145
   6.13.1 Translations for fully spelled strings ............................ 145
   6.13.2 Transcriptions for fully spelled strings .......................... 146
6.14 Discussion ........................................................................... 147
   6.14.1 Growth in the use of roman letter strings in CAN ............. 147
   6.14.2 Growth in the use of roman letter strings in XIN ............... 150
   6.14.3 Numerous RL zì strings, fewer fully spelled strings ...... 152
   6.14.4 Initialisms vs. spelled words in CNA and XIN ............... 155
   6.14.5 Short length of roman letter strings .............................. 158
   6.14.6 Why upper case is preferred ...................................... 159
   6.14.7 Preferred modifiers and heads ..................................... 161
   6.14.8 Translations and transcriptions for RL strings ............. 162

7. Concluding remarks .................................................................. 164
   7.1 Summary of findings ..................................................... 164
   7.2 Lessons from the corpus study ......................................... 167
   7.3 Issues for future research ............................................... 169

Appendix A: Samples of XIN and CNA newswires ......................... 172
Appendix B: Sample of a coded XIN newswire ............................... 178
Appendix C: Sample of a coded CNA newswire ............................ 184
Appendix D: Types of roman letter strings in the XIN and CNA samples 189
Bibliography ............................................................................. 204
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The Initials of Mandarin</td>
<td>18</td>
</tr>
<tr>
<td>2.2</td>
<td>The Finals of Mandarin</td>
<td>18</td>
</tr>
<tr>
<td>3.1</td>
<td>Semantic-phonetic Compounds</td>
<td>37</td>
</tr>
<tr>
<td>4.1</td>
<td>Writing Foreign Names in English</td>
<td>57</td>
</tr>
<tr>
<td>4.2</td>
<td>Examples of Simultaneous Phonetic and Semantic Adaptation</td>
<td>60</td>
</tr>
<tr>
<td>4.3</td>
<td>Examples of Phonetic Adaptations with Explicative Adjuncts</td>
<td>62</td>
</tr>
<tr>
<td>4.4</td>
<td>Separate Application of Adaptation Processes in Lettered Words</td>
<td>68</td>
</tr>
<tr>
<td>4.5</td>
<td>Lettered Words Containing Explicative Adjuncts</td>
<td>70</td>
</tr>
<tr>
<td>4.6</td>
<td>Roman Letter Native Creations</td>
<td>71</td>
</tr>
<tr>
<td>4.7</td>
<td>Single Letter Units in English</td>
<td>79</td>
</tr>
<tr>
<td>4.8</td>
<td>Pronunciation of Letter Names in Mandarin</td>
<td>86</td>
</tr>
<tr>
<td>5</td>
<td>Summary of the Sample data for the Corpus Study</td>
<td>108</td>
</tr>
<tr>
<td>6.1</td>
<td>Proportion of RL zì Strings and FS Strings in XIN and CNA</td>
<td>123</td>
</tr>
<tr>
<td>6.3</td>
<td>Summary Data for FS and RL zì Strings</td>
<td>133</td>
</tr>
<tr>
<td>6.4</td>
<td>Length of RL Strings and Counts in XIN and CNA</td>
<td>136</td>
</tr>
<tr>
<td>Section</td>
<td>Title</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>----------------------------------------------------------------------</td>
<td>------</td>
</tr>
<tr>
<td>6.5</td>
<td>Proportion of UC, MC, and LC in XIN and CNA</td>
<td>137</td>
</tr>
<tr>
<td>6.6</td>
<td>Case Combinations and String Types in XIN and CNA</td>
<td>138</td>
</tr>
<tr>
<td>6.7</td>
<td>Case Combinations Used to Write Initialisms in XIN and CNA</td>
<td>139</td>
</tr>
<tr>
<td>6.8</td>
<td>Hybrid String Types in XIN and CNA</td>
<td>140</td>
</tr>
<tr>
<td>6.9</td>
<td>RL Modifiers and Heads in [M N] Nominals in XIN and CNA</td>
<td>140</td>
</tr>
<tr>
<td>6.10</td>
<td>Roman Letter Modifier Types in XIN and CNA</td>
<td>141</td>
</tr>
<tr>
<td>6.11</td>
<td>Roman letter heads in [M N] nominals in XIN and CNA</td>
<td>142</td>
</tr>
<tr>
<td>6.12</td>
<td>RL Zi and FS Units in Hybrid Strings in XIN and CNA</td>
<td>143</td>
</tr>
<tr>
<td>6.13</td>
<td>Translations for Initialisms in XIN and CNA</td>
<td>144</td>
</tr>
<tr>
<td>6.14</td>
<td>Translations for Fully Spelled Strings in XIN and CNA</td>
<td>145</td>
</tr>
<tr>
<td>6.15</td>
<td>Transcriptions for Fully Spelled Strings in XIN and CNA</td>
<td>146</td>
</tr>
<tr>
<td>Figure</td>
<td>Description</td>
<td>Page</td>
</tr>
<tr>
<td>--------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>6.1</td>
<td>Rate of Increase in the Use of Roman Letter Strings in CNA and XIN</td>
<td>120</td>
</tr>
<tr>
<td>6.2</td>
<td>Probable Initialisms vs. Spelled Words in CNA and XIN</td>
<td>126</td>
</tr>
<tr>
<td>6.3</td>
<td>Proportion of Probable Initialisms in CNA and XIN 1991–2005</td>
<td>130</td>
</tr>
<tr>
<td>6.4</td>
<td>Length of Fully Spelled and RL Zi Strings</td>
<td>133</td>
</tr>
<tr>
<td>6.5</td>
<td>Length of Roman Letter Strings in XIN</td>
<td>135</td>
</tr>
<tr>
<td>6.6</td>
<td>Length of Roman Letter Strings in CNA</td>
<td>136</td>
</tr>
</tbody>
</table>
CHAPTER 1

WHAT ARE ‘LETTERED WORDS’?

1.1 Introduction

This is a study of lettered words in contemporary Putonghua and Guoyu, the two standard varieties of Mandarin Chinese that are the national languages of the People’s Republic of China (PRC) and Taiwan. Lettered words, or roman letter words, are words in any variety of Chinese written partly or fully with roman letters. Examples include initialisms (e.g. WTO), spelled-out words (e.g. internet), and hybrid words with both roman letter and Chinese character components (e.g. X 光 ‘X-ray’). Many lettered words are introduced into mainstream use through print rather than speech (Zhang 2005). The printed form may serve as Mandarin speakers’ primary exposure to many lettered words, while the spoken form may be infrequent. Individuals may see and read a variety of lettered words much more often than they say or hear them.

The fact that lettered words are in certain ways born of the written language and are created by literate speakers makes them unusual. The core definition of ‘word’ in linguistics is a spoken form that may or may not have a corresponding written form, and whose written form, if one exists, is secondary to the spoken form. A meaningful analysis
of lettered words requires a discussion of both their spoken form and meaning as well as their corresponding written form and how it may impact the spoken form and meaning. The written form of lettered words is therefore an integral part of the investigation of lettered words in this study.

A limited number of lettered words in Mandarin are spelled out words of English such as <buckeye>. Most are abbreviations such as <OSU>. Many spelled words are *hapax legomena*, or items that appear only once in a corpus. Newspaper stories include fully-spelled proper names written in roman letters, but those names are not necessarily used in other written or spoken contexts. A smaller number of spelled words are borrowings, mostly from English, such as *fax* and *e-mail*.

Most lettered words are words in which the individual roman letters can be pronounced as discrete phonological units. Usually they are pronounced with their respective letter names, whose pronunciation is familiar to educated speakers of Mandarin. (The pronunciation of lettered words is discussed in Chapter 4.) Initialisms such as *WHO* and *MBA* also fall into this category because they are pronounced letter by letter. What English speakers consider to be acronyms, such as *SARS* and *AIDS*, also fall into this group because the roman letters can be pronounced individually if speakers are unaware of the source-language pronunciation or if for other reasons give the word a letter-by-letter pronunciation. Most initialisms and acronyms in English are written in all capital letters, which provide a clue that the word is an initialism or an acronym rather than an ordinary word of English.
1.2 The special nature of roman letter characters

Lettered words are unusual for at least two reasons. First, their written form reflects aspects of their morphology and expression of meaning. Although lettered words are certainly used in speech as well as in writing, the actual graphs used to write them also contribute to understanding their meaning. Although any speaker can use lettered words in speech, an illiterate speaker would have no way of understanding how their role as graphs in the writing system contributes to their meaning. This is best explained with an example. (Although the following example from Mandarin is not a lettered word, it is an example of how literacy interacts with aspects of how speakers can understand words.)

When using 十字路口 shí zì lùkòu ‘intersection’ lit. ‘ten character street-mouth’, illiterate Mandarin speakers would understand the expression in speech, but they would not necessarily know that the morpheme shí ‘ten’, written as 十 and resembling an intersection, is used to form the compound since they does not know the compound’s written form. In other words, the speakers would know the morpheme’s spoken form but not why that morpheme is used in the expression meaning ‘intersection’. English speakers use words such as T-shirt and A-frame that have similarly iconic bases. While illiterate speakers understand these words in speech, they presumably do not know what T and A have to do with their meaning. To take another example, English speakers use the euphemism the F-word and other words of this type to avoid saying taboo words. Other speakers invariably understand which words are being avoided, including illiterate speakers. However, it is conceivable that illiterate speakers would not understand why the euphemisms are formed as they are since they are unfamiliar with the written form of the
letters of the alphabet and the spellings of taboo words. Similarly, if someone referred to the “ph-word,” literate speakers would know that it could not possibly be the word ‘f___’ because of the spelling.

Second, lettered words depart from earlier forms of borrowing and coinage in Chinese that used only Chinese morpheme-syllables, all of which are written with Chinese characters. Two productive traditional borrowing processes are loan translation (e.g. 热狗 régōu ‘hot dog’, lit. ‘hot-dog’), in which Chinese morphemes are used to translate the meaning of the morphemes in the original, and (rebus-based) transcription (e.g. 沙发 shāfā ‘sofa’, using the graphs for the two morphemes meaning ‘sand’ and ‘disperse’), in which Chinese characters are used only for their phonetic value and the individual meanings of the morphemes that they otherwise write are suppressed. New terms in Chinese are coined by creating compounds with Chinese morphemes. The written form is then a straightforward stringing together of the characters used to write the component morphemes. Since the entry of roman letter characters into Chinese, however, roman letter units replace some or all of the Chinese morphemes and their associated characters in borrowings and coinages.

1.3 Comparison with orthographic borrowing in English

Although lettered words are used in spoken Mandarin, a wider variety is used in writing, especially in the Chinese print media (Wang 2004, Zhang 2005) and in computer-mediated communication (Gao 2007). To compare this phenomenon to a similar situation of orthographic borrowing in American society, one can think of the
familiar example of using Greek letters, rather than their romanized forms, to write the names of fraternities, sororities, and honor societies at American universities. Americans expect to see Greek letters on the outside of a fraternity house rather than their romanized forms. A more unusual example of orthographic borrowing, and one that more accurately mirrors the situation in Chinese society with respect to lettered words, is the following hypothetical example: imagine a language contact situation in which English-speaking societies borrow heavily from Chinese, and in more ‘sinicized’ societies, there happened to be the widespread use of Chinese characters to write Chinese borrowings in English, such as 鍋 in place of <wok> and 豆腐 in place of <tofu>, and to write neologisms in English, such as 豆腐burger in place of <tofuburger> and 豆腐rkey in place of <tofurkey>. Chinese characters would be used in writing these words on product packages, menus, and elsewhere. Americans would need to learn to write the Chinese characters used in these words and to read and say them aloud. It may be difficult to imagine such a situation occurring in American society, but it helps us understand the striking visual and linguistic impact that roman letters and lettered words are having on the written Chinese language.

1.4 Significance of lettered words

The use of roman letters in Chinese involves new issues regarding the role of the script in morphology, lexical semantics, and lexical borrowing that have yet to be explored fully. First, and most significantly, the script reflects aspects of the formation of lettered words. It provides indications of the linguistic process of forming these words
rather than merely transcribing their pronunciation in writing once they have been created. The script itself reveals characteristics of the word-formation process in ways that, to my knowledge, have yet to be articulated systematically in the study of Chinese words. Second, lettered words use a foreign script in all or part of their written form rather than native Chinese characters. This represents an aspect of their form that cannot be accounted for in traditional analyses of lexical contact (e.g. Haugen 1950, Weinreich 1953), whose main concern is the spoken form of borrowings and native creations rather than their adaptation in writing. Finally, certain lettered words reflect the resolution of a mismatch in the English and Chinese writing systems and the linguistic units represented in each. Whereas English writing is morphophonemic, Chinese writing is morphosyllabic, such that the individual graphs and combinations of graphs in the two writing systems represent different linguistic units. The most common types of lettered words resolve these incompatibilities in innovative ways that I explore in this study.

1.5 Lettered words in Chinese linguistics

Although lettered words have become popular in Mandarin, they are just beginning to receive the attention of Chinese linguists in the United States. Descriptions of the morphology of Mandarin and borrowing in Mandarin either do not mention lettered words at all (e.g. Lü 1963, Lu 1964, Li & Thomson 1981, Duanmu 2000, Packard 2000, Hsu 2002) or discuss them only briefly (Chen 1999, Sun & Jiang 2000, Lin 2001). Riha (2006a) was a preliminary study of the morphology and semantics of lettered words. Studies of new words in Mandarin (e.g. Sun & Jiang 2000, Hu & Hsu 2003, Li 2004,
Wang 2004, Zhang 2005, Gao 2007) acknowledge the existence of lettered words in Mandarin and identify different types of lettered words and processes for their formation, but they do not systematically examine the question of how these words compare with other native words and older borrowings in Chinese or how roman letter graphs are incorporated into Chinese writing.

1.6 Research questions

In this study I investigate four aspects of lettered words that pertain to their formation and use in a Chinese discourse context. First is the nature of the Chinese lexicon and the Chinese writing system and how some types of lettered words reflect the characteristics of both. I suggest that certain types of lettered words can be seen as Chinese in both their morphological structure and in the way they are written. They have characteristics that are analogous to those of Chinese morpheme-syllable-characters (zì) and their combinations.

Second is the status of lettered words as products of language contact. Since foreign linguistic units, morphological processes, and orthographic units are either used in the formation of lettered words or serve as models, I examine them as products of the mixture of Chinese and English linguistic processes and orthographies.

Third are sociolinguistic factors contributing to the formation of different types of lettered words, which help to explain why certain kinds of lettered words may be more common in one Mandarin-speaking society than another. Both the PRC and Taiwan use Mandarin as their national language but have different social and political systems,
cultures, and histories of language use. I describe similarities and differences in the types of lettered words used in the PRC and in Taiwan and propose motivations for their patterns of lettered word use based on sociopolitical factors.

Finally, I consider one communicative context in which lettered words commonly occur – Chinese newswires. I identify lettered words in newswires from the PRC and Taiwan and examine translations and transcriptions, two types of contextual support frequently provided alongside lettered words in newswires. I examine patterns in the use of this additional contextual information.

I investigate my research questions by way of a corpus study lettered words in Chinese newswires.

1.7 Corpus study

I examine a large corpus of Chinese newswires to investigate the following aspects of lettered words: 1. the nature of lettered words as entities with Chinese linguistic and orthographic characteristics, 2. their status as products of language contact, 3. sociopolitical factors that shape the types of lettered words used in the PRC and in Taiwan, and 4. the communicative context of the newswires in which they occur. I used the LDC Chinese Gigaword Corpus Third Edition, which includes newswires from both the PRC and Taiwan, for the corpus study. The corpus contains news stories from the PRC’s Xinhua News Agency and Taiwan’s Central News Agency. I examined newswires in the sixteen-year period from 1991, the first year of the corpus, to 2006, the last year of the corpus.
1.8 Understanding speaker intuitions about lettered words

In this dissertation I do my best to represent what I believe are Chinese speakers’ intuitions about lettered words. It is difficult to know the internal cognitive representations of a given speaker without evidence from psycholinguistic experiments, however, which to my knowledge have yet to be conducted regarding the processing and production of lettered words in Chinese speech or writing. To truly understand how Chinese speakers comprehend and create lettered words, one would need to consider at least the following few factors. First, since lettered words are used frequently in writing, comprehension and production in both speech and writing would have to be investigated. Second, speakers’ familiarity with English would have to be ascertained, and the effects of this familiarity, or the lack of it, would have to be considered. Third, the use of roman letters to write English-based units would need to be differentiated from their use in writing romanized Chinese units. Finally, the effects of literacy in Chinese (in Chinese characters and in pīnyīn romanization) and in English would have to be examined carefully. Since I do not have individual speaker information regarding these factors, I make educated guesses about speaker intuitions based on patterns in the lettered word data I have collected and on information from previous research on lettered words.

1.9 Organization of this dissertation

The rest of this dissertation is organized as follows. Chapter 2 discusses the structure of the Mandarin lexicon. Chapter 3 describes the Chinese writing system. Chapter 4 explains traditional forms of lexical borrowing and coinage in Mandarin using Chinese zì
and innovative forms of borrowing and coinage that use roman letter characters in place of Chinese zi. Chapter 5 describes the methodology used for the corpus study of roman letter strings in PRC and Taiwan newswires. Chapter 6 presents the results of the corpus study. Chapter 7 summarizes the findings of the dissertation and discusses issues for future research.
CHAPTER 2

STRUCTURE OF THE MANDARIN LEXICON

2.1 Nomenclature

The terms referring to present-day Mandarin Chinese are controversial and complex, so I begin this chapter by discussing the different names used for Mandarin Chinese. The formation of a standard language in Chinese society has been a difficult process that gave rise to numerous names for the standard language, each of which has historical and sociolinguistic connotations. The following are some of the names that are used to describe the standard language and that appear in this dissertation: Xiàndài Hânyǔ ‘Modern Chinese’, used primarily in the PRC, Pùtōnghuà ‘Common Speech’, used in the PRC to refer to the spoken standard language, and Guóyǔ ‘National Language’, used in Taiwan to refer to the spoken standard language.

English names for the standard language are varied as well. The term Mandarin refers to the largest variety of Chinese, the one to which the Beijing regional variety belongs. Mandarin is a translation of guānhuà, the language of officials, or mandarins, employed in imperial China. Standard Mandarin is derived from and loosely based on the Beijing variety of Mandarin but excludes its idiosyncratic features. The terms Standard Mandarin, Mandarin, and Mandarin Chinese are usually avoided by Chinese language
authorities and many Chinese scholars writing in English, however. Rather, the terms
Chinese, Modern Chinese, and Standard Modern Chinese are preferred by Chinese
language authorities and many Chinese scholars. Norman (1988) suggests that the term
Mandarin is avoided because of its connection with China’s imperial past. For the sake of
convenience, in this dissertation I use the term Mandarin to mean Standard Mandarin,
and I also use it as a cover term for the Chinese terms Xiàndài Hânyǔ, Pùtōnghuà, and
Guóyǔ. In cases where the PRC and Taiwan varieties of the standard language need to be
differentiated, I use Pùtōnghuà (Putonghua) and Guóyǔ (Guoyu). I use the term Chinese
to refer to all the varieties of Chinese as a group.

2.2 Mandarin lexicon or Chinese lexicon?

In the following sections I describe the Mandarin lexicon so that I can show the role of
lettered words in it in the next chapter. The general descriptive properties I outline here
actually apply to the Chinese lexicon as a whole, however, rather than just to Mandarin.
The reason for this is that the seven varieties of Chinese – Mandarin, Wu, Xiang, Gan,
Hakka, Min, and Yue – are customarily differentiated on the basis of their phonologies
rather than their morphology or syntax (Li 1937), which are considered to be similar.

A further point is that it is necessary for me to speak of a Chinese writing system
rather than a Mandarin writing system. Aside from vernacular characters used only in
particular varieties of Chinese, standard characters represent morphemes that can be
pronounced and understood in any variety of Chinese. If one writes a string of characters
in Mandarin, each of them can be pronounced and understood by literate speakers of
most varieties of Chinese. Whether the string is understood as a meaningful word, phrase, or sentence is not always certain, however. There are differences among the varieties at the level of syntax that may make sentences written in Mandarin, the variety used in writing the standard language, awkward for speakers of other varieties to understand. Interestingly, this is traditionally not a matter of concern for Chinese scholars. The ‘Great Tradition’ of unifying speakers of different Chinese varieties through the use of ‘logographic’ Chinese characters overlooks differences among the varieties in syntax. Y. R. Chao, for example, considered the Chinese varieties to be homogenous in terms of their syntactic characteristics, stating that there is “practically one universal Chinese grammar” (1968a:13) with only minor differences and close equivalents across the varieties.

Saillard (2004:166) counters Chao’s traditional view, pointing out that “‘modern written Chinese’ is often considered universal at the meta-topolectal level, despite the fact that there still exists a writing tradition in some Chinese topolects, reflecting each topolect’s specific syntactic structures and lexical peculiarity”. In her observations of Putonghua classes at a primary school in Wenzhou, Zhejiang Province, where the Wu variety of Chinese is spoken, Saillard noticed that the teaching of Putonghua in China’s educational system considers only the phonetic needs of speakers of non-Mandarin varieties rather than the teaching of syntactic structures, leading to difficulties in achieving fluency rapidly in Putonghua. The assumption reflected in Chinese schools’ pedagogical approach is that the Great Tradition requires literate speakers of Chinese to overcome differences in the grammar of their variety and the standard language to
participate in the highly valued cultural practice of being able to read and understand Chinese characters. Cheung (1992:216) states this expectation concisely in saying that “the Chinese script is a system understood by speakers of all regional dialects, many of which are mutually unintelligible. The written Chinese language, having been deprived of its phonic elements, is an approximation towards a written language system accessible to all dialect users.” While this may be true, speakers of non-Mandarin varieties must make adjustments with respect to syntax to access the standard written language.

2.3 Mandarin syllables

2.3.1 The importance of syllables in Mandarin

To describe the Mandarin lexicon, it is necessary to begin with the syllable, the phonological building block of linguistic units in the Mandarin lexicon, including lettered words, as I show in Chapter 3. Kratochvil (1968:23) suggests that the position of the syllable in all varieties of Chinese is more important than one might expect on the basis of experience with European languages. Indeed, syllables are the “basic natural unit” in modern Mandarin. Kratochvil comments (but does not document) that if asked to divide an utterance into smaller segments, educated Mandarin speakers will divide the utterance into syllables, whereas speakers of European languages will divide it into larger units composed of more than one syllable, usually words.

I would add that speakers of European languages may also be more aware of segments than speakers of Mandarin as well. Studies of English and Mandarin reading acquisition, for example, show that while awareness of phonemes is a strong predictor of subsequent
reading ability in English, such awareness is not required for the development of reading ability in Mandarin (Lin and Akamatsu 1997), suggesting that awareness of phonemes may not be as important for Mandarin speakers as for English speakers. Li, Anderson, and colleagues (2002) state that phonological awareness in Mandarin consists of a knowledge of the onset, rhyme, syllable, and tone, rather than consisting of awareness of phonemes as in English. And Bertelson et al. (1997) found, for example, that the zhùyīn phonetic symbols taught in Taiwan predisposed subjects to segmentation based on syllable initials and rhymes rather than phonemes. Although the teaching of reading in the PRC through the use of official pīnyīn romanization increases phonemic awareness (Read et al. 1986), the pedagogical focus still emphasizes onsets, rhymes, syllables, and tones, as I discuss below.

In attempting to answer why the syllable is the critical phonemic unit in Mandarin, Kratochvíl describes it as a “kind of crosspoint where different levels of MSC [Modern Standard Chinese] grammar intersect: the overwhelming majority of morphemes in MSC are monosyllabic and most MSC morphemes occur as basic syntactic units (or words)” (1968:24). This description fails to mention the fact that Modern Mandarin contains numerous polysyllabic words. Even in these words, however, literate speakers are aware of the meanings of the morphemes corresponding to the individual syllables (Hoosain 1992), a phenomenon that may be reinforced by the Chinese writing system, in which each morpheme-syllable is written with a separate character.

2.3.2 Mandarin syllable structure
Sequences of consonants and vowels in Mandarin syllables can be described with the formula (C1)V(C2), proposed by Kratochvíl (1968), where V denotes an optional medial and the obligatory vowel nucleus (a simple vowel, diphthong, or triphthong), and the parentheses denote an optional initial consonant C1, and an optional final consonant C2. Four types of segmental combinations occur in Mandarin syllables: C1VC2, C1V, VC2, and V.

Tone, a relative and contrastive pitch pattern occurring in syllables, is also an obligatory part of Mandarin syllable structure, with each syllable having a distinctive tone. Tones differentiate syllables that are identical in segmental structure and form an integral component of morphemes. Mandarin has four tones: high level (e.g. yī ‘clothes’), high rising (yí ‘to suspect’), dipping/falling-rising (yí ‘chair’), and high falling (yì ‘meaning’) (Li and Thompson 1981:8). There is also a “neutral” tone, that is, an absence of tone, on a weakly stressed syllable that follows a syllable with a full tone (e.g. bàba ‘father’, the reduplicated form of bà ‘father’). Since neutral-tone syllables cannot be pronounced in isolation, the neutral tone is generally not considered to be one of the contrastive tones. When a syllable-morpheme that normally has weak stress is cited in isolation, it is said with its citation tone. For example, to say the second ba in bàba in isolation, one would pronounce it as bà. In cases where the citation tone is a neutral tone (e.g. for grammatical particles such as le or ma), one of the four tones must be used, and the high tone is customarily chosen in such cases (Norman 1988).

Collapsing across tones, Mandarin has just 400 legal segmental shapes for syllables, which when fully specified for tone comprise only about 1,250 different syllable shapes.
It is important to point out how these syllables are normally conceptualized in Chinese society, an issue that highlights once again the problem of phonemic awareness. Rather than being described as a series of phonemes, Mandarin syllables, and Chinese syllables in general, are traditionally described as combinations of *initials,* or initial consonants, *finals,* or rhymes containing the medial, main vowel, and final consonant, and tone. Outside the context of linguistic studies, the individual phonemes that make up Mandarin syllables normally do not receive as much attention as the initials and finals. This may provide further support for the view that syllables are the more salient phonological unit for Mandarin speakers rather than phonemes.

Mandarin has twenty-one initials and thirty-five finals that combine to form legal syllables. Mandarin dictionaries and many linguistic studies of Mandarin phonology present information on the initials and finals by providing a table showing the initials, followed by a table showing the finals. This presentation is given in Tables 2.1 and 2.2 using *pīnyīn* romanization. IPA transcriptions of the phonemes represented by the romanized forms are also given.
labial | b [p] | p [pʰ] | m [m] | f [f]  
alveolar | d [t] | t [tʰ] | n [n] | l [l]  

| dental sibilants | z [ts] | c [tsʰ] | s [s] |  
| palatal | j [tʃ] | q [tʃʰ] | x [ç] |  
| velar | g [k] | k [kʰ] | h [x] |

<table>
<thead>
<tr>
<th>kāikōuhū ‘open mouth’</th>
<th>qíchihū ‘even teeth’</th>
<th>hékōuhū ‘closed mouth’</th>
<th>cuōkōuhū ‘puckered mouth’</th>
</tr>
</thead>
<tbody>
<tr>
<td>i [i]</td>
<td>u [u]</td>
<td>ü [y]</td>
<td></td>
</tr>
<tr>
<td>a [a]</td>
<td>ia [ia]</td>
<td>ua [ua]</td>
<td></td>
</tr>
<tr>
<td>o [o]</td>
<td>uo [ua]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>e [ə]</td>
<td>ie [iɛ]</td>
<td></td>
<td>üe [yɛ]</td>
</tr>
<tr>
<td>ai [ai]</td>
<td></td>
<td>uai [uai]</td>
<td></td>
</tr>
<tr>
<td>ei [ei]</td>
<td></td>
<td></td>
<td>uei [uei]</td>
</tr>
<tr>
<td>ao [au]</td>
<td>iao [iau]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ou [ɔu]</td>
<td>iou [iɔu]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>an [an]</td>
<td>ian [ian]</td>
<td>uan [uan]</td>
<td>üan [yan]</td>
</tr>
<tr>
<td>ang [əŋ]</td>
<td>iang [ian]</td>
<td>uang [uan]</td>
<td></td>
</tr>
<tr>
<td>eng [əŋ]</td>
<td>ing [in]</td>
<td>ueng [uəŋ]</td>
<td></td>
</tr>
<tr>
<td>ong [ɔŋ]</td>
<td>iong [yŋ]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Tables 2.1 and 2.2 show the initial and final combinations that can be formed with Kratochvil’s (1968) formula representing Mandarin syllable structure, (C1)V(C2). The initials are placed in the C1 slot, and the finals in the V and C2 slots. Syllables that do not contain an initial consonant are called ‘zero-initial’ syllables. The finals are traditionally
divided into four groups, kāikōuhū ‘open mouth’, qìchīhū ‘even teeth’, hékōuhū ‘closed mouth’, and cuōkōuhū ‘puckered mouth’. Open mouth finals lack a medial, even teeth finals have the medial ɨ [i], closed mouth finals have the medial u [u], and puckered mouth finals have the medial ü [y] (Lin 2001:31). The letter ɨ used in pīnyīn represents three different sounds: a weak syllabic [z] in sibilants (zi, ci, si), a weak syllabic retroflex continuant [ʈ], in retroflexes (zhi, chi, shi), and a high front unrounded [i] after all other initials (Norman 1988:142). The consonant in “C2” can only be n or ng. A final retroflex r [ʐ] representing the er suffix also occurs, as in 跑儿 pǎor ‘to run’, but since the suffix is added secondarily to a syllable and does not form a syllable of its own, it is not listed in the finals table. Only certain combinations of initials and finals form phonotactically legal syllables in Mandarin. Possible combinations are described in Kratochvíl (1968), Li and Thompson (1981), Norman (1988), and Duanmu (2007), among others.

2.4 Morphemes in Mandarin

In both classical and modern Chinese, most morphemes are one syllable in length, and most syllables can be analyzed as independent morphemes. Chinese characters also represent single syllables and single morphemes. The difference between classical and modern Chinese is that polysyllabic words might have been more rare in the classical language than in modern Chinese. This is illustrated in example 2.1, adapted from Duanmu (2007:3), in which a quotation from Confucius is shown in the original classical Chinese (a) and in a modern Mandarin translation (b). In the example Q indicates a
question marker and the hyphen in ‘far-off place’ indicates that the combination may be a compound. The word ‘far-off place’ and other polysyllabic words are marked with square brackets.

2.1

a. 有 朋 自 [远方] 来 不 亦 乐 乎
   have friend from [far-place] come not also joy
   ‘If you have friends coming from far away, isn’t it also a joy?’

   [if-result] have [friend-friend] from [far-place] come
   不 是 也 很 [快乐] 吗
   not is also very [happy-happy]
   ‘If you have friends coming from far away, isn’t it also a joy?’

Example 2.2 shows that the original sentence in classical Chinese in (a) had just one possible polysyllabic word, 方远, while the version in modern Mandarin in (b) has four. As shown by the morpheme-by-morpheme gloss in (b), each of the polysyllabic words can be analyzed as combinations of individual morpheme-syllables. Modern Mandarin contains thousands of disyllabic and trisyllabic words such as these. Polysyllabic morphemes also exist in Mandarin, such as 玫瑰 méiguì ‘rose’, 柠檬 níngméng ‘lemon’, 蟑螂 zhāngláng ‘cockroach’ and others, most of which are borrowings, but Norman (1988:154) suggests that there are not many more of them in the modern language than in classical Chinese.

Judging from the appendix of new words in the 2005 edition of the PRC’s authoritative dictionary of the standard language, Xiàndài Hányǔ Cídiǎn ‘Dictionary of Modern Chinese’, the vast majority of new words in Mandarin are polysyllabic as well.
This is also the conclusion obtained in a study of new words in Mandarin by Sawer (1995), who found that less than one percent of new words are monosyllabic. Most new words can be analyzed as strings of monosyllabic morphemes like the polysyllabic words in 2.2, and they are formed with existing morphemes using established word formation processes. The compound 网吧 wângbā ‘internet bar’, for example, is a new word in Mandarin formed with the morphemes wâng ‘net’ and bā ‘bar’. The compound is created by analogy with other bā-headed compounds such as 酒吧 jiūbā ‘bar’ lit. ‘liquor-bar’, 水吧 shuībā ‘juice bar’ lit. ‘water-bar’, and 氧吧 yângbā ‘oxygen bar’ lit. ‘oxygen-bar’.

2.5 Zi in Chinese

The term zì is usually translated as ‘character’, meaning a written Chinese character, but this translation does not suitably invoke the role of zì in defining what Chao (1968a:136) aptly terms the ‘sociological word’ in Chinese speaking societies, which he defines as “the unit, intermediate in size between a phoneme and a sentence, which the general, nonlinguistic public is conscious of … and is practically concerned with in various ways.” A more suitable translation of zì might be ‘morpheme-syllable-character’, since Chinese speakers use zì to refer to the morpheme-syllable as well as to the written character itself. This is similar to the way that English speakers use word to refer to space-separated letter strings and not just to the spoken forms that these orthographic units represent. Bauer and Benedict (1997:285-86) elaborate on the relationship between spoken and written zì as follows:
“Because the Chinese characters have been so highly esteemed by Chinese people throughout the history of Chinese civilization, they have shaped the way both literate and non-literate Chinese people think about the spoken and written forms of the Chinese language. When most speakers of Chinese dialects … reflect upon the Chinese language, what comes first to mind is not its spoken form but rather Chinese characters … as far as the layman is concerned, the Chinese characters equal the Chinese language.”

Bauer and Benedict suggest that morpheme-syllables and the characters used to write them shape Chinese speakers’ metalinguistic awareness whether or not they are literate. That is, all Chinese speakers, literate or not, are influenced by Chinese culture, in which written zì are an ever-present part of linguistic and social interaction.

Chao (1968a) asserted that even in the case of polysyllabic words, it is still the individual zì with which they are composed that serve as the primary unit of metalinguistic awareness for Chinese speakers. Following Chao and others, I consider zì to be the anchor for Chinese speakers’ metalinguistic knowledge, and I use this term in the remainder of the dissertation to refer to Chinese morpheme-syllable-characters. In Chapter 3, I consider how lettered words relate to Chinese zì.

2.6 Defining word in Chinese

The technical term word is translated by the Mandarin term cì. This term is more difficult to define for Chinese and continues to be a matter of controversy in Chinese linguistics. Although words exist in Chinese, I follow Chao and others in assuming that
they are not the basic units of metalinguistic awareness that they are in English. Chinese
speakers do have a robust implicit knowledge of words, as shown in psycholinguistic
experiments suggesting that bimorphemic units are stored and perceived as single units
Taft and Zhu 1997), but words are not as salient to them as to English speakers. Rather, I
suggest that the zì that are the components of words are more salient to Chinese speakers
(Riha 2008).

Since Chinese does have words, however, scholars have taken pains to define what
they are. In working toward a rigorous definition of word in Chinese, Packard (2000) first
provides a summary of various general definitions of ‘word’, including ‘orthographic
word’, ‘lexical word’, ‘semantic word’, ‘phonological word’, ‘psycholinguistic word’,
and so on, and shows why each of these is inadequate. He concludes that the notion of
‘syntactic word’, that is, “a form that can stand as an independent occupant of a syntactic
form class slot, … commonly designated in the literature as $X_{0}$” (p. 12), is most accurate
in describing words both in Chinese and cross-linguistically. In this study I follow
Packard in defining Chinese words as ‘syntactic words’.

What is perhaps most confusing about words in Chinese is that they can consist of
either one zì or more that one zì. There are many words in Chinese that consist of one zì
that is a free morpheme (e.g. 我 wǒ ‘I’, 人 rén ‘person’, etc.; Chao (1968a:190) gives
other examples). This class of words forms the most basic elements of the lexicon, but
such words constitute only a small part of modern Mandarin usage. A Chinese
government study found that in a total of 3,624 words that represent about eighty percent
of all occurrences of words in modern Mandarin, monosyllabic words constituted just twenty-nine percent of the total (ZWGW 1959). Curiously, the Chinese words that Americans are most often exposed to are exactly this set of monosyllabic words. Americans frequently encounter certain monosyllabic Chinese words on gift items, such as jewelry with the words 福 ‘good fortune’ and 寿 ‘(long) life’ and decorative items with the words 和 ‘harmony’ and 爱 ‘love’. This probably reinforces the common Western misperception that Chinese words are primarily monosyllabic.

Words formed with two or more 字 are either compounds (e.g. bīngshān ‘iceberg’, lit. ‘ice-mountain’) or derived words (e.g. fāngzi ‘house’, lit. ‘house-SUFFIX’). Words containing two 字 are the most common polysyllabic words in Chinese (Sawer 1995, Taft and Zhu 1997), and most are compounds rather than derived words (Duanmu 2007). As for why disyllabic words are preferred, one popular explanation is ambiguity avoidance. Disyllabic words are thought to be favored as a compensatory measure for dealing with the large number of homophonous syllables in Chinese (Lü 1963, Li and Thompson 1981). Duanmu (2007:190) argues against this view, stating that metrical considerations are more important than ambiguity avoidance. In his view, words with phrasal stress are required to be disyllabic, whereas those without phrasal stress can be monosyllabic. This has lead to the creation of a dual vocabulary in Chinese in which each word has a monosyllabic form and a disyllabic form. Duanmu (2007:162) notes that dual vocabulary items are generally formed either through compounding (e.g. 耳朵 ěrduo ‘ear’ lit. ‘ear-petal’, 耳 ěr ‘ear’) or truncation (e.g. 中学 zhōngxué ‘middle school’, 中 zhōng ‘middle school’). In addition to metrical considerations, Duanmu points out that context also
plays an important role in the use of the dual vocabulary. In situations where zhōng ‘middle’ abbreviates zhōngxué ‘middle school’, for example, speakers must be aware that the topic under discussion is “types of schools” in order to know that zhōng refers to ‘middle school’ rather than simply to ‘middle’, its usual meaning.

2.7 Compounding

Just as simple words are difficult to define in Chinese, so are compound words. The definition of ‘compound’ in Chinese is hard to pin down due in part to the problem of determining the boundedness of Chinese morphemes. Although some morphemes that were free in classical Chinese are no longer free in modern Chinese, classical sayings and grammatical constructions are used widely in the modern language, making it difficult to state categorically that a morpheme is always free or always bound. Thus, linguistic register plays a role in determining whether a morpheme is free or bound (Packard 2000:68). The morpheme 言 yán ‘speak, speech’, for example, is bound in modern Mandarin but free in classical Chinese and in idiomatic phrases from classical Chinese that are borrowed into the modern language, such as 言简意赅 yán jiǎn yì gāi ‘brief and to the point’, lit. ‘speech-brief meaning-inclusive’. Polysemy also contributes to variation in boundedness, with some entries for a highly polysemous form involving a bound morpheme and others a free morpheme.

Some linguists sidestep the issue of boundedness to argue for a broad definition of compounds. Li and Thompson (1981:46) define compounds as “all polysyllabic units that have certain properties of single words and that can be analyzed into two or more …
morphemes” even if the morphemes are not free in modern Mandarin. Duanmu (2000:97) also defines a ‘compound’ as any word composed of two or more morphemes, regardless of whether they are both free.

Other linguists use boundedness to argue for a more narrow definition of ‘compound’. Dai (1997:126) defines a compound as a syntactic word composed of two or more syntactic words. Since only free morphemes can be syntactic words, this leaves out the possibility that polymorphic words that include a bound morpheme could also be called ‘compounds’. For example, 电脑 diànnǎo ‘computer’ lit. ‘electric-brain’ would not qualify as a compound since diàn is a free morpheme but nǎo is a bound morpheme.

Packard’s (2000) categorization of disyllabic complex words in Mandarin similarly restricts the notion of ‘compound’ only to words composed of two root words (e.g. 马路 mǎlù ‘street’ lit. ‘horse-road’ in which mǎ and lù are both free morphemes). Packard categorizes polysyllabic words that contain bound roots and affixes as other types of words. *Bound root words* contain a root word and bound root or two bound roots (such as diànnǎo ‘computer’ above); *derived words* contain a bound root or root word and a derivational affix (e.g. 椅子 yǐzǐ ‘chair’, lit. ‘chair-SUFFIX’); and *grammatical words* contain a word and a grammatical affix (e.g. 我们 wǒmen ‘we’, lit. ‘I/me-PLURAL’).

Packard suggests that different types of polysyllabic words need to be distinguished because words of each type share properties at a certain level of abstraction and because their shared properties reflect the rules of their formation. In my study of lettered words, I have not found it beneficial to categorize lettered words into Packard’s four categories. Rather, I find it more informative to think of them simply as compounds or derived words.
The reason for this is that I have thus far found no patterns in lettered word formation that would necessitate distinguishing compounds from bound root words, nor have I found any grammatical words that are lettered words. There are a small number of derived words that are lettered words, such as e 电子 ‘e-ize’ (e < electronic), but I did not find any words of this type in my corpus study. In the discussion of lettered words in the following chapters, I focus simply on compounds rather than on the distinctions among Packard’s four categories of word types.

2.8 Abbreviation

Abbreviation is a common way of forming new words in Mandarin. Sawer (1995) found that after compounding, abbreviation is the second most frequent word formation process for creating news words. In Sawer’s study, compounds constituted 70 percent of new words in Mandarin and abbreviations 19 percent. Abbreviations are compound words that are formed by selecting a few key morphemes from a compound or idiomatic phrase. In a study of Mandarin abbreviations, Chen (1998) found that an overwhelming majority of abbreviations are disyllabic and that they are usually formed by taking the first morpheme of each word in a four-syllable compound name made by combining two disyllabic words. For example, 北京大学 Běijīng Dàxué ‘Peking University’ (where the two component words are literally ‘north-capital’ and ‘big-learning’), is abbreviated as 北大 Běidà. This method may be avoided, however, when it can result in ambiguities or non-grammatical forms. Lin (2001:81) observes, for example, that 清华大学 Qīnghuá Dàxué ‘Qinghua University’ in Beijing is abbreviated as 清华 Qīnghuá rather than as as
because Qingdà could potentially refer to the names of at least two other universities in China. Qinghai University’ in the capital of Qinghai Province and Qingdao University’ in Qingdao, Shandong Province. Lin notes that in Taiwan, which also has a Qinghuà Dàxué but no competitor universities for the abbreviation Qingdà, Qingdà is used as the abbreviation of the name of the university.

Another common type of abbreviation, briefly discussed by Lin, is the combination of a referential number morpheme followed by a content morpheme from the original expression. For example, the maxim “学习好 xuéxi hǎo, 工作好 gōngzuò hǎo, 身体好 shēntǐ hǎo” ‘study well, work well, maintain good health’ is abbreviated as 三好 sānhǎo ‘three virtues’, lit. ‘three-good’. In a slightly different type of abbreviation that is also frequent in Mandarin, a morpheme not common to all of the words in the expression is used in the abbreviation and is often a hypernym for the group. For example, a set of spices called 五香 wǔxiāng ‘five spices’ is commonly used in Chinese cooking, but only two of the spices have 香 xiāng ‘spice’ in their names, 丁香 dīngxiāng ‘clove’ and 花香 huāxiāng ‘fennel’. Similarly, the abbreviation for the five basic flavors (sweet, sour, bitter, spicy, and salty) is 五味 wǔwèi ‘five flavors’, but the flavors do not have 味 wèi ‘flavor’ in their names.

The patterns for forming abbreviations are relevant for the formation of pīnyīn initialisms, which I discuss in Chapter 4. Some pīnyīn initialisms are formed with these same principles of abbreviation.
2.9 Distinguishing compounds from phrases

A final issue to discuss is the vexing problem of distinguishing compounds from phrases in Mandarin. This is yet another difficult area of Mandarin grammar, one that I mention because it pertains to the corpus study of lettered words discussed in Chapters 5 and 6. Of concern is the difference between a noun-noun compound and a noun phrase consisting of two nouns (actually noun phrases) linked by an assumed 的 de particle, which is omitted from the phrase and makes it look like a compound. My data on lettered words includes numerous examples of this type, and I needed to determine whether it is meaningful to distinguish lettered strings as compounds and phrases or whether it would be just as effective to treat them as one group.

The problem is best understood by first describing the nature of noun phrases with the de particle. Li and Thompson (1981) state that noun phrases containing the de particle consist of an associative phrase followed by a second noun phrase. The associative phrase consists of a noun phrase together with de, which indicates a connection, or association, between the two phrases. The associative phrase modifies the second noun phrase, and the second noun phrase is the head. Li and Thompson (1981:113) schematize this as shown in 2.2.

2.2 \[\text{NP1-de} \quad \text{NP2}\]

associative phrase  head noun

One type of associative phrase is the possessive phrase, an example of which is given in 2.3a. The de particle indicates the genitive class of relationships, including the possessive meaning, and is glossed as POSS in 2.3a. The second type of associative phrase indicates a
broader notion of semantic association and is glossed as ASSOC by Li and Thompson, as in 2.3b and c.

2.3  
a. *wǒ* -de *chénshān*  
I POSS shirt ‘My shirt’  
b. *kēxué* -de *fāzhǎn*  
science ASSOC development ‘the development of science’  
c. *chénshān* -de *kòuzì*  
shirt ASSOC button ‘the buttons (on) the shirt’

Although the relationship between *kēxué* and *fāzhǎn* in 2.3b could be interpreted as ‘science’s development’ and that in 2.3c as ‘the shirt’s button’, the relationship in these examples is clearly different from that in 2.3a.

Noun phrases like those in 2.3 can be distinguished easily from compounds by the presence of the *de* particle. There are many instances, however, when the *de* particle is omitted for semantic or metrical reasons, giving the two noun phrases the appearance of an noun-noun compound. Chao (1968a:288) calls noun phrases in which the *de* particle is omitted ‘zero’ *de* phrases. He illustrates the flexibility with which *de* can be included or omitted using the example shown in 2.4, which contains the *de* particle and the following words: 好 *hǎo* ‘good’, 学生 *xuéshēng* ‘student’, and 宿舍 *sùshè* ‘dormitory’.

2.4  
a. 好 学生的宿舍  
*hǎo* 学生的宿舍  
‘dormitory for good students’  
(lit. ‘[[good students’] [dormitory]]’)

b. 好的学生 宿舍  
*hǎo* 的学生 宿舍  
‘good dormitory for students’  
(lit. ‘[[good] [student dormitory]]’)

c. 好 学生 宿舍  
*hǎo* 学生 宿舍  
ambiguous; most likely interpreted as a

Chao suggests that ‘zero’ marks smaller constituents and that *de* marks larger ones. Thus, in his view *de* is not needed in 2.4a between *hǎo* ‘good’ and *xuéshēng* ‘student’ because
they form a small constituent, but it is added between hào xuéshèng ‘good student’ and sùshè ‘dormitory’ because they form a larger constituent. Similarly, in 2.4b de is omitted between xuéshèng ‘student’ and sùshè ‘dormitory’ because they form a small constituent but is added instead between hào ‘good’ and xuéshèng sùshè ‘student dormitory’ because these words form a larger constituent. Example 2.4c is open to interpretation as either a or b. Chao suggests that speakers are more likely to assume the interpretation in a than b. Finally, 2.4d can be interpreted as either a or b.

From the perspective of my study, the point of Chao’s example is that words joined with ‘zero’ de, such as xuéshèng sùshè ‘student dormitory’ and hào xuéshèng sùshè in 2.5c look just like compounds, and I am faced with the question of whether their structure is essentially the same as that of compounds or different in a way that is meaningful for my study. Linguists have grappled with the question of whether phrases that use de are semantically different from those that do not (e.g. hào de xuéshèng ‘good student’/‘student that is good’ vs. hào xuéshèng ‘good student’), but have not reached a consensus. Duanmu (2007) provides a review of the literature and suggests that all nominals with the structure [modifier de noun] ([M de N]) are phrases and that those with the structure [M N] are compounds. Thus, he treats ‘zero’ de nominals as compounds. In Duanmu’s view, hào xuéshèng ‘good student’, xuéshèng sùshè ‘student dormitory’, and hào xuéshèng sùshè ‘dormitory for good students / good dormitory for students’ would all be compounds. Conversely, hào de xuéshèng ‘good student’, xuéshèng de sùshè ‘student dormitory’, and hào de xuéshèng de sùshè ‘dormitory for good students / good dormitory for students’ would all be phrases.
My study does not aim to resolve the debate about whether ‘zero’ *de* phrases are compounds. I find that it is not necessary to delve into this question for the purpose of analyzing the lettered word data I present in Chapter 6. Rather, what is of greater interest in my study is the endocentric, ‘modifier + head noun’ structure shared by both compounds and ‘zero’ *de* phrases in Mandarin. Lettered compounds and lettered ‘zero’ *de* noun phrases have the same [M N] structure, and it is this structure that is ultimately of interest in lettered strings. I find that the type of unit that occupies the modifier and head slots, that is, whether it is roman letter-based or Chinese *zi* based, is more informative to analyze than aiming to determine whether lettered compounds should be distinguished from lettered ‘zero’ *de* phrases. In Chapters 5 and 6, I conflate the two types of lettered strings into one category, modifier-noun nominals ([M N] nominals), a term I borrow from Duanmu (2007). Following Duanmu’s analysis of Mandarin noun compounds and phrases, I treat all lettered strings containing a modifier and a head noun as [M N] nominals.
CHAPTER 3

THE CHINESE WRITING SYSTEM

3.1 Chinese writing as ‘real’ writing

The Chinese writing system can be divided into five main types of units (Chao 1968b:103). They are characters that originated as: 1. pictographs, or pictures of objects (马 ‘horse’), 2. ideographs, or symbols for ideas (一 ‘one’), 3. compound ideographs, in which the meaning of the whole character is a combination of its parts (亼 ‘honest’, consisting of 人 ‘person’ and 言 ‘word’), 4. loan characters, or rebus characters, that are borrowed to write a homophonous word (来, originally a pictograph meaning ‘(a type of) grain’ was used to write the homophonous word ‘come’), and 5. phonetic compounds, or what DeFrancis (1989) calls semantic-phonetic compounds, that consist of a semantic radical (signific) and a phonetic (蝗 huáng ‘locust’ consists of the radical 虫 ‘insect’ and the phonetic 黄 huáng, used independently to mean ‘emperor’). A commonly believed view in the West is that Chinese characters are primarily pictographs and ideographs, and little attention is paid to rebus characters or semantic-phonetic compounds. Only a small number of Chinese characters can be described as pictographs and ideographs, however. They currently form less than three percent of characters in Chinese (Chao 1968b).
DeFrancis argues that the vast majority of characters are structured to encode primarily phonetic information.

DeFrancis suggests that the encoding of phonetic information is a prerequisite for a ‘full’ or ‘real’ writing system. It must be based on speech and cannot involve the manipulation of soundless symbols, that is, pictographs or ideographs, to obtain meaning directly without the use of language. To qualify as a real writing system, Chinese writing must represent speech rather than pictures or ideas. DeFrancis claims that real writing systems do not differ in the basic requirements for the features of writing but rather in the details of how they represent the phonetic aspects of speech. Thus, Chinese and English writing are essentially the same in that they both represent the phonetics of speech, but they differ in how they do so.

DeFrancis (1989:49) proposes the “Duality Principle” to explain how writing systems are structured. They use symbols that represent the sounds in speech, as well as using symbols that add nonphonetic information. These two aspects are combined in different proportions in different scripts. The “phoneticity” of a writing system, that is, the fit of symbols to sounds, can range from high to low. The less efficient a system is in its phonetic representation, the more it compensates by the use of nonphonetic devices. DeFrancis (1989:58, 2002:16) proposes a continuum from “pure” phonetic writing, in which speech is represented faithfully, to “meaning-plus-sound” scripts in which semantic information is added to supplement less effective sound-to-symbol correspondences. Languages that have relatively good sound-to-symbol correspondences are divided as follows: pure phonemic systems such as Finnish, Greek, and Latin; pure
consonantal systems such as Arabic and Hebrew; and pure syllabic systems such as Cherokee and Linear B. Languages with poorer sound-to-symbol correspondences requiring supplemental semantic information are also divided into three groups:

morphophonemic systems such as English, French, and Korean; morphoconsonantal systems such as Egyptian; and morphosyllabic systems such as Chinese and Mayan.

Japanese is conveniently left out of DeFrancis’s framework, most likely because it has a mixed writing system consisting of pure syllabic kana and morphosyllabic Chinese characters. The borrowing of foreign scripts resulting in a mixed writing system poses problems for the classification of writing systems, as noted by Hansell (2002). It is indeed a mixed writing system that is used to write many lettered words in Chinese, however. This mixing of scripts is a new issue to grapple with in the classification of the Chinese writing system. I return to the question of script mixture in Chinese in Chapter 4, where I discuss how the morphophonemic system of English is combined with the morphosyllabic system of Chinese to write certain types of lettered words.

3.2 Rebus writing in Chinese

Although pictures can be used as abstract representational symbols in writing, their usefulness is limited since it is difficult to express a large number of concepts with pictures. Rather, the use of pictures as rebus symbols was needed for the development of writing in Chinese.

Rebus symbols are commonly defined as representations of a word or syllable by pictures of objects whose names resemble the sounds of the word or syllable. The
morpheme ‘I’ in English, for example, can be represented with a picture of an eye. Crystal (2001:284) defines the term *rebus* more broadly to include not just pictures but letters, symbols, and logograms as well. Thus, *IOU* ‘I owe you’, *H&* ‘hand’, and *K9* ‘canine’ all qualify as writing composed of rebus symbols. Crystal’s broader definition of rebus writing is useful for describing transliteration with Chinese *zì*, as well as one of the functions of roman letters in Chinese lettered words, an issue I discuss in the next chapter.

Boltz (1994) describes the evolution of Chinese writing that led to use of the rebus principle in three stages. First, in the pictographic stage, pictures were used to represent those parts of the language that were picturable. For example, a picture of an elephant was used to represent the word in Ancient Chinese that developed into the word 象 (xiàng) ‘elephant’ in Putonghua. Second, in the ‘multivalent stage’, graphs were used as rebus symbols to represent homophonous words (e.g. 象 ‘elephant’ was used to write 像 ‘image’) or to write words that were semantically related but had different pronunciations (e.g. 眼 ‘eye’ was used to write 见 ‘to see’). Third, the ambiguity of the second stage was resolved through the use of semantic and phonetic determinatives. For example, a variant of the 人 ‘person’ radical was added to the phonetic base 象 to distinguish the morpheme 像 ‘image’, and another variant of the same radical was added to 目 ‘eye’ to distinguish 见 ‘to see’. The semantic-phonetic compounds created in this third stage of Chinese writing account for 97% of Chinese characters in the modern language (DeFrancis 1989:99).
3.3 Semantic-phonetic compound characters

DeFrancis (1989) estimates that almost all of the characters encountered in modern texts are combinations of 214 radicals and 4,000 phonetics, a large but manageable number. He considers the phonetics to be a kind of syllabary to which radicals are added to distinguish among homophonous (or similar-sounding) forms. For example, one of the characters in the syllabary is 门 mén ‘door’. Some of the radicals that can be added to it and the resulting derivatives are shown in Table 3.1.

<table>
<thead>
<tr>
<th>syllabic element</th>
<th>门 mén</th>
<th>derivative</th>
</tr>
</thead>
<tbody>
<tr>
<td>signific</td>
<td>金 ‘metal’</td>
<td>鎏 mén ‘mendelevium (Md)’ (a chemical element)</td>
</tr>
<tr>
<td></td>
<td>心 ‘heart’</td>
<td>門 mēn ‘stuffy’</td>
</tr>
<tr>
<td></td>
<td>口 ‘mouth’</td>
<td>間 wèn ‘ask’</td>
</tr>
<tr>
<td></td>
<td>耳 ‘ear’</td>
<td>閱 wén ‘hear’</td>
</tr>
</tbody>
</table>

Table 3.1 Semantic-phonetic compounds

An important characteristic of semantic-phonetic compound characters is that the phonetic element may indicate the syllable pronunciation or tone with only partial accuracy. DeFrancis (1989:113) estimates that 25% of characters have ‘completely’ useful phonetics representing all the phonemes in the derivative (鎏 mén); 17% have ‘generally’ useful phonetics representing all the segments but not the tone (閤 mēn); 24% have ‘contextually’ useful phonetics that represent most of the segments (閏 wèn and 閵
wén); and 33% have ‘useless’ phonetics representing no segments ((artist: xià, in which the phonetic is artist: bū).

The examples in 3.1 show that the radical frequently provides just a general clue to the meaning of the character rather than a specific description. DeFrancis (1989:107) describes it as a “thesaurus-like clue.” Coulmas (2003) similarly suggests that it is useful primarily as a lexical ordering principle, rather than a productive semantic component, since dictionary entries of characters are organized according to their radicals. (The first Chinese dictionary to be organized by radicals is the Shuōwén Jiēzì, completed in AD 100 (Norman 1988).) Both DeFrancis and Coulmas maintain that radicals have limited utility, and Coulmas goes so far as to ask whether they serve any useful function. Boltz (1994:122) claims that “in many cases no more than an aesthetic appeal underlies the use of semantic determinatives.”

I disagree with all three authors and suggest that radicals have considerable utility beyond their aesthetic appeal. Although radicals generally do not provide precise information about the meanings of morphemes written with particular characters, they do make a variety of indirect contributions to meaning and function. An important visual effect of the presence of radicals in characters is that they help to group morphemes that form a constituent at some level of meaning. This can be illustrated with an example using one of the 214 radicals. In 3.1 the ++ ‘grass’ radical indicates visually that the strings of characters should be grouped together to form semantic units.

3.1 茉莉 mòli ‘jasmine’ – a monomorphic, disyllabic word
茉莉花 mòli huā ‘jasmine flower’ – a bimorphic compound
茉莉花茶 mòli huā chá ‘jasmine flower tea’ – a trimorphic compound
The characters in 3.1 illustrate a regularity in the composition of characters that is helpful for recognizing the components of complex characters, as well as the complex characters themselves. In semantic-phonetic compounds, radicals are frequently on the left or on top, while the phonetic is frequently on the right or below the radical (Gao & Kao 2002). (In 3.1 the ‘grass’ radical is on top and the phonetics are below it.) Peng et al. (1997) show that even schoolchildren are attuned to these conventions and use them to comprehend characters. The authors conclude that radicals and phonetics and their relations are the primary information sources used for Chinese character recognition. Thus, I suggest that far from being uninformative, radicals contribute considerably to meaning and character recognition. In fact, one of the limitations that roman letters may have as graphs in the Chinese writing system is that they are not differentiated with radicals in the same way as Chinese semantic-phonetic characters. Roman letters indicate visually that they themselves form a group (e.g. $p$ and $H$ in pH 值 ‘pH level’), but they are not linked visually to the Chinese characters with which they form a constituent.

3.4 Creation of new characters

Mair (1996:200) notes that the current edition of the comprehensive dictionary of Chinese characters published in the PRC, Hányǔ Dà Zìdiǎn (1986-90), lists about 60,000 characters, a formidable number. The number of characters in daily use is much smaller than that, however. Recent statistical surveys show that 1,000 characters cover approximately 90% of all occurrences in typical texts; 2,400 cover 99%; and 5,200 cover
99.99% (Mair 1996). Less than 3,000 characters are considered as essential for basic reading and writing.

What is not often mentioned, however, is that even though several thousand characters form the basic stock of characters in common use, the creation of new characters is infrequent. There appears to be a resistance to creating new characters rather than using existing ones for new purposes. Chen (1999:116) states that new characters are coined primarily for dialect writing but that very few “have gained acceptance by people other than the inventors.” The creation and use of new characters is in fact discouraged in the PRC, where language authorities have worked to simplify and standardize Chinese characters since the 1950s. Hansell (2002) points out that when new characters are created, they tend to be formed as semantic-phonetic compounds, the most common type of character.

The preference for using existing characters is similar, in a sense, to English writing. New words are written with combinations of letters from of the existing stock of twenty-six letters in the English alphabet rather than with newly created letters, although diacritics are sometimes used to augment existing letters in the writing of foreign terms and names (e.g. é, è, ñ).

The stock of graphs used in Chinese can be expanded even though new characters are created infrequently. Chinese writing is open to borrowing graphs from other writing systems and incorporating them into the Chinese writing system. For example, some native Japanese characters (kokuji) created with the same components as Chinese characters and with the same principles have been borrowed into writing in Hong Kong.
and Taiwan. The Japanese hiragana letter の [no] that writes the morpheme ‘of’ is also popular in the commercial sectors of Hong Kong and Taiwan in place of the Chinese characters 之 and 的 used to write the morpheme ‘of’ (Chan and Kwok 1982). Greek letters are used in certain scientific terms, such as α粒子 ‘alpha particle’ and β射线 ‘beta ray’. But perhaps most importantly, the twenty-six letters of the roman alphabet have been borrowed into written Chinese as a new set of graphs that Hansell (1997) calls the *Sino-alphabet*. I suggest that roman letters are not simply a set of new graphs in the writing system, however, but actually a new set of Chinese zi, an idea I elaborate on in the next chapter.

### 3.5 Graphemes and frames in writing

In order to discuss how roman letters are used as new characters in Chinese writing in Chapter 4, I need to describe the *graphemes* and *frames* of writing in Chinese and English, since it is these units that must be coordinated to allow for the writing of lettered words in Chinese. The classification systems proposed by DeFrancis (1989) and Unger (1992, 2004) rely on clearly distinguishing the graphemes and frames of writing in various writing systems. Mixed writing systems need to harmonize their use of these basic units in order to mix the graphs from the different writing systems effectively.

The *grapheme* is the meaningless graphic unit that corresponds to the smallest unit of speech represented in writing, and the *frame* is the basic unit of writing surrounded by white space or other delimitative structure on a printed page (DeFrancis 1989). In English, graphemes are single letters or combinations of letters that represent phonemes
Individual letters may also constitute frames, such as the article *a* and the pronoun *I*, but they are more commonly combined to form frames representing orthographic words. In Chinese, graphemes are characters that represent single syllables. They may themselves constitute frames (口中 ‘mouth’) or combine with nonphonetic elements to form complex characters that represent frames (唱 ‘to sing’, which has 口 ‘mouth’ on the left as the radical). In English, frames almost always consist of more than one grapheme and take up differing amounts of space (e.g. *I* or *a* vs. *antidisestablishmentarianism*). In contrast, Chinese frames consist of only one simple or compound grapheme, and each grapheme uses the same amount of space, as in 3.2.

3.2 simple grapheme 口 ‘mouth’
complex grapheme 唱 ‘to sing’

The graphic discreteness of Chinese graphemes illustrated in 3.2 is referred to by Boodberg (1957) as ‘equidimensionalism’. No matter how complex the internal form of a grapheme is, with some having upwards of twenty individual ‘strokes’ (marks), each grapheme takes up the same amount of space on a page.

While DeFrancis (1989) asserts that graphemes are the most important units of a written language, Unger (1992) maintains that frames have more functional value. I would claim that attention to frames may be more intuitive for native speakers than attention to graphemes. In English, the frame is the orthographic word, while in Chinese, it is the character. English speakers may consider the orthographic word to be the ‘basic unit’ of writing, while Chinese speakers consider the character to be the basic unit. Hoosain (1992) found, for example, that when Chinese speakers were asked to mark word boundaries in Chinese sentences, there were many discrepancies in their responses.
I suggest that one reflection of this difference between basic units in the two languages is that the units that are entries in dictionaries may tend to represent the frames in the respective language. Most entries in English dictionaries are orthographic words, while those in Chinese dictionaries are individual characters, with subentries for complex words and idiomatic phrases containing those characters. For example, a subentry under 唱 chàng ‘sing’ would be 唱片 chàngpiàn ‘record’ lit. ‘sing-disk’. In considering the composition of lettered words in the next chapter, an important issue I discuss is how Chinese speakers come to terms with the difference in the frames of the two respective languages.

It is important to point out that the Chinese writing system delimits characters, that is, the syllable-morpheme level, but does not delimit words. All the characters in a sentence are spaced out evenly rather than having white space between polysyllabic words written with multiple characters. (So, 他是画家 ‘He is (an) artist’) is not written as 他 是 画家.) Readers simply have to know which character strings form words and phrases because these divisions are not indicated in the writing system. In contrast, English writing delimits units above the morpheme level, that is, words, but is normally underspecified at the morpheme level. Orthographic words are separated by white space, but morphemes generally are not. Although some compounds in English are morpheme-delimited (<coffee table>), and hyphenation helps to delimit morphemes in some derived words (<e-mail>) and compounds (<low-tech>), in most cases readers cannot rely on the English writing system to identify morphemes. Just as Chinese readers rely on their linguistic knowledge to identify constituents above the morpheme level (i.e. to know that
唱片 chànghào forms a compound meaning ‘record’), readers of English rely on their linguistic knowledge to divide written words into their morphemes (i.e. to know that <classmate> is composed of the morphemes class and mate). Both writing systems require readers to use linguistic knowledge beyond what is given in the orthography to recognize constituents at various levels of grammar.

In sum, underspecification in the two writing systems occurs at opposite levels. Generally speaking, Chinese writing delimits syllables but not segments, while English writing delimits segments but not syllables. Chinese writing underspecifies at the word level, while English writing underspecifies at the morpheme level. In my view, these characteristics of the writing systems may either reflect or may have shaped the intuitions of native speakers about salient linguistic units in their language. The frame, in particular, may be set at the psycholinguistically ‘basic level’ in each language, the anchor for speakers’ metalinguistic knowledge. If this is indeed true, then it appears that a mismatch exists between the graphemes and frames of Chinese and English writing with respect to both the phonetic and morphological units represented in each. I discuss how the mismatch is overcome in lettered words in Chapter 4.

3.6 The development of Hányǔ pīnyīn romanization

After the founding of the People’s Republic of China in 1949, attention continued to be paid to language reform in China, including simplification of the traditional script (Chen 1999). Although some reformers favored abolishing traditional characters in favor or romanization, character simplification won out as the more important goal to be
pursued in language reform. An official list of simplified characters was published in 1956 and subsequent lists were published after that time (Ramsey 1987). While plans for the alphabet were considered less critical than character simplification, a new phonetic scheme, Hányǔ pīnyīn ‘Chinese phonetic alphabet’, or pīnyīn, was promulgated in 1958. Following the earlier latinized writing systems guóyǔ luōmázi ‘national language romanization’ and latinxua sin wenz ‘latinized new script’, which were not widely adopted, Hányǔ pīnyīn is phoneme based and uses the roman alphabet. Unlike latinxua sin wenz, however, which aimed to provide separate schemes for the major varieties of Chinese, Hányǔ pīnyīn is based entirely on the pronunciation of the Beijing variety of Mandarin, making it a useful tool for promoting pǔtōnghuà (Chen 1999), which also uses the Beijing variety for its pronunciation.

Like the earlier schemes, Hányǔ pīnyīn was designed not as a replacement for Chinese characters but rather as an auxiliary sound-annotating tool to facilitate the learning of characters in the standard language. In fact, DeFrancis (1950) notes that when the first draft of the scheme was circulated, it was called Hányǔ pīnyīn wénzì ‘Chinese phonetic writing system’, but by the time the scheme was promulgated in 1958, the word wénzì ‘writing system’ had been deleted, and the name of the system had been changed to Hányǔ pīnyīn fāngàn ‘Chinese phonetic scheme’. The renaming of the pīnyīn system is similar to the renaming of guóyǔ zìmǔ ‘national phonetic alphabet’ to be zhùyǔn fùhào ‘symbols for phonetic notation’ in 1930 to indicate definitively that the system was not to be a replacement for Chinese characters.
Also in 1958, Premier Zhou Enlai stated explicitly in a definitive report that pīnyīn would be used indicate the pronunciation of Chinese characters and to promote the standard language and that it was not to be used as a substitute for Chinese characters (Ramsey 1987:145). With that official pronouncement, the direction of language planning formally shifted emphasis from alphabetic writing to character simplification. Even with this change in the direction of language planning, however, Hányǔ pīnyīn functioned as a useful aid in promoting the standard language, particularly among speakers who were not familiar with the Beijing variety that serves as the basis for pǔtōnghuà.

Since its inception, Hányǔ pīnyīn has served as the official phonetic scheme for writing Chinese in the PRC. It is employed extensively in textbooks and reference works and is used as a common input system in Chinese word processing. The system was also accepted by the International Standardization Organization in 1982 as the standard form of transcription for Chinese words (Chen 1999).

3.7 Learning Hányǔ pīnyīn

3.7.1 Writing pīnyīn letters

Hányǔ pīnyīn is currently used widely in the PRC as an auxiliary system for transcribing Chinese personal and place names and for annotating Chinese characters, as well as in education for the deaf and blind and in library cataloguing (Rohsenow 2001). Essentially, it is used only “where Chinese characters are inconvenient to use or cannot be used” (Rohsenow 2001:128). Pīnyīn is also employed in the initial stages of teaching Chinese children to read Chinese. Tao and Zuo (1997) and Ingelsrud and Allen (1999)
describe the process, which is useful to know in considering how Chinese people are initially exposed to roman letters. Ingelsrud and Allen observed a primary school class on the first day that pīnyīn was introduced. They noticed that the characters 汉语拼音 were written on the board rather than their romanized form, Hányǔ pīnyīn. This suggested to them that the emphasis in instruction was on literacy as literacy in Chinese characters rather than roman letters. Tao and Zuo (1997:657) confirmed this view in discussing how pīnyīn is taught in primary school. They stress that the teaching of characters is “always a part of the language lesson” and state further that “reading and writing [characters] are never separate in Chinese education at primary levels.” Clearly, Chinese characters are learned as the primary writing system in Chinese.

Students learn the letters of the pīnyīn alphabet as pictographs, and only lower case letters are learned even though upper case letters are also commonly used in writing pīnyīn. In the class Ingelsrud and Allen observed, h was pictured as a chair, ü as a fish, and b as a transistor radio, with the curved section of the letter being the radio itself and the left-hand tail being the antenna (Ingelsrud and Allen 1999:87). The same terminology used to describe the strokes of Chinese characters was used to describe the parts of letters. The shape of b was called 树和圆 shù hé yuán ‘tree and circle’, for example.

Letters were also learned with verses that served as mnemonic devices. The letters n and m, d and b, and ü were learned with the verses in 3.3, which describe both the shape of the letters and their letter name pronunciations (Ingelsrud and Allen 1999:88).

3.3  a. 一扇小门 nǎn nǎ nǎ, 两扇小门 mǎ mǎ mǔ
    yīshàn xiǎomén nǎ nǎ nǎ, liǎngshàn xiǎomén mǎ mǔ mǔ
    ‘one small gate nǎ nǎ nǎ, two small gates mǔ mǔ mǔ’
b. 左下半圆 哆 - 哆 - 哆, 右下半圆 玻 - 玻 - 玻
zuòxià bànyuán duō duō duō, yòuxià bànyuán bō bō bō
‘lower left half circle duō duō duō, lower right half circle bō bō bō’

c. 竖右弯竖点点 向 - 向 - 向
shù yòu wān shù diǎn diǎn ù - ù - ù
‘vertical line, bend to the right, vertical line, dot dot ù ù ù’

The verses for the letters $b$, $d$, and $ü$ use the technical terms that are normally describe the strokes in Chinese characters. While chanting these verses, students wrote the letters in the air with their fingers and then on their desks, just as they do in learning Chinese characters. Ingelsrud and Allen point out that visual mnemonic devices such as the pictures illustrating the letters are used only to introduce the letter shapes. Students are quickly expected to use the ‘metalinguage’ of character shapes to describe the shapes of the letters. From the pedagogical methods used for teaching roman letters, the reference point in both teaching and learning appears to be Chinese characters and the way they are structured and written. Roman letters seem to be learned as supplementary graphs with the limited function of writing pīnyīn.

3.7.2 Pronouncing and writing pīnyīn

In addition to learning how to write the letters used in pīnyīn, students must also learn how to pronounce letter names and syllables formed with pīnyīn letters. The current method used to teach the names of the consonants (i.e. initials) is 直呼 zhíhū, which focuses on individual phonemes rather than syllables (Ingelsrud and Allen 1999). This is unlike the traditional method, less widely used now in the PRC but still used in Taiwan, which teaches the names of the initials as syllables. The letters $b$ and $p$, for example, are
now introduced as [pə] and [pʰə] rather than as the Mandarin syllables [puo] and [pʰuo]. (Although [pə] and [pʰə] are actually syllables rather than segments, they are not thought of as such since they are not legal syllables of Mandarin like [puo] and [pʰuo].) Simple finals are taught as segments (e.g. a in ba), but nasal finals such as -in, -ian, and so on are taught as clusters. Ingelsrud and Allen observed that the zhīhū method was not rigidly applied, however. Some students that had been taught pīnyīn at home or in kindergarten had learned to pronounce the consonant names as syllables, and their teacher did not correct them.

At the schools that Ingelsrud and Allen visited, the concept of the syllable was central to teaching the reading and writing of pīnyīn. The units that students practiced reading and writing were syllables rather than segments, and ‘words’ were introduced inconsistently in reading materials. Dictation exercises focused on initials, finals, and syllables rather than on individual letters or words. Moreover, although education officials claimed that pīnyīn education aimed to encourage a sense of word formation, Ingelsrud and Allen (1999) found that while bi- and trisyllabic words were written as single words on slates, in examining student test papers, they noticed that students tended to write words and sentences as strings of space-separated syllables instead and were not corrected. In my view, the lack of emphasis on teaching lexical boundaries in the use of pīnyīn most likely reflects the genuine difficulties of identifying ‘words’ in Chinese texts and Chinese speakers’ greater concern with the morpheme-syllables that make up words and phrases rather than with words themselves.
3.8 Use of pīnyīn once literacy is acquired

Although pīnyīn is a useful pedagogical tool in teaching reading in Chinese characters, Chen (1999) states that as Chinese students progress to higher grades and learn more characters, they lose their skills in pīnyīn and stop using it almost completely by the time they finish primary school. They do not need to continue reading or writing in pīnyīn because they use Chinese characters almost exclusively after primary school. According to Rohsenow (1996), in the standard curriculum students are taught pīnyīn for just two months purely as a phonetic notational device for the pronunciation of Chinese characters. After that, they read primarily Chinese characters. As a contrast to the standard curriculum, Rohsenow mentions a government-sponsored experiment in reading, in place since 1982, that uses pīnyīn for the first two years of schooling and claims considerable success in teaching students to read Chinese characters. Ultimately, however, most students read and write almost exclusively in Chinese characters once pīnyīn ceases to be needed in learning to read Chinese characters.

Apart from the teaching of reading, pīnyīn has become useful as an auxiliary tool in a new area, the computerization of Chinese. Chinese characters present challenges to the processing of Chinese by computers, and the use of pīnyīn to write Chinese characters provides a ready solution. Chen (1999) reports that there has been a steady increase in the PRC in the design of computer input/output systems for Chinese since the 1980s that has been fueled in part by the central government’s longstanding goal of promoting modernization. In writing on script standardization for information processing, Feng Zhiwei, a researcher at the Chinese Academy of Social Sciences, explains that China’s
aim is to construct an information superhighway similar to the one being designed by the United States. Feng states that if pīnyīn is used for information transmission in computers, China will have the same writing system in its computers as the “majority of developed countries in the world, improving the efficiency of China’s information network” (1997, quoted in Rohsenow 2001:134). Feng argues, however, that it is not necessary to replace Chinese characters with latinized writing, nor is it necessary to promote latinized writing among all Chinese citizens. Rather, he argues that digraphia should be promoted only for use in computers to improve China’s information network and link it with the networks of other developed nations.

Although the goal of language planners is to promote the use of pīnyīn in computer processing of Chinese, in actuality some Chinese in the PRC prefer to use character-based input systems in writing Chinese, such as the ‘five stroke’ system (wǔ bǐ xíng) (Rohsenow 2001, Su 2003). In this system the user analyzes characters into their most significant strokes, and the computer program then converts the coded input into Chinese characters. Rohsenow (2001) asserts that since pīnyīn is based on the pronunciation of putonghua, using pīnyīn for input is convenient only for those who know both putonghua and the spelling conventions for writing it in pīnyīn. This limitation has been overcome to some degree with an input system called liánpīn ‘connected pinyin’ that allows speakers to retrieve commonly occurring collocations by typing in just the initials of syllables or just one full syllable followed by the initials of the other syllables. For example, the compound 中国 Zhōngguó ‘China’ is retrieved by typing in zh and g, the compound 建设 jiànsè ‘construct’ by typing in jian and sh, and the idiomatic phrase 岂有此理 qǐ yǒu cǐ
lì ‘outrageous’ (lit. ‘where have this reason’) by typing in the initials of the four syllables, q, y, c, l. Liánpīn has become popular in the PRC for computer input and also for typing text messages on cell phones. Chinese friends of mine have pointed out, however, that it is used primarily by younger individuals familiar with pīnyīn. Older individuals not familiar with pīnyīn tend to prefer stroke-based input systems or systems that allow them to hand-write characters.

It is interesting to note that the segment-based nature of pīnyīn appears to be reduced as much as possible with liánpīn. Rather than requiring users to type in whole syllables, words, and phrases letter by letter as is properly done in pīnyīn writing, the initials of the constituent syllables are used as much as possible as abbreviations for the larger units. I suggest that this may be another example of the preference for the traditional division of Chinese syllables into initials and finals rather than their division into segments.

3.9 Use of zhûyûn fûhào in Taiwan

Unlike the PRC, Taiwan does not use Hânyû pīnyīn for phonetic annotation. The system is avoided in Taiwan for political reasons. When the Nationalist Chinese government retreated to Taiwan after the end of the Chinese civil war in 1949, language reform schemes that were developed during the republican era were transplanted to Taiwan as part of the Nationalist government’s effort to be the “preserver of traditional Chinese culture” (Rohsenow 2001:130). Guûyû, the ‘national language’, that is, Standard Mandarin based on the Beijing variety of the 1930s, became the national language in
Taiwan. Annotation of Chinese characters was done with zhùyīn fúhào ‘symbols for phonetic notation’ rather than with pīnyīn.

Zhùyīn fúhào is a revised version of guóyīn zìmǔ ‘national phonetic alphabet’, created in China in 1913 by the Ministry of Education’s Commission for Unifying Reading Pronunciation and adopted in 1918 for use in promoting the national language (Norman 1988). In 1930 the name of the alphabet was changed to zhùyīn fúhào ‘symbols for phonetic notation’ to explicitly define the system as an auxiliary means of indicating the pronunciation of Chinese rather than as a new writing system (Norman 1988). The phonetic alphabet was used widely in language instruction and continues to be used in Taiwan. The earlier romanized writing system, guóyǔ luómáizi ‘national language romanization’, created in 1928, is used in place of pīnyīn to transcribe Chinese names in a romanized form. Although the pīnyīn system was rejected entirely, it is now used in a limited way to write mainland Chinese personal and place names (Rohsenow 2001).

Unlike pīnyīn, the zhùyīn alphabet is modeled on Japanese katakana, one of the two native Japanese syllabaries. The symbols are simple graphs that are adaptations of ancient Chinese characters with few strokes and simple forms (Chen 1999). Although zhùyīn symbols are normally used in Taiwan mostly for phonetic notation, Su’s (2003) study of a group of Taiwan university students’ computer-mediated communication shows that the symbols can be used for informal written communication as well. Su calls this type of stylized writing zhùyīn wén ‘phonetic symbol writing’. In zhùyīn wén Chinese characters are replaced by zhùyīn phonetic symbols. The phonetic representation of characters is reduced to the initial, or less commonly, the vowel. It is frequently difficult to recover the
meanings expressed by zhùyīn wén since the reader has to supply the syllable-final information and the tone of the syllable, but the system is acceptable to use for in-group communication.

Su notes that zhùyīn wén is popular among the students she observed because it conveys a sense of “naivete reminiscent of compositions written by elementary school students” (p. 15). Although awkward as a writing system, zhùyīn wén is yet another way that Chinese speakers broaden the set of graphs at their disposal for writing Chinese. I suspect that zhùyīn wén may be used by Taiwanese people for word play in situations where individuals in the PRC might use pīnyīn writing.
CHAPTER 4

LEXICAL BORROWING IN CHINESE

4.1 Borrowing processes

4.1.1 Introduction

Chinese has several ‘traditional’ processes of borrowing that I describe in the following sections. Three basic principles underlie each of the processes. First is that foreign terms are normally incorporated into Chinese using the existing set of zi in the language. One or more morpheme-syllables are used to render the foreign term into Chinese, and the adapted term is written with the characters that write those morpheme-syllables. Second, with regard to the adaptation process itself, individual morphemes, syllables, or other sound groupings in the original term created through folk etymology, are mapped onto one or more zi. In each case, the parts of the word, whether it is divided up in terms of its meaning units, sound units, or both are ‘fit’ onto a string of one or more zi. Finally, in the adaptation process there is a trade-off with regard to the types of information about the original word that can be represented since the zi themselves have inherent pronunciations and meanings that influence the adaptation. If Chinese zi are used for their pronunciation but not their meaning, the foreign word or morpheme’s original
pronunciation is expressed, but not its meaning. If Chinese zi are used for their meaning but not their pronunciation, the foreign term’s meaning is expressed, but not its pronunciation. And finally, if Chinese zi are used for both their meaning and their pronunciation, some approximation of both the meaning and the pronunciation of the foreign term can be achieved. The different ways that words are parsed and fit onto strings of zi are the processes of adaptation described below.

The three principles described above also govern the formation of lettered words, as I show in later sections. Generally speaking, each process helps to ‘sinicize’ foreign terms to make them more like Chinese morphemes and words. In the words of anthropologist Melville Herskovits (1938, 5f.), “the fact that the receiving group can find in the traditions of the donor something understandable in terms of their own patterns makes for a lack of resistance and accelerates acceptance.” The process of using zi to adapt foreign terms makes them more acceptable in Chinese and easier to assimilate into the language. Anttila (2003:425) makes a similar point as Herskovits in his discussion of analogy in language use, stating that “lining up secure similarities gives an anchor for going into the uncertainties.” Using Chinese zi to represent foreign terms makes them understandable in terms of native linguistic patterns and provides an anchor for integrating the terms and concepts into Chinese.

4.1.2 Phonetic adaptation

Phonetic adaptations, or phonetic loans, interpret the pronunciation of a foreign word in terms of existing Chinese syllables but do not express its meaning. President Nixon’s
name, for example, was interpreted variously in Mandarin as *ni-ke-song*, *ni-ke-xun*, and *ni-ke-sun* (Norman 1988:138), but of course, no meaning of ‘Nixon’ could be conveyed with the morphemes used in the adaptation. A secondary process of selecting characters to write phonetic loans is usually also undertaken. When characters are chosen to write phonetic loans, the morphemes they normally write must also be taken into consideration even though their meanings are irrelevant in phonetic adaptations. Characters representing morphemes with negative, inappropriate, or unusual meanings are normally avoided, and infrequent characters are generally also avoided. DeFrancis (1984:9) provides an example. In transcribing the personal name *Anna*, for example, the most suitable syllables in Mandarin are *an* and *na*. The name is conventionally written with characters representing the morphemes 安 *ān* ‘peace’ and 娜 *nà* ‘elegant’. It could be represented equally well with characters representing the morphemes 暗 *àn* ‘dark, gloomy’ and 纳 *nà* ‘to patch’, but the contextually inappropriate meanings of the morphemes these characters represent preclude their use in transcribing the name.

A similar situation exists in English, as shown by the possible alternative transcriptions of the foreign names shown in Table 4.1.

<table>
<thead>
<tr>
<th>Foreign Name</th>
<th>Homophonous Written Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caesar</td>
<td>Seizer</td>
</tr>
<tr>
<td>Schubert</td>
<td>Shoebert</td>
</tr>
<tr>
<td>Cantonese surname</td>
<td>Lim</td>
</tr>
</tbody>
</table>

Table 4.1 Writing Foreign Names in English
Caesar, Schubert, and Lim could conceivably be written with the homophonous forms in Table 4.1, but this would require speakers to suppress the meanings of the morphemes in those forms and to use the spellings simply for their phonetic value, which would be awkward. To avoid similar situations in Mandarin, characters used to write phonetic adaptations are chosen with an awareness of the morphemes they normally represent so that inappropriate morphemes can be avoided.

Personal names are almost always adapted as phonetic loans because they are not amenable to calques. At the same time, examples of common nouns borrowed as phonetic loans in Mandarin are also not rare. Words such as 坦克 tǎnkè ‘tank’, 沙发 shāfā ‘sofa’, 打 dǎ ‘dozen’, and others express the pronunciation of the original English words, but the characters used to write them represent morphemes that are unrelated to the meaning of the word. In tǎnkè ‘tank’ tān ‘level’ and kè ‘restrain’ are unrelated to the meaning of ‘tank’, in shāfā ‘sofa’ shā ‘sand’ and fā ‘disperse’ are unrelated to the meaning of ‘sofa’, and so on. Certain characters in Chinese have also been created specifically to accommodate lexical importation (T’sou 2001). The characters 咖 kā/gā, 喝 lǐ, and 啤 fēi in 咖啡 ‘coffee’ and 啤酒 ‘curry’, for example, and 卡 kǎ in 卡路里 kǎlùlǐ ‘calorie’ are generally used primarily to write transcriptions of foreign words and do not have independent meanings apart from their function as phonetic characters.

4.1.3 Semantic adaptation
Borrowing of foreign terms is also done through semantic adaptation, which expresses the meaning of the foreign term but not its pronunciation. Semantic adaptation has two general forms, *loan translations* and what Chen (1999) calls *semantic translations*. Loan translations usually have a one-to-one mapping between the morphemes in the original and the Chinese morphemes used to adapt them. Examples are shown in 4.1.

4.1

<table>
<thead>
<tr>
<th>hot dog</th>
<th>honeymoon</th>
</tr>
</thead>
<tbody>
<tr>
<td>→ →</td>
<td>→ →</td>
</tr>
<tr>
<td>rè gǒu</td>
<td>mì yuè</td>
</tr>
<tr>
<td>hot dog</td>
<td>honey moon</td>
</tr>
<tr>
<td>‘hot dog’</td>
<td>‘honeymoon’</td>
</tr>
</tbody>
</table>

In contrast to loan translations, which generally translate the meaning of the foreign term closely morpheme by morpheme, semantic translations describe the meaning of the foreign term but do not translate it directly. They do not have a one-to-one mapping between the morphemes in the original word and its adaptation in Chinese. Examples include 抽水机 *chōushuījī* ‘pump’ lit. ‘draw-water-machine’, 女强人 *nǚqiángrén* ‘feminist’ lit. ‘female-power-person’, 汽车 *qìche* ‘car’ lit. ‘steam-vehicle’, and 话筒 *huàtōng* ‘microphone’ lit. ‘speech-tube’.

### 4.1.4 Simultaneous phonetic and semantic adaptation

T’sou (2001) suggests that the most successful cases of adaptation are those that convey both the pronunciation and the meaning of the original. One or more *zi* are selected to approximate the pronunciation of the original, and the meanings of the morphemes represented by the characters used to write the word simultaneously provide
clues about its meaning. I have found that the characters selected to convey the meaning of the term do not always do so, however. Rather, they may simply describe images, ideas, or associations with the foreign term that are thought to be relevant to its meaning. Examples are shown in Table 4.2 (adapted from T’sou 2001).

<table>
<thead>
<tr>
<th>English</th>
<th>Mandarin</th>
<th>Literal Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>vitamin</td>
<td>维他命 wèitāmìng</td>
<td>wèi ‘sustain’, tā ‘one’s’, mìng ‘life’</td>
</tr>
<tr>
<td>neon</td>
<td>霓虹 nǐhóng</td>
<td>ní ‘secondary rainbow’ hóng ‘rainbow’</td>
</tr>
<tr>
<td>logic</td>
<td>逻辑 luójì</td>
<td>luó ‘collect’, jī ‘codify’</td>
</tr>
<tr>
<td>mini</td>
<td>迷你 mínì</td>
<td>mì ‘charm’, nǐ ‘you’</td>
</tr>
<tr>
<td>hacker</td>
<td>黑客 hēikè</td>
<td>hēi ‘wicked’, kè ‘visitor’</td>
</tr>
<tr>
<td>Coca-Cola</td>
<td>可口可乐 kèkǒu kèlè</td>
<td>kèkǒu ‘ingestible’, kèlè ‘enjoyable’</td>
</tr>
</tbody>
</table>

Table 4.2 Examples of Simultaneous Phonetic and Semantic Adaptation

The main difference between the examples in Table 4.2 and phonetic adaptations is that the zi in these examples are used for both their sound value and their meaning, while the zi in phonetic adaptations are used just for their sound value. The primary difference between the words in the table and semantic adaptations is that these examples aim to convey the original pronunciation of the foreign term and its meaning, while semantic adaptations convey only its meaning.

The examples in Table 4.2 do not have a one-to-one correspondence between the morphemes in the English word and its Mandarin adaptation, but they do succeed in mapping the word onto strings of zi according to the original word’s syllable structure,
which helps to convey the rhythm of the word’s original pronunciation. The examples use the same number of Mandarin syllables in their adaptation as they have in English original. This is not always true in adaptations, however, and appears to be especially problematic for monosyllabic English words that include consonant clusters or word-final consonants. Such words tend to be adapted with two or more Mandarin syllables: *punk* is adapted as *péng-kè*, *waltz* as *huā-ěr-zī*, *Jeep* as *jí-pū*, *jazz* as *jué-shì*, *AIDS* as *ài-zī*, and so on. The consonant clusters and word-final obstruents of English syllables cannot be mapped onto individual Mandarin syllables, so more than one Mandarin syllable is needed to approximate the original pronunciation.

4.1.5 Separate application of processes

As noted at the beginning of this chapter, adaptation of foreign terms is usually done with a string of one or more Chinese *zì* that correspond to a parsing of either the word’s morpheme structure or its sound structure. Since each *zì* is pronounced as an independent syllable and can also be independent of the other *zì* in the string with respect to meaning, different adaptation processes can be used with different *zì* in the string. Thus, one part of the adapted borrowing may be a phonetic adaptation, while another part may be a semantic adaptation. For example, in *英*国 *yīngguó* ‘England’ lit. ‘*yīng*-country’, *yīng* is a phonetic adaptation, while *guó* ‘country’ is a loan translation. Similarly, *internet* is rendered as *因特网* *yīntèwǎng* lit. ‘*yīntè*-net’, in which the first two *zì* are phonetic adaptations and the third is a loan translation. The fact that each *zì* can be an independent linguistic unit allows for flexibility in the adaptation of foreign terms. The use of separate
adaptation processes for different parts of borrowed words is important for explaining the formation of lettered words as well, as I show later.

4.1.6 Addition of explicative adjuncts

The adaptation of foreign terms may involve the addition of an extra linguistic unit, as discussed by Cheung (1972), Chen (1999), T’sou (2001), Wang (2004), and others. I follow Hansell (1989) in calling the unit an explicative adjunct. The adjunct is a morpheme or word added to an adaptation that indicates the semantic category of the term. Addition of the adjunct helps to clarify the meaning of the term, particularly in the case of phonetic loans, but also in loan translations whose meaning is not evident from a literal translation of the term. For example, in 鸡尾酒 jīwěijiǔ ‘cocktail’ lit. ‘cock-tail-alcohol’, the literal meaning of jīwěi ‘cocktail’ in the context of the word is not transparent, so jiǔ ‘alcohol’ is used as an explicative adjunct to specify the semantic category of the term. Examples of phonetic adaptations with explicative adjuncts are shown in Table 4.3.

<table>
<thead>
<tr>
<th>English</th>
<th>Chinese</th>
<th>Phonetic Adaptation</th>
<th>Explicative Adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeep</td>
<td>吉普</td>
<td>吉普 jìpù</td>
<td>车 chē ‘vehicle’</td>
</tr>
<tr>
<td>khaki</td>
<td>卡其布</td>
<td>卡其 kāqí</td>
<td>布 bù ‘cloth’</td>
</tr>
<tr>
<td>waltz</td>
<td>华尔兹舞</td>
<td>华尔兹 huáèrzǐ</td>
<td>舞 wǔ ‘dance’</td>
</tr>
<tr>
<td>marathon</td>
<td>马拉松赛跑</td>
<td>马拉松 mǎlāsòng</td>
<td>赛跑 sài pǎo ‘running race’</td>
</tr>
</tbody>
</table>

Table 4.3 Examples of Phonetic Adaptations with Explicative Adjuncts
Most often the adjunct is added in the head position, as in the examples in Table 4.3. However, in some cases it is added as the modifier instead, as in 车胎 chētāi ‘tire’ lit. ‘car-tāi’ and 酒吧 jiǔbā ‘bar’ lit. ‘liquor-bā’, in which the modifiers 车 ‘car’ and 酒 ‘liquor’ are the adjuncts. Based on examples such as these, it appears that when a concept represented by a phonetic loan does not fit easily into a native Chinese semantic category that could serve as a head, an alternative way to clarify the term’s meaning is by adding a modifier representing a native semantic category instead. In other words, since it is difficult to classify a ‘tire’ and a ‘bar’ as types of Chinese things, these words are classified as items that at least pertain to ‘vehicles’ and ‘alcohol’ in Chinese, even if they are not types of vehicles and alcohol themselves.

4.2 Innovations in borrowing

The adaptation processes described above apply to certain kinds of lettered words as well. Although roman letters appear in place of Chinese zi, their use can be seen as an extension of the original borrowing processes that express the foreign word’s meaning, pronunciation, or a mixture of the two.

4.2.1 Extending phonetic adaptation

Individual roman letters have letter names in Mandarin that approximate the pronunciations of letter names in English, an issue I discuss in detail in section 4.8. One or more Mandarin syllables are used to approximate the English letter names, just as Mandarin syllables are used to approximate the original pronunciations of phonetic
adaptations. The effect with regard to lettered words is that letters in English words that can be pronounced as independent units with their letter names, that is, those that are not linked with other letters to pronounce the word or morpheme, are now dealt with differently in speech and writing. The original graphs are retained, so the letters are no longer sinicized with regard to their written form, but they are still sinicized with respect to pronunciation. Phonetic adaptation is used to convey the pronunciation of the letter as a string of one or more Mandarin syllables, just as in other phonetic adaptations. Examples are shown in 4.2.

4.2

<table>
<thead>
<tr>
<th>English original:</th>
<th>X ray T cell</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chinese adaptation:</td>
<td>X 光 T 細胞</td>
</tr>
<tr>
<td>Mandarin pronunciation:</td>
<td>ài kè sǐ guānɡ tí xì bāo</td>
</tr>
<tr>
<td>Gloss:</td>
<td>‘X’ ‘ray’ ‘T’ ‘cell’</td>
</tr>
</tbody>
</table>

The combination of the original written form and a phonetic adaptation in examples such as those in 4.2 represents a new twist on the traditional process of forming phonetic adaptations. Whereas both the pronunciation and the written form were sinicized in examples such as those in Table 4.3, now the pronunciation continues to be adapted as in the past, but the written form is imported directly. In contrast, the ‘spelled out’ parts of the original terms (ray, cell) are still adapted in a traditional manner as loan translations.

4.2.2 Eliminating semantic adaptation

Examples such as those in 4.2 show that while the pronunciation of letter names is sinicized with Chinese zì, the written form of the letters is not. I suggest that the meaning
of the morphemes represented by the letters is not sinicized either. The letter morphemes do not undergo semantic adaptation in the manner of the words in 4.1 and in Table 4.2. Rather, the meaning is left as is, and this is indicated in writing through the use of the original roman letter graphs.

I suggest several reasons why semantic adaptation is not used to adapt the letter morphemes. First, the meaning of the morpheme represented by the letter may already have been imported into Chinese, so that it actually exists as a productive morpheme in the language and does not need to be sinicized. For example, *The American Heritage Dictionary (4th edn.*) lists the following meaning for *A* as a graph: ‘the first letter of the English alphabet’, and it lists the following meanings for *A* as a letter morpheme: ‘the first in a series’, ‘the best in quality or rank’, ‘the sixth tone in the scale of C major’, and ‘a type of blood in the ABO system’. Both the graph *A* and the letter morpheme *A* have been imported into Chinese with these same meanings and therefore do not need to be sinicized. Chinese speakers associate the same meanings with *A* as English speakers and understand the various meanings of *A* in the context of its use.

Second, the meaning of the letter morpheme may not be known, or it may be either obvious or difficult to translate or describe. Translating or describing it may require multiple morphemes, making adaptation awkward (Sun and Jiang 2000). In *T恤* ‘T-shirt’, for example, the meaning of *T* may not have been known at the time of borrowing, in which case it may have been imported directly without an attempt to translate or describe it. Alternatively, since *T* is potentially being used as an iconic symbol in *T-shirt*, its meaning could be thought of as either too obvious or too difficult to translate or
describe, so that the letter morpheme would be retained as is. In X 光 X guāng, X originally referred to ‘an unknown quantity’, but this meaning may not have been known at the time of borrowing, or it may have required multiple morphemes to translate or describe so that it may have been easier to simply import X directly instead.

Third, as noted by Liu (2001), one of the chief attractions of lettered words is that they are short. (Liu appears to be referring to initialisms in particular.) In cases where letters are used as initialisms, translating the meanings of the words abbreviated by the letters would defeat the purpose of using the initialisms, since initialisms are used because they are shorter than the original. For example, translating the words abbreviated by DNA results in 脱氧核糖核酸 tuōyǎng hétáng hēisuăn ‘deoxyribo nucleic acid’, which has twice as many syllables as DNA (dī ēn ēi). Although abbreviations of zì-based compounds are common in Chinese (such as 北大 běidà for 北京大学 běijīng dàxué ‘Beijing University’) and provide an effective way of shortening compounds and phrases, initialisms are nonetheless perceived to be an effective way of creating short words and are preferred in some cases.

Finally, in the absence of full information about a foreign term, letters may simply be interpreted as graphs with the default meaning ‘the n\textsuperscript{th} letter of the English alphabet’. For example, although the letters in the names of aircraft such as A-320 and F-16 are initialisms, Mandarin (and English) speakers may think of them simply as ‘letters of the alphabet’ rather than as abbreviations. (This would, however, require psycholinguistic tests or some other type of linguistic investigation to confirm.) When viewed this way in Mandarin, the letters are somewhat like the phonetic characters 咖 kā/gā, 啊 lá, 卡 kǎ, etc.
that have the meaning ‘character used for transcribing foreign words’ rather than specific meanings as morphemes.

With regard to this uncertainty about the meaning of individual letters, I suggest that whether a roman letter is understood as a graph representing a letter of the alphabet or as an initialism or a roman letter morpheme with a contextual meaning (such as A) depends on one’s perspective as a Mandarin speaker rather than on the actual meaning of the letter in a particular term. For example, individuals knowledgeable about music know the contextual meaning of C in C 大调 ‘scale of C major’, but others may think of C simply as ‘the third letter of the alphabet’. Similarly, individuals who are knowledgeable about international trade most likely know that WTO is an initialism and, depending on their level of English fluency, may know the English words that the letters abbreviate, but other speakers may think of it simply as a string composed of the arbitrary letters <W>, <T>, and <O> to which a particular meaning has been assigned. This is similar to English speakers’ understanding of the initialism DHL, the name of an express mail company. I expect that most speakers do not realize that the word abbreviates the names of the founders of the company (Dalsey, Hillblom, and Lynn) but rather memorize it as a string of three arbitrary letters that has a particular meaning. In sum, whether speakers view roman letters simply as graphs representing the letters of the alphabet or as abbreviations or letter morphemes with contextual meanings depends on the understanding they have of the term rather than on the true meanings of the roman letters in particular lettered words.

4.2.3 Retaining separate application of processes
An important reason why the words in 4.2 have a mixed orthography is that different parts of the compound are adapted differently, which results in a hybrid written structure. Just as *internet* is parsed as *inter* and *net*, with *inter* adapted as the phonetic loan 因特 *yīntè* and *net* as the loan translation 網 *wǎng*, English compounds that contain one or more independent roman letters are also adapted differently. *Independent* roman letters are letters that can be pronounced with their letter names and function as initialisms, letter morphemes, phonetic characters, or symbols. They tend to be retained as is, and the rest of the compound is rendered into Chinese as a phonetic, semantic, or simultaneous phonetic and semantic adaptation. Examples are shown in Table 4.4.

<table>
<thead>
<tr>
<th>English</th>
<th>Chinese</th>
<th>Letters (Type)</th>
<th>Adaptation (Type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB gun¹</td>
<td>BB 枪</td>
<td>BB (letter morph.)</td>
<td>枪 qiāng ‘gun’ (loan translation)</td>
</tr>
<tr>
<td>vitamin A</td>
<td>维他命 A</td>
<td>A (letter morph.)</td>
<td>维他命 wéitānìng ‘vitamin’ (simultaneous adaptation)</td>
</tr>
<tr>
<td>ABC weapons²</td>
<td>ABC 武器</td>
<td>ABC (initialism)</td>
<td>武器 wǔqì ‘weapon’ (loan translation)</td>
</tr>
<tr>
<td>e-business</td>
<td>e 商务</td>
<td>e (initialism)</td>
<td>商务 shāngwù ‘business’ (loan translation)</td>
</tr>
</tbody>
</table>

¹ BB: ‘a small size of lead pellet used in air rifles’; ² ABC < a(tomic), b(iological), c(hemical)

Table 4.4 Separate Application of Adaptation Processes in Lettered Words

The examples in Table 4.4 all exhibit the same pattern: the independent roman letters are imported directly and the spelled-out parts of the compound have been converted into Chinese morphemes and characters. This may be a commonly occurring pattern in
Chinese. All of the examples of hybrid words in Xiàndài Hányǔ Cídiǎn ‘Dictionary of Modern Chinese’ and Xiàndài Hányǔ Guǐfàn Cídiǎn ‘Standard Dictionary of Modern Chinese’ follow the same pattern. My corpus study of hybrid words considers whether this is a common pattern in newswires as well.

How speakers interpret the meanings of the independent letters depends on their knowledge of what the letters represent. If speakers do not have any information about them, they may think of the letters simply as graphs meaning ‘the nth letter of the alphabet’. If they do have more information about them (for example, by looking them up in a dictionary), they may know that some are letter morphemes and others initialisms, and they may know what their the contextual meanings are or what the letters abbreviate.

4.2.4 Adding explicative adjuncts

The addition of explicative adjuncts to roman letter units is a useful way to provide a clue about their meaning in Chinese. The addition of adjuncts to roman letter units is analogous to adding explicative adjuncts to phonetic adaptations, which are also opaque with respect to meaning. Examples of roman letter units with added explicative adjuncts are shown in Table 4.5.
<table>
<thead>
<tr>
<th>English</th>
<th>Chinese</th>
<th>Letters (Type)</th>
<th>Explicative Adjunct</th>
</tr>
</thead>
<tbody>
<tr>
<td>PC¹</td>
<td>PC 机</td>
<td>PC (initialism)</td>
<td>机 jī ‘machine’</td>
</tr>
<tr>
<td>MS²</td>
<td>MS 病</td>
<td>MS (initialism)</td>
<td>病 bìng ‘disease’</td>
</tr>
<tr>
<td>T-shirt</td>
<td>T 衫</td>
<td>T (letter morph. ) (恤 xù: phonetic loan for shirt)</td>
<td>衫 shān ‘upper garment’</td>
</tr>
<tr>
<td>GRE³</td>
<td>GRE 考试</td>
<td>GRE (initialism)</td>
<td>考试 kǎoshì ‘exam’</td>
</tr>
</tbody>
</table>

¹PC < p(ersonal) c(omputer), ²MS < m(ultiple) s(clerosis), ³GRE < G(raduate) R(ecord) E(xamination)

Table 4.5 Lettered Words Containing Explicative Adjuncts

The examples in Table 4.5 function as independent words in English, but they are opaque when imported into Chinese. The addition of an explicative adjunct helps to sinicize the words and the concepts they represent by pigeonholing them into native Chinese semantic categories. In the examples of hybrid words in Xiàndài Hànyǔ Cídiǎn ‘Dictionary of Modern Chinese’ and Xiàndài Hànyǔ Guīfān Cídiǎn ‘Standard Dictionary of Modern Chinese’, for example, the adjuncts are added as heads, as in the examples in Table 4.3, rather than as modifiers, as in 车胎 chētāi ‘tire’ lit. ‘car-tāi’ and 酒吧 jiǔbā ‘bar’ lit. ‘liquor-bā’. My corpus study of lettered words considers this issue as well by examining whether explicative adjuncts in hybrid words are added primarily as modifiers or as heads.

4.3 Native creations

4.3.1 Native creations based on English
The previous sections described the adaptation of lettered words imported into Chinese from English. The same principles apply to lettered words that could be considered *native creations* in Haugen’s (1950) and Winford’s (2003) typologies of lexical contact phenomena, that is, new compounds created with native and foreign morphemes or through the creative use of foreign morphemes. The word formation principles in section 4.1 apply to the formation of native creations in the same way as to the adaptation of foreign words. Examples are shown in Table 4.6.

<table>
<thead>
<tr>
<th>Coinage</th>
<th>Gloss</th>
<th>Letters (Type)</th>
</tr>
</thead>
<tbody>
<tr>
<td>BB</td>
<td>‘baby’</td>
<td>BB (phonetic characters)</td>
</tr>
<tr>
<td>B 超</td>
<td>B chāo¹ ‘ultrasound’ (lit. ‘B-wave’)</td>
<td>B (letter morpheme)</td>
</tr>
<tr>
<td>A 拷</td>
<td>A kào ‘first copy of a videotape’ (lit. ‘A-copy’)</td>
<td>A (letter morpheme)</td>
</tr>
<tr>
<td>e 化</td>
<td>e huà ‘e-ize’ (e &lt; electronic)</td>
<td>e (initialism)</td>
</tr>
<tr>
<td>STS 教育</td>
<td>STS jiào yù ‘STS education’</td>
<td>STS (initialism)</td>
</tr>
<tr>
<td>AA 制</td>
<td>AA zhì ‘AA system’ (each person pays their own share of a meal)</td>
<td>AA (letter morphemes)</td>
</tr>
</tbody>
</table>

¹‘ultrasonic diagnosis B’, ²STS < s(cience), t(chnology), s(ociology)

Table 4.6 Roman Letter Native Creations

In the examples in Table 4.6, the roman letter component of native creations is composed of individual roman letters that can be pronounced with letter names and that may have independent meanings. Spelled out morphemes or words may be used much less often, and the component that would need to be spelled out as an English word may be expressed with Chinese morphemes instead. This is an issue that I investigate in my corpus study. The examples in the table also show that roman letters in native creations
have the same functions as in adapted borrowings of the type in Tables 4.4 and 4.5. They are used as letter graphs, letter morphemes, and initialisms. In cases where roman letters are used as letter morphemes, they have the same meanings as in English, since these meanings have been borrowed into Chinese. (So A in A kǎo means ‘first’, just as A does in English.)

4.3.2 Native creations based on pīnyīn romanization

Judging from the lettered words in Xiàndài Hànyǔ Cídiǎn ‘Dictionary of Modern Chinese’ and Xiàndài Hànyǔ Guīfàn Cídiǎn ‘Standard Dictionary of Modern Chinese’, most native creations appear to be based on foreign words or morphemes, especially ones from English, but some have their basis in Chinese pīnyīn romanization. These are used in the PRC more often than in Taiwan, since Taiwan generally does not use pīnyīn romanization for phonetic annotation. Rather than abbreviating English words or morphemes, roman letters abbreviate morphemes written in pīnyīn romanization. Examples include RMB (< r(én) m(ín) b(i) 人民币), the name of the Chinese currency (lit. ‘people’s currency’), YZ (< y(ing) z(uō) 硬座) ‘hard seat’ (on a train), and RW (r(uān) w(ò) 软卧) ‘soft berth’. The letters are also used to express romanized abbreviations of compounds and phrases. The romanized abbreviations do not necessarily correspond to legal native abbreviations in Chinese zì, however, as shown in 4.3.

4.3

国家标准 guójìā biāozhǔn ‘national standard’
国家标准 > 国标
guójìā biāozhǔn > guóbiǎo > GB OR
guójìā biāozhǔn > GB
The abbreviation GB in 4.3 could be interpreted as an initialism formed from the zi-based abbreviation guóbiāo, formed by using the first morpheme in the compounds guójiā ‘nation(al)’ and biǎozhǔn ‘standard’, but it could also be interpreted as an initialism that abbreviates the full compounds (guójiā and biǎozhǔn). It is the latter pattern that is probably used to form HSK. HSK is most likely not derived from a zi-based abbreviation since, according to several Beijing language authorities I queried, the abbreviation for hàn yǔ shuǐ ping kāoshi ‘Mandarin proficiency is hàn kāo rather than hàn shuǐ kāo, since, according to their explanation, disyllabic abbreviations are preferred. Thus, in HSK, the first letter of each word in the three-word compound is most likely used to abbreviate the compound. This pattern of forming initialisms is exactly like the pattern in English and other Western languages of using the first letter in each of the words in the compound or phrase to create abbreviations (e.g. WTO). It would be useful to investigate whether this pattern of abbreviation was borrowed from English or other Western languages. Chinese lexicographers may know how the pattern came to be used in Chinese.

Based on examples from Liu (1999) and Gao (2007), it appears that pīnyīn-based lettered words may tend to be initialisms rather than fully-spelled words. For example, kinship terms such as 哥哥 ‘elder brother’ and 弟弟 ‘younger brother’ are written in computer-mediated communication (CMC) as the romanized abbreviations GG and DD but rarely as their fully-spelled romanized forms gēge and didi. One motivation for this is
surely the desire to abbreviate frequently occurring written forms as much as possible, which English speakers do as well in CMC. Another, more insightful explanation may be that writers reanalyze spelling and writing with pīnyīn letters to be more like spelling and writing with Chinese zì. Rather than writing fully-spelled words in pīnyīn, writers may abbreviate pīnyīn spellings so that they essentially become strings of zì rather than strings of letters connected into orthographic words. Whereas gē and dìdī require a “Western” way of writing, that is, joining letters together into orthographic words, GG and DD abbreviate the pīnyīn spellings to write the morphemes as two discrete roman letter zì, each of which fits into the imaginary equidimensional square of Chinese writing.

4.4 Roman letter modifier, Chinese zì head

An important characteristic of the examples in Tables 4.4 to 4.6 is that the modifier is expressed in roman letters and the head in Chinese zì. This pattern has the critical semantic function of categorizing foreign concepts into native Chinese semantic categories, which is an important way of “lining up secure similarities [as an] anchor for going into the uncertainties” (Anttila 2003:425). For example, a PC jī ‘PC’ is interpreted as a type of jī ‘machine’ rather than as an entirely foreign category of items with no Chinese semantic equivalent (i.e. PC). BB gun is adapted as BB qiāng. It is categorized in the native category qiāng ‘gun’ rather than being imported directly as the foreign item BB gun that would not necessarily be linked with the native category qiāng ‘gun’. This process is illustrated in 4.4. with English words for types of guns that are interpreted as types of qiāng in Mandarin.
Example 4.4 shows three types of adaptations of English terms for types of guns: 
láifuqiāng ‘rifle’, a phonetic adaptation with an explicative adjunct; jīqiāng ‘machine gun’ and shǒuqiāng ‘hand gun’, both of which are loan translations, and BB qiāng, in which the letters BB are imported directly as letter morphemes and gun is translated as qiāng ‘gun’. Qiāng is the hypernym, and the various types of guns are its hyponyms. In each case, the foreign item is added to the qiāng ‘head family’, a term I borrow from Krott and Nicoladis’s (2005) study of children’s parsing of compounds. The new compounds have qiāng as the head and an element expressing a new ‘type of qiāng’ as the modifier. Example 4.4 shows how the innovative letter-based adaptation BB qiāng fits neatly into the Chinese semantic space. Its modifier-noun structure is essentially no different from that of other types of qiāng, except that its modifier is expressed with roman letter morphemes rather than Chinese zì.

The ‘roman letter modifier-Chinese zì head’ structure of many lettered words has the additional benefit of specifying the classifier that is to be used with the word. Classifiers (CL) are obligatory elements that go before nouns in Mandarin noun phrases. They occur in phrases that use a number (e.g. 一个人 yī ge rén ‘one CL person’), a demonstrative (e.g. 这个人 zhèi ge rén ‘this CL person’), or certain quantifiers (每个人 měi ge rén
‘every CL person’). Mandarin has several dozen classifiers, listed in Chao (1968a).

Speakers must memorize which classifiers are paired with which types of nouns. The Chinese ｚｉ head in hybrid words such as ＢＢ ｑｉāｎｇ takes the guesswork out of determining which classifier is to be paired with a roman-letter based noun. This is illustrated in 4.5 by comparing PC 机 ‘PC (machine)’ with PC.

4.5

<table>
<thead>
<tr>
<th>这台 PC 机 vs. 这台 PC</th>
</tr>
</thead>
<tbody>
<tr>
<td>this CL PC machine</td>
</tr>
<tr>
<td>‘this PC’</td>
</tr>
<tr>
<td>这台 PC</td>
</tr>
<tr>
<td>this CL PC</td>
</tr>
<tr>
<td>‘this PC’</td>
</tr>
</tbody>
</table>

Example 4.5 shows that one drawback of lettered words without a Chinese ｚｉ head is that one does not immediately know which classifier is to be used with the word. Speakers must know the meaning of the word to be able to determine its Chinese classifier. One must know what a ＰＣ is, or more specifically, that it is a type of ｊī ‘machine’, in order to assign it a proper classifier. In contrast, if the word has a Chinese ｚｉ head, either via loan translation (as in ＢＢ ｑｉāｎｇ) or through the addition of an explicative adjunct (as in ＰＣ ｊī), the classifier is determined by the Chinese ｚｉ head. Speakers simply need to know the meaning of the Chinese ｚｉ head to assign the word a classifier; they do not need to know the meaning of the modifier to do so.

**4.5 Tendencies in the formation of lettered words**

Based on my examination of lettered words, I propose that there are two tendencies in the types of lettered elements that are borrowed into Chinese and in the structure of lettered words. First, many, possibly most, borrowed and natively formed lettered words
are like those in the tables above, that is, initialisms and hybrid words containing independent letters, rather than fully-spelled words. English *fashionista* and *tankini*, for example, are unlikely to be imported directly in speech or writing and would most likely be sinicized as phonetic or semantic adaptations through the use of Chinese *zi*. English *SUV* and *MP3*, on the other hand, are likely to be imported directly into Chinese and may or may not have sinicized equivalents.

A second tendency is that in the borrowing and creation of compounds containing an initialism or a letter morpheme, this component is used directly, while the other component is expressed in Chinese *zi*. So, *ATM machine*, for example, was adapted as *ATM机* *ATM jī* ‘ATM machine’, which retains the roman letter morphemes in the initialism but translates *machine* as the morpheme *机 jī* ‘machine’. Another adaptation of *ATM machine* is the semantic translation *提款机* *tīkuānjī*, lit. ‘withdraw-money-machine’, which is used frequently in Mandarin. An unlikely alternative to either of these adaptations would be *提款 machine*, in which the modifier is expressed with Chinese *zi* (*提款*) and the head (*machine*) with roman letters. The adaptation *提款 machine* is made even more unlikely because the head is a spelled out English word rather than an initialism or a letter morpheme. In sum, the unacceptability of *提款 machine* reflects what may be a strong preference for using the ‘roman letter modifier – Chinese *zi* head’ pattern to form hybrid compounds, and also for using roman letter elements that are not spelled out words. This issue is also investigated in my corpus study.
4.6 Congruence between Chinese zì and ‘single letter units’

The innovative outcomes of Chinese-English lexical contact that employ roman letter units in place of Chinese zì are possibly due to congruence, or typological fit, between Chinese zì and certain kinds of linguistic units in English. Congruence is discussed with regard to code switching in studies of language contact (e.g. Sankoff and Poplack 1981, Muysken 1997, Myers-Scotton and Jake 2001), which focus primarily on describing the alignment of grammatical categories in utterances involving code switching. I suggest, however, that the notion also applies to explaining the acceptability of lettered words in Chinese. Muysken (1997:362) describes the concept of congruent lexicalization, a situation “where two languages share a grammatical structure which can be filled lexically with elements from either language.” If we take the view that Chinese utterances consist of strings of hierarchically ordered Chinese zì, the “slots” in which roman letter elements can be used in place of Chinese zì are the points of congruence between the two languages. These slots are now filled lexically with elements from either Chinese or English, and they are increasingly being filled with certain kinds of roman letter-based elements from English. Partially filled [M N] compounding patterns such as [__ qiāng] ‘__ gun’, which contain an open modifier slot and a Chinese zì head from a particular head family, are one of the main examples of a grammatical structure that is now being filled lexically with a roman letter-based element.

It should be noted that the notion of ‘congruence’ involves controversies about whether the grammatical/structural borrowing of linguistic material will occur only between congruent grammatical systems. Thomason and Kaufman (1988) and Friedman
and Joseph (To appear: Ch. 3.2.1.7) argue against this, suggesting that given the appropriate social circumstances, any element of one language can be borrowed into another. I concur with this assessment and want to emphasize that the “open slots” mentioned above lend themselves to being filled either with Chinese \( \text{zi} \) or with certain kinds of English units. This does not mean, however, that Chinese cannot borrow from English aspects that are not congruent with Chinese. I am claiming here simply that \( \text{zi} \) and certain kinds of English units are congruent as both grammatical and orthographic units; I am not claiming that non-congruent structures cannot be borrowed.

The English units that can fill slots in strings of Chinese \( \text{zi} \) are what I refer to as *single letter units*. These units are single letters of the English alphabet used in speech (#1–3) and writing (#1–8) as in Table 4.7.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>graph meaning ‘the ( n )th letter of the alphabet’ (( ABCs ))</td>
</tr>
<tr>
<td>2.</td>
<td>morpheme (( A)-list (A ‘top-ranked’))</td>
</tr>
<tr>
<td>3.</td>
<td>initialism (( PG ) (( PG &lt; ) parental guidance [movie rating]))</td>
</tr>
<tr>
<td>4.</td>
<td>syllable (( EZ &lt; ) easy)</td>
</tr>
<tr>
<td>5.</td>
<td>rebus symbol (( K9 &lt; ) canine)</td>
</tr>
<tr>
<td>6.</td>
<td>icon (( Z = ) indicates ‘sleep’, ( W = ) ‘George W. Bush’)</td>
</tr>
<tr>
<td>7.</td>
<td>mathematical symbol (( x = ) multiplication symbol)</td>
</tr>
<tr>
<td>8.</td>
<td>roman numeral (( V = 5 ))</td>
</tr>
</tbody>
</table>

Table 4.7 Single Letter Units in English

In certain instances, single letters can be combined to indicate intensity or length (e.g. \( ZZZZ \), which indicates ongoing sleep), or they can be combined on a Likert-type scale to indicate degree (e.g. \( AAA \) bonds). Although the linguistic use of single letter units as
morphemes and initialisms should be separated from their non-linguistic use as symbols in writing, both types of single letter units are congruent with Chinese zì in one or more ways. They can therefore be substituted for zì in either writing or speech. Quoting Herskovits (1938, 5f.) again, it is in single letter units that Chinese speakers have found “something understandable in terms of their own patterns,” which increases acceptance. These ways of using letters are much more readily assimilated into Chinese than the use of letters as graphs to spell orthographic words and may be more functional in the language. I suggest that it is the strong similarities between single letter units and Chinese zì that most likely explain the frequent use of single letter units in Chinese.

Single letter units share one or more of the following characteristics of Chinese zì that make them amenable to use as replacements for zì. First, each letter can be pronounced independently of any surrounding letters just as each Chinese zì is pronounced separately from surrounding zì. Further, single letter units can be pronounced with their Chinese letter name, which is one syllable in length for 19 of the 26 letters. The pronunciations of Chinese zì are similarly one syllable in length. Second, the single letter units may have individual meanings, as shown in Table 4.7, in the same way as Chinese zì have individual meanings. Finally, single letter units function as independent graphs, taking up one equidimensional square in the manner of Chinese characters. Example 4.6, adapted from Hansell (1989), illustrates these points.

4.6

<table>
<thead>
<tr>
<th>A</th>
<th>餐</th>
<th>B</th>
<th>餐</th>
</tr>
</thead>
<tbody>
<tr>
<td>ēi</td>
<td>cān</td>
<td>bī</td>
<td>cān</td>
</tr>
</tbody>
</table>

‘meal A’  ‘meal B’  (set meals in a restaurant)
The single letter units $A$ and $B$ in $A$ and $B$ have the same properties as the morpheme $cān$ ‘meal’, written as the character 餐. They are pronounced with their letter names as Mandarin syllables ($ēi, bī$); each fills one equidimensional square; and each has an independent meaning, with $A$ designating the first of the two set meals (say, chicken), and $B$ designating the second (say, fish).

4.7 Roman letter zi

What I have called ‘single letter units’ in English have been incorporated into Chinese as a new set of zi in the language, roman letter zi (RL zi). Hansell (1989) argues for a similar view with his concept of the Sino-alphabet, which he defines as roman letters that have been assimilated into written Chinese so that “no special rules are needed to use them.” Hansell suggests that the letters are assimilated by “making the individual graphs graphically and phonologically discrete and by allowing free directionality” (p. 170).

My view of RL zi goes beyond the Sino-alphabet. I suggest that letters of the English alphabet have not simply been added to Chinese as a new set of graphs that are individually pronounceable and have free directionality. Rather, the meanings of letters as morphemes and symbols, shown in Table 4.7, have also been incorporated into Chinese together with their use of as graphs. RL zi have been added to the existing stock of Chinese morpheme-syllable-characters as an additional resource in the language, albeit one with a more limited set of functions. Just as each Chinese zi has an associated
pronunciation, meaning, and graph, RL zì also have these same defining characteristics. Practically speaking, RL zì could each have an entry in Chinese zìdiān ‘zì dictionaries’ listing their pronunciation, graph, and most of the same definitions that are listed for single letters in English dictionaries. All that would be missing would be an indication that individual letters have the additional important function of abbreviating spelled words in English (and other Western languages) and in pīnyīn romanization.

Roman letter zì and their uses represent an important point of congruence between Chinese and English grammar, as well as between the two writing systems. Chinese exploits this congruence by favoring the use of RL zì over the more ordinary use of letters as graphs to spell words. Some functions of RL zì are the following: 1) initialisms (NBA, composed of three RL zì), 2) zì with contextual meanings (in 陈 X ‘X Chen’, X means ‘unknown quantity’ and anonymizes the individual’s name), and 3) as phonetic characters (K他命 kèitāming ‘ketamine’, a performance-enhancing drug used by athletes). (K他命 is possibly formed by analogy with 维他命 wěitāming ‘vitamin’. The letter K, pronounced as the syllable [kʰei], may be substituted for the Chinese zì 维, pronounced in Mandarin as the syllable wéi). Such uses of letters are considered marginal in English as compared with their ordinary use in simply spelling words, but it is these very uses that may form the point of congruence between Chinese and English and that appear to be assimilated and built upon most easily in Chinese. My corpus study tests whether these are the uses that appear most frequently in Chinese newswires rather than the use of letters to spell words.
4.8 Pronunciation of roman letter zi in Mandarin

4.8.1 Initialisms vs. acronyms

An important characteristic of RL zi is that they can be pronounced as discrete phonological units. Usually each RL zi is pronounced with its respective letter name, whose pronunciation is familiar to average speakers of Mandarin. Mandarin letter names are approximations of English letter names. In initialisms, that is, abbreviations formed with the first letters of words or phrases and pronounced letter by letter, such as GDP, each of the RL zi is pronounced separately with its Mandarin letter name, ji di pi. The pronunciation of what English speakers treat as acronyms, that is, abbreviations formed with the initial letter or letters of words or phrases and pronounced as a word, such as SARS, RAM, and FED, is potentially more complex. The vast majority of acronyms in Mandarin are borrowed rather than natively coined. Since non-native spelling-to-pronunciation conventions are needed for their formation, one has to be at least marginally bilingual in English and Mandarin to create them in Mandarin, so that not many natively formed lettered acronyms currently exist.

Speakers approximate the English pronunciation of acronyms borrowed from English using one of three basic strategies. First, if the word has a Chinese zi equivalent in the form of a phonetic or semantic adaptation, the pronunciation of the Chinese zi equivalent is used to pronounce the lettered word. For example, TOEFL can be pronounced as its Chinese zi equivalent 托福 tuòfú. Second, if there is no Chinese zi equivalent, speakers pronounce the word as they hear it pronounced by others (e.g. SARS is pronounced as [sà.sz]). This of course depends on actually hearing the word pronounced by others,
which will not necessarily be the case for many lettered words since a wider variety of
them is used in print than in speech. Finally, if speakers are familiar with English, they
can construct the acronym’s pronunciation based on their knowledge of the spelling and
pronunciation conventions of English.

If Mandarin speakers do not know that a particular initialism is pronounced as a word,
they pronounce it letter by letter, creating a plausible alternative to the acronymic
pronunciation. This strategy is used by English speakers as well when they do not know
whether an initialism is pronounced as an acronym. The correct pronunciation of
initialisms is established by convention rather than by rule. For example, the American
standardized college entrance exams SAT and ACT could be pronounced as acronyms
([sæt] and [ækt]), but by convention both words are pronounced as initialisms. To take
another example, my university has a division called Faculty and TA Development,
abbreviated as FTAD. I have heard the abbreviation pronounced as both an initialism and
as the letter name-plus-word combination [eftæd]. The acronymic pronunciation is used
within FTAD, while the letter-by-letter pronunciation appears to be used by outsiders
new to FTAD.

The English conventions for pronouncing initialisms either letter by letter or as
acronyms appear to have been carried over into Mandarin, as well. Either a word-based
or a letter-based pronunciation is acceptable. In most cases, it depends on what has been
imported from English, since most acronymic pronunciations are imported from English.
If Mandarin speakers do not know whether an initialism is pronounced as an acronym or
if they do not know the acronymic pronunciation, they will tend to pronounce the initialism letter by letter, which is also true of English speakers.

An additional point I want to stress here with regard to the pronunciation of initialisms as acronyms is that Mandarin speakers either have to hear the acronym pronounced by others and then imitate it, which does not require a knowledge of English, or they have to figure out the acronymic pronunciation themselves, which requires a knowledge of the sound-to-spelling rules of English. Aside from the rules of pīnyīn romanization, there are no native sound-to-spelling rules for pronouncing strings of letters that spell words. Rather, the only pronunciation convention that has so far been firmly nativized in Mandarin with regard to roman letters is the pronunciation of individual letter names. As a result, the pronunciation of strings of letters in which the letters can be pronounced with their individual letter names is much easier to negotiate for Mandarin speakers who do not know English than strings of letters that must be pronounced as words. If non-English speaking Mandarin speakers encounter the words VIP and glitterati in a Chinese text, they will most certainly have an easier time pronouncing the former than the latter.

4.8.2 Letter name pronunciations

Letter name pronunciations in Mandarin generally approximate the pronunciations of letter names in English (primarily in American English). Although roman letters have official letter name pronunciations in the PRC that are given in dictionaries using zhùyīn fūhào ‘Chinese phonetic symbols’, Chinese friends have informed me that these pronunciations are generally not used to pronounce roman letters that represent RL zì. An
alternate set of letter name pronunciations modeled on English pronunciations of letter names has been conventionalized for this purpose. The approximate pronunciations are given in Table 4.8.

<table>
<thead>
<tr>
<th>Roman letters</th>
<th>Pronunciation (in IPA)</th>
<th>Roman letters</th>
<th>Pronunciation (in IPA)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A, a</td>
<td>[œi]</td>
<td>N, n</td>
<td>[œn]</td>
</tr>
<tr>
<td>B, b</td>
<td>[pɪ]</td>
<td>O, o</td>
<td>[œu]</td>
</tr>
<tr>
<td>C, c</td>
<td>[cɪ] or [sɪ]</td>
<td>P, p</td>
<td>[pʰti]</td>
</tr>
<tr>
<td>D, d</td>
<td>[tɪ]</td>
<td>Q, q</td>
<td>[kʰiau]</td>
</tr>
<tr>
<td>E, e</td>
<td>[jɪ]</td>
<td>R, r</td>
<td>[er]</td>
</tr>
<tr>
<td>F, f</td>
<td>[œ.fu]</td>
<td>S, s</td>
<td>[œ.sz]</td>
</tr>
<tr>
<td>G, g</td>
<td>[tœi]</td>
<td>T, t</td>
<td>[tʰti]</td>
</tr>
<tr>
<td>H, h</td>
<td>[ai.tœhjy]</td>
<td>U, u</td>
<td>[jœu]</td>
</tr>
<tr>
<td>I, i</td>
<td>[ai]</td>
<td>V, v</td>
<td>[wei]</td>
</tr>
<tr>
<td>J, j</td>
<td>[tœi]</td>
<td>W, w</td>
<td>[tœ.po.ju]</td>
</tr>
<tr>
<td>K, k</td>
<td>[kʰœi]</td>
<td>X, x</td>
<td>[ai.kʰa.sz]</td>
</tr>
<tr>
<td>L, l</td>
<td>[œ.lu]</td>
<td>Y, y</td>
<td>[waï]</td>
</tr>
<tr>
<td>M, m</td>
<td>[œ.mu]</td>
<td>Z, z</td>
<td>[tsœ]</td>
</tr>
</tbody>
</table>

Table 4.8 Pronunciation of Letter Names in Mandarin

The pronunciations in Table 4.8 are based on my recollection of the pronunciations of letter names by Beijing speakers and the pronunciations of my colleagues from China. They provide a general sense of how letter names are pronounced in Beijing Mandarin and possibly Beijing-influenced Mandarin. Although letter name pronunciations have been conventionalized in Mandarin, they have not been codified and are not given in authoritative dictionaries of Mandarin. Similarly, lettered words are listed as entries in an
appendix in Mandarin dictionaries, but their pronunciations are not given. Speakers either know the pronunciations from having heard them, or they have to ask others — they cannot consult a Chinese dictionary to find them. In part because speakers are “on their own” in this regard with only other speakers to rely on for figuring out conventionalized pronunciations of letter names and lettered words, the pronunciation of letter names and lettered words appears to vary depending on a number of factors, which may include speakers’ familiarity with English, pronunciation conventions in particular regions of the Chinese-speaking world, and pronunciations specific to particular lettered words. The pronunciation of letter names may also be influenced by regional varieties of Chinese.

Mandarin speakers normally use one of several strategies for approximating the pronunciation of English letter names (Riha 2006b). In cases where the English letter name corresponds to the pronunciation of a Mandarin syllable, that syllable’s pronunciation is adopted as the letter name. The pronunciation of 18 of the letter names is created this way: A, B, C, D, E, G, I, J, N, O, P, Q, R, T, U, V, Y, Z. A is pronounced as the Mandarin syllable əi [əi], B as bī [pi], and so on. In the case of F, L, M, and S, [ε] is epenthesized before the Mandarin syllables fu [fu], lu [lu], mu [mu], and si [sz] to give approximations of the English pronunciation. The epenthized vowel appears to be an approximation of the English vowel [ε] in the English letter name pronunciations, while the Mandarin syllables that follow the epenthized vowel constitute native syllable substitutions for the remaining part of the English letter name.
In $H$ two Mandarin syllables are joined together, and in $W$ and $X$ three syllables are joined together, to approximate the English letter name pronunciation for each letter. This is necessary due to phonological constraints in Mandarin, which does not allow consonant clusters or syllable final consonants except /n/ and /ŋ/. Consonant clusters in these letter names are divided up into separate syllables in which the consonant becomes the onset (e.g. [eks] --> [ai.kʰə.sz]). Syllable-final consonants in English letter names are similarly repositioned to become onsets (e.g. [eitʃ] --> [ai. tʃə]). In the pronunciation of $W$, the English syllable [dʌ] is approximated with the Mandarin syllable da [ta], the syllable [bɔl] is approximated with [pɔ], which is somewhat like the Mandarin syllable bo [puə], and [ju] is pronounced either approximately as [ju] or somewhat like the Mandarin syllable you [jɔu].

The letter names for $C$, $K$, $Q$, and $Z$ tend to be pronounced with innovative combinations of Mandarin initials (initial consonants) and finals (rhymes). For example, $C$ is pronounced by some speakers as [sɪ] and by others as [sɿ], in which the Mandarin initial [s] is combined with the final [ɿ]. $K$ is pronounced as [kʰei], in which the initial [kʰ] is combined with the final [ei]. (The syllable [kʰai] exists in Standard Mandarin, but not [kʰei]. This syllable does exist in certain other varieties of Chinese, however, and it is possible that it is borrowed to pronounce the letter name for $K$.) Innovative syllables such as these combine initials and finals into combinations that are not legal in Mandarin.
phonology and do not exist as possible syllables in the standard language. Such innovative syllables are also used in pronouncing English acronyms.

There is also the issue of tone in the pronunciation of roman letter names. In Table 4.8 I indicate that each single-syllable letter name is pronounced with a high level tone, shown as a macron over the main vowel in each IPA transcription. It is my impression that monosyllabic letter names are usually pronounced with a high level tone on the syllable, and I have marked this as their ‘lexical tone’. As for the letter names of F, H, L, M, S, W, and X, my impression is that the first syllable of each letter name is pronounced with a high level tone but that the subsequent syllables may be pronounced with a neutral tone. I have left the tone of the subsequent syllables unmarked. Coincidentally, the pronunciations of Cantonese monosyllabic letter names in Bauer and Benedict (1997) are given with a high level tone as well. The letter name pronunciations they provide are also similar to those in Table 4.8.

In subsequent sections and chapters, I provide romanized pronunciations of Chinese morphemes but do not provide the pronunciations of roman letters since they are given in Table 4.8. When indicating pronunciation, I simply write the roman letters with the understanding that they are to be pronounced as in the table. The pronunciation of Mandarin morphemes is given in italics. The four tones of Mandarin are indicated over the main vowel of each syllable in the customary manner of pīnyīn romanization. A macron is used for the high level tone; an acute accent for the rising tone; a hacek for the dipping tone; and a grave accent for the high falling tone. For example, the pronunciation of a hybrid word such as PC 机 ‘PC’ lit. ‘PC-machine’ is given as PC jī. P and C are
pronounced as in Table 4.8, and the pronunciation of the Mandarin morpheme 機 ‘machine’ is given in romanized form as jī. Neutral tone syllables are indicated by the lack of a diacritic.

4.9 Going beyond the familiar

Thus far, I have stressed the ways in which Roman letter zì are similar to Chinese zì with respect to their pronunciation, meaning, and written form. Their congruence with Chinese zì encourages their assimilation into Chinese. The process of integrating roman letters into Chinese does not simply involve the adoption of those aspects of roman letters and their uses that are similar to those of Chinese zì, however. Speakers are also able to go beyond something “understandable in their own terms” to incorporate dissimilar structures into the language as well. In addition to using English-based roman letter elements that are congruent with those in Chinese, there is also a sense of going beyond the familiar to adopt structures that are less zì-like and more English-like as well. At the level of phonology, this involves the adoption of novel phonemes, phoneme combinations, and syllables. Such innovative combinations are already used in the pronunciation of the certain letter names, as shown in the previous section, and in the pronunciation of some English acronyms. At the level of morphology, new structures are being introduced, such as the rules of abbreviation that apply to pīnyīn letter-based initialisms. In addition, although the ‘RL zì modifier - Chinese zì head’ structure tends to dominate hybrid compound formation, as certain lettered words expressed with RL zì become increasingly frequent, they will most likely also begin to serve as heads. Finally,
spelled out English words are increasingly being integrated into Mandarin. As societal familiarity with English in both the PRC and Taiwan continues to increase, speakers are gaining enough fluency in English to accept more ordinary English words, that is, spelled words, that are not congruent with Chinese zì and words formed with Chinese zì. Although RL zì will continue to be used frequently in lettered words because they are highly functional in Chinese, the adoption of a wider variety of spelled English words is also increasing. If current trends continue, we can look forward to the day when not just VIP is used by average Mandarin speakers in speech and writing, but glitterati will be as well.
5.1 Rationale

5.1.1 Selecting a corpus

Prior to embarking on the current study, in Riha (2006a) I analyzed the structure of about one hundred lettered words found in dictionaries, newspapers, and literature on new words in Mandarin. My aim was to conduct a preliminary study of lettered words to identify patterns in their morphology and semantics. Neither I nor the linguists whose writings on lettered words I consulted (Kang 1999, Wang 2004, Li 2005, etc.) had conducted a detailed examination of a large number of lettered words, nor did there exist a study of the development of lettered words over time or studies of lettered word use in different Chinese societies. To better understand the role of lettered words in Mandarin, it was necessary to conduct a larger and more systematic study of these words to investigate the range of possible patterns in their formation, including how the Chinese and English scripts were combined, changes in types of lettered words over time, and differences between the PRC and Taiwan, two Chinese societies with Mandarin as their national language.
I determined that a corpus study of written texts would enable me to conduct a meaningful comparison of a large number of lettered words used in the PRC and Taiwan. I wanted to be certain to choose an appropriate corpus to conduct such a comparison. The considerations that helped focus my search for a corpus are summarized in Dell Hymes’ (1974) model for analyzing the components of communication. The framework characterizes communicative events in terms of dimensions such as the setting, participants, and so on, which informally guided my choice of a corpus.

Any set of texts that I decided to analyze are the products of communicative events. Thus, I wanted to select a corpus whose components of communication were familiar to me so that I could interpret the language of the texts correctly. Equally important, I wanted the corpora to be comparable in PRC and in Taiwan to allow for an “apples to apples” comparison of lettered words contained in them.

5.1.2 Focus on Standard Mandarin

I also wanted the texts to be representative of Standard Mandarin used in both societies rather than the Mandarin of specific social groups, such as the young people in the studies of computer-mediated communication conducted by Su (2003) and Gao (2007). My interest is in identifying patterns of lettered word usage and types of lettered words that are so common and frequent in Mandarin that they have made their way even into the standard language. From a practical standpoint, the corpus also needed to be transparent, or understandable, to me as the analyst, and Standard Mandarin is the variety with which I am most familiar. In researching Chinese computer-mediated
communication, for example, I found that media such as blogs and chatrooms included a considerable amount of in-group jargon that I, as an outsider, could not interpret.

I determined that a corpus of newswires from the PRC and Taiwan would be an appropriate choice based on my criteria. Newswires are a good choice because they are written for a broad audience and are representative of Standard Mandarin in the two Chinese societies. As such, they are transparent and accessible to me as an analyst and represent the type of Mandarin of interest in my study.

5.2 Corpus

5.2.1 Chinese Gigaword Third Edition

I chose the Linguistic Data Consortium’s (LDC) newest corpus of Chinese newswires, the *Chinese Gigaword Third Edition*, released in August 2007, for the corpus study. The corpus contains a comprehensive archive of newswire text data from the Xinhua News Agency (PRC), Central News Agency (Taiwan), Agence France Presse, and Zaobao Newspaper (Singapore). I elected to use the Xinhua (XIN) and Central News Agency (CNA) texts since they are most representative of news writing in the PRC and Taiwan, the two Chinese societies of interest in my study. Additionally, Xinhua and the Central News Agency are both state news agencies. They have similar objectives, content, styles of writing, readers, and so on, all of which are aspects that I wanted to control as much as possible in the study.

Original data archives received by the LDC from Agence France Presse, Xinhua, and Zaobao were encoded in GB-2312, while those from CNA were encoded in Big-5. To
avoid problems that could result from differences in character-set specifications, LDC converted all text files in the corpus to UTF-8 character encoding, which I retained for the study.

The data in the corpus is categorized by LDC into the following four types: 1. story: this the most frequent type and represents the most typical newswire item: a report on a topic or event consisting of paragraphs and full sentences; 2. multi: this type of document contains a series of unrelated blurbs, each of which briefly describes a particular topic or event; 3. advis: (“advisory”) this type of document is addressed by news services to news editors and is not intended for publication; and 4. other: these are documents that clearly do not fall into any of the above types, such as sports scores, stock prices, temperatures in different cities, and so on. Expecting that there might be differences in the writing styles and language choice used in the four types of texts, I elected to use only stories for my study to avoid complications in the analysis that could result from mixing different types of newswire items.

5.2.2 Data file markup

The data files in the corpus are saved as text files and are presented in Standard Generalized Markup Language (SGML) form using a simple, minimal markup structure. (SGML is a standard for markup languages.) The file name, or DOC id, for each document consists of a prefix identifying the news source and a 6-digit date containing the year and month in which the file contents were originally published by the news source, followed by a “.gz” file extension indicating that the file contents were
compressed using the GNU “gzip” compression utility. In the online “readme” file for the corpus, LDC points out that each file contains “all the usable data received by LDC for the given month from the given news source.”

The markup structure in each document contains six types of tags: the DOC id, type (story, blurb, advis, other), headline, dateline, text, and paragraphs, as shown in 5.1, adapted from the LDC’s “readme” file for the corpus.

5.1  <DOC id= "..." type= "...">
   <HEADLINE>
   The Headline is optional -- not all documents have one.
   </HEADLINE>
   <DATELINE>
   The Dateline is optional -- not all documents have one.
   </DATELINE>
   <TEXT>
   <P>
   Paragraph tags are only used if the ‘type’ attribute is story
   </P>
   <P>
   All data files use the UNIX-standard “\n” form of line termination, and text lines are generally wrapped to a width of 40 characters or less.
   </P>
   </TEXT>
   </DOC>

The document ID uniquely identifies each document in the corpus through a combination of elements: the prefix for the news source, the year, month, and date of publication, and a sequence number. The dateline is a brief string that gives the location the report is coming from and sometimes the news service and date. Examples of newswires from XIN and CNA are given in Appendices A and B.
5.2.3 Time frame and data quantity

Although the corpus contains a full sixteen years’ worth of data for XIN (1991-2006), data for the years 2003, 2004, and 2006 for CNA are incomplete. I excluded these years from parts of my analysis, as noted in the results in the next chapter, and I used 2005 as the final year for which I sampled data. Files covering the entire period from 1991 to 2006 were sampled to find trends in lettered word use over time, and the first, middle, and final year of the two parts of the corpus (1991, 1998, 2005) were sampled to identify types of lettered words in XIN and CNA. According to the online LDC “data statistics” files for the corpus, the story portion of the XIN part of the corpus for 1991 to 2005 contains 1,026,057 document files, 483,539 K-words (thousands of Chinese characters), and 1,521,962 KB of text. The comparable CNA part of the corpus contains 1,778,426 document files, 777,250 K-words, and 2,435,202 KB of text.

5.3 Research questions and hypotheses

5.3.1 Introduction

The goal of the corpus study was to examine a randomly chosen subset of lettered words in XIN and CNA newswires to determine whether the characteristics of lettered words described in Chapter 4 hold across a large sample of lettered words. In addition, I also aimed to identify similarities and differences between lettered words used in the PRC and Taiwan, to find trends in lettered word use in the two sets of newswires during the 1991-2006 period, and to examine how the newswire context affects lettered word use.
The corpus study provides insights into the different types of lettered words used in Chinese newswires, and more broadly, what the written form of lettered words reveals about the process of adapting borrowed English words and concepts into Mandarin. In particular, I wanted to determine whether the points of congruence between English and Chinese morphology and between the two writing systems are exploited in the large sample I examined in the corpus study. Equally important, answers to the research questions help to show the degree to which PRC and Taiwan newswires use lettered words that diverge from the points of congruence between the two languages. Finally, the context of newswires itself provides supplementary information about lettered words that makes them easier to understand in Mandarin, namely translations and transcriptions of the lettered items.

My research questions and hypotheses are presented in the following sections.

5.3.2 Use of roman letter strings in the PRC vs. Taiwan

Research question

I suggest that fully spelled strings require greater familiarity with English than RL-zì strings to be used widely. In Chapter 4, I noted that Taiwan has had a longer and more intense period of Mandarin-English language contact in the post World War II era than the PRC. Are more fully spelled words used in CNA than XIN?

Hypothesis

Based on Taiwan’s longer period of language contact, I expect that Taiwan has greater use of FS strings than the PRC. I test this as follows:
The proportion of FS strings in total roman letter strings is greater in CNA than XIN during 1991, 1998, 2005. (Conversely, the proportion of RL zi strings is greater in XIN than CNA during the same period.)

5.3.3 Use of roman letter strings over time

Research question

As societal familiarity with English increases, so does the use of a wider variety of English words, most of which are spelled out. Does the use of spelled words increase over the sixteen-year time period in CNA and XIN as familiarity with English increases?

Hypotheses

During the period from 1991 to 2006, English became increasingly important in both Chinese societies, which may have contributed to the increased use of lettered strings. However, Taiwan has a longer period of use and possibly greater overall use of English than mainland China. Based on this, I hypothesize that:

a. The use of roman letter strings increased from 1991 to 2006 in both XIN and CNA.

b. CNA has a greater overall use of roman letter strings from 1991 to 2006 than XIN.

c. CNA has a faster increase in the use of roman letter strings from 1991 to 2006 than XIN.
5.3.4 Characteristics of roman letter strings

Research question

Roman letter words that are zi-like appear to resolve the mismatch in the Chinese and English writing systems and appear to be integrated into Chinese more easily than fully spelled words. Do words written entirely in roman letters in the newswires tend to be zi-like rather than fully spelled?

Hypotheses

1. In Chapter 4, I proposed that from the perspective of Mandarin grammar and the Chinese writing system, RL zi are more congruent with Chinese zi than fully spelled units. Strings of RL zi may therefore be more common in Mandarin than fully spelled strings. I test this with the following hypothesis:

   For both CNA and XIN, the proportion of roman letter strings composed of RL zi is greater than the proportion of FS strings.

2. Observation of RL zi in the corpus shows that many appear to be initialisms borrowed from English. I tested whether the majority of RL zi in CNA and XIN are initialisms.

3. Shorter roman letter strings may be preferred to longer strings. Fully spelled strings appear to be longer on average than RL zi strings. To compare the length of RL zi strings and FS strings, I tested:

   a. FS strings are longer than RL zi strings.

   b. The most frequent length of RL strings is closer to the average length of RL zi strings than the average length of FS strings.
5.3.5 Case Combinations Used in Writing RL Strings

Research question

Three types of case combinations are used in writing English words: all upper case, all lower case, and mixed case. Are the same types used in Chinese newswires? I expect that the distribution of the three types may be different in Chinese newswires for at least two reasons. First, Chinese writing emphasizes different roman letter-based linguistic units from English writing (initialisms rather than spelled words), and second, the three types of letter combinations may have fewer functions in Chinese writing than in English writing and may not be used equally often.

Hypotheses

I expected to observe that all upper case (UC) are the most common case combination in Chinese writing and that all lower case (LC) and mixed case (MC) are used less frequently. I also expected that UC is used primarily to write RL zì strings, and that most of these are initialisms. I expect that LC and MC are used to write mostly FS strings. I tested these claims as follows:

a. UC is used more frequently than LC or MC in both XIN and CNA.

b. Most MC and LC strings are FS strings.

c. Most UC strings are RL zì strings.

d. Most UC strings are initialisms.

5.3.6 Characteristics of Hybrid Strings

Research questions
In Chapter 4, I described the structure of [M N] nominals, or endocentric noun compounds and phrases. Most of the examples I provided contain a roman letter modifier and a Chinese zì head, which I believe may be representative of roman letter-based [M N] nominals as a whole. In these types of structures, is the Chinese zì component in the primary, head position and the roman letter component the secondary, modifier position? In hybrid strings containing both roman letter and zì components, are the roman letter components zì-like or fully spelled?

Hypotheses

1. In hybrid strings that are [M N] nominals, I expect that modifiers expressed with roman letters will tend to be more common than heads expressed with roman letters. I test this as:

   In CNA and XIN, the proportion of [M N] nominals in which the roman letter modifier is expressed with roman letters is greater than proportion in which the head is expressed with roman letters.

2. In hybrid strings that are [M N] nominals, I expect that the modifier will tend to be formed with RL zì rather than FS units. This is tested as:

   In CNA and XIN, the proportion of [M N] nominals in which the roman letter modifier is RL zì is greater than the proportion in which it is FS.

3. Similarly to 2, I expect that the head will tend to be formed with RL zì rather than FS units. This is tested as:
In CNA and XIN, the proportion of [M N] nominals in which the roman letter head is formed with RL zì is greater than the proportion in which it is formed with FS units.

4. Following from 1 to 3, I have suggested that hybrid strings formed with RL zì rather than FS units are most congruent with Mandarin word formation patterns and with the frames of Chinese writing. I expect that hybrid strings will generally contain primarily RL zì rather than FS units. I test this with the following:

In CNA and XIN, the proportion of hybrid strings in which the roman letter component is RL zì is higher than the proportion in which it is FS.

5.3.7 Contextual support in newswires

Research question

The lettered words in the corpus occur in the context of newswires, one of whose functions is to introduce and explain specialized terms. Are translations or transcriptions provided for lettered words as a form of ‘contextual support’ for these words?

Hypotheses

The hypotheses below regarding contextual support provided in newswires are divided into those for initialisms and those for fully spelled strings. The division is based on my initial observations of the corpus, which showed that the two types of strings appear to be treated differently as to whether a translation or a transcription is given.
Initialisms

1. It appears that initialisms that are one letter in length tend not to be given translations. This may be because single-letter RL zi have a high degree of polysemy, so that they can be accommodated by reanalysis, whereas strings of upper case RL zi tend to be abbreviations. I test this claim with the following:

   In CNA and XIN, most translations for initialisms are given for those that are more than one letter in length.

2. Initialisms appear not to be given a transcription in newswires. This is tested as:

   In CNA and XIN, most initialisms do not have transcriptions.

Fully Spelled Strings

1. My initial observations showed that newswires seem to provide translations for FS strings that are not proper names. This observation is tested as:

   In CNA and XIN, most FS strings that are proper names do not have a translation.

2. Following from 1, I noticed that when transcriptions are provided in newswires, they appear to be provided for FS strings that are proper names. I test this as:

   In CNA and XIN, most transcriptions are provided for FS strings that are proper names.

5.4 Finding lettered strings

I examined the corpus before beginning the corpus study and quickly found that not all the items written in roman letters can necessarily be called words as we commonly
think of them. Examples include the roman letters used in 乔治·W·布希‘George W. Bush’, 陈 X ‘X Chen’ (an anonymized name used to refer to a suspected criminal), and 2005 年 ‘(the) year 2005’ in which the two zeroes are written using the capital letter O rather than the numeral. I also found that instead of instances of a single letter or lettered word occurring on its own in a Chinese phrase or sentence, multiple lettered words occurred together (e.g. Hungarian Trade Office). Interestingly, a fair number of these multiple-word strings were not space-separated into words. I found examples such as CONFERENCEBOARD, in which the words were not marked at all, and DifferentPoliticalViewer, in which the words were marked through the use of mixed case letters but were not separated by white space. I wanted to count all instances of multiple word groupings, space-separated or not, as one lettered token since they occur as a single, coherent unit of meaning in the sentence. Thus, I decided that it would be more appropriate to extract roman letter strings rather than roman letter words. The strings could then be analyzed to determine whether they contained words, and if so, what types of words.

I define a lettered string as any one of the following types of meaning units in a sentence or a headline. First are individual words, mostly from English, written in roman letters, such as WTO, Christie’s, M&M, etc. Second are multiple words written in roman letters and occurring together as a unit in the Chinese context, such as Lord Geoffrey Howe, Asian Forum, and thank you. Third are alphanumeric strings, including 200CC, G7, H5N1, and others. Fourth are the roman letter components of hybrid strings that contain both roman letters and Chinese characters. Some of these may be considered
compounds, such as 维生素 A ‘vitamin A’, B 组 ‘group B’, and VCD 视盘机 ‘VCD player’, and others as phrases lacking the *de* particle (discussed in Chapter 2), such as IPV6 标准 ‘IPV6 standard’ and KGB 工作人员 ‘KGB staff’, in which *de* could optionally be added between the lettered element and the Chinese *zi* element. Since it is difficult to separate the compounds from the phrases, as discussed in Chapter 2, I grouped them together as [M N] nominals. Finally, the sample data also included two Internet web addresses in CNA 1998 and five computer path names in CNA 2005.

### 5.5 Sampling procedures

Two corpus searches were conducted to answer the research questions and test the hypotheses. First, a sample was extracted that would be representative of the lettered words occurring in the XIN and CNA newswires from 1991 to 2005. Second, all of the roman letter strings in the XIN and CNA newswires from 1991 to 2006 to show changes in their use over time and to show differences in use over time in initialisms and spelled words. These procedures are described below.

#### 5.5.1 Sample for 1991, 1998, 2005

A simple random sample was taken of newswires in the XIN and CNA parts of the corpus in the first, middle, and final years of the corpus (1991, 1998, 2005), creating six sets of data in which lettered strings were analyzed. A Python script {http://python.org} was used to randomly select 25 articles containing roman letter sequences from each of
three years of data (1991, 1998, 2005) from the XIN and CNA sections of the corpus. In UTF-8, the English alphabet occupies a fixed code range, so roman letter sequences could be extracted as regular expressions. Roman letter strings were defined as the regular expression

\[(A-Za-z0-9\-\.]+)?[A-Za-z](A-Za-z0-9\-\.]+)?\]

which matches any instance of a roman letter [A-Za-z] optionally preceded by one or more letters, digits, hyphens, apostrophes or periods and optionally followed by one or more letters, digits, hyphens, apostrophes or periods.

The aim of the sampling procedure is twofold: first, to extract whole newswires of the type story containing lettered strings, and second, to obtain an approximately equal number of tokens of lettered strings in each of the six sets of newswires (XIN 1991, 1998, 2005 and CNA 1991, 1998, 2005). In articles containing one or more lettered strings, the headline and all of the sentences in the article were extracted. The roman letter strings in the headline (if any) and in the sentences were highlighted to aid readability and to ensure that no lettered strings would be missed in coding. Each instance of a lettered string in an article, be it a word (A 股 ‘A (stock) share’), phrase (thank you), alphanumeric string (200CC), or other kind of string (2005 年 ‘(the) year 2005’ was coded as one token.

The original 25 articles sampled from each of the six sets of newswires produced differing numbers of tokens. More articles were added with the randomized sampling procedure until approximately equal numbers of tokens were obtained for each set of newswires. A summary of the sample data is shown in Table 5.

---

1 Both corpus searches were conducted in collaboration with Kirk Baker using Python

107
Table 5 Summary of the sample data for the corpus study

<table>
<thead>
<tr>
<th>News agency and year</th>
<th>No. of newswires</th>
<th>No. of tokens</th>
<th>News agency and year</th>
<th>No. of newswires</th>
<th>No. of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>XIN 1991</td>
<td>42</td>
<td>115</td>
<td>CNA 1991</td>
<td>55</td>
<td>113</td>
</tr>
<tr>
<td>XIN 1998</td>
<td>41</td>
<td>116</td>
<td>CNA 1998</td>
<td>32</td>
<td>117</td>
</tr>
<tr>
<td>XIN 2005</td>
<td>28</td>
<td>114</td>
<td>CNA 2005</td>
<td>25</td>
<td>112</td>
</tr>
<tr>
<td>Total XIN</td>
<td>111</td>
<td>345</td>
<td>Total CNA</td>
<td>112</td>
<td>342</td>
</tr>
</tbody>
</table>

Grand Total

| Newswires | 223 |
| Tokens    | 687 |

The samples for XIN 2005 and CNA 2005 contain fewer articles than the other sets of newswires, and CNA 1991 contains the greatest number. This discrepancy occurred because each newswire had a different number of lettered strings in it. Some had only one, while others had many. Since each instance of a lettered string was counted as one token, several articles were needed in some cases to obtain the same number of tokens as in one article containing multiple lettered strings.

5.5.2 RL strings in XIN and CNA from 1991 to 2006

In addition to the sample of newswires for 1991, 1998, and 2005 in XIN and CNA, I also wanted to see changes in the use of roman letter strings during the entire sixteen-year period of the corpus and to see differences over time in the use of initialisms and spelled words in XIN and CNA. To do this, a Python script was used to extract all the roman scripts that he created after we discussed my research questions and hypotheses.
letter strings from XIN and CNA from 1991 to 2006. Figures 6.1, 6.2, and 6.3 in the next chapter were created with the data collected for these years.

Figure 6.1, which shows the number of roman letter strings per ten thousand Chinese characters, was created with data obtained by counting all occurrences of roman letter strings in each year from 1991 to 2006 in XIN and CNA using the regular expression given in the previous section. Figure 6.2, which compares the counts of probable intialisms and spelled words in XIN and CNA over time, was created with totals per year in XIN and CNA obtained by counting all occurrences of ‘all upper case’ strings and all occurrences of ‘mixed case’ strings in each year. ‘All upper case strings’ were counted using the regular expression

\[(\text{[A-Z0-9-\'.]+})?\text{[A-Z]}(\text{[A-Z0-9-\'.]+})?\]

which matches any instance of a capital letter roman letter optionally preceded by one or more capital letters, digits, hyphens, apostrophes, or periods and optionally followed by the same.

Preliminary analysis of the corpus showed that ‘all upper case’ strings are normally initialisms rather than spelled words. All upper case strings were therefore used to represent initialisms. There were some instances of spelled words written in all-caps, however, such as NIKE and MILD 7. Thus, in Chapter 6 when I refer to the results of the corpus search obtained from the regular expression for all upper case strings (above), I call the resulting strings ‘probable initialisms’ rather than ‘initialisms’.
Preliminary analysis of the corpus also showed that ‘mixed case strings’ are nearly always spelled words rather than initialisms. Thus, mixed case strings were used to represent spelled words. ‘Mixed case’ strings were counted using the regular expression

```
([A-Za-z0-9.]+)[A-Z](?[A-Za-z0-9]+)?
```

which matches any instance of a capital roman letter optionally preceded by one or more capital letters or lower case letters, digits, hyphens, apostrophes, or periods and optionally followed by the same.

Figure 6.3, which shows the proportion of probable initialisms in XIN and CNA over time, uses the same data as Figure 6.2 but converts the counts to the proportion: (number of all upper case) / (number of all upper case + number of mixed case). This represents the proportion of probable initialisms in total roman letter strings as approximated by the count of upper case and mixed case strings. As mentioned above, upper case strings were used to represent initialisms, since strings written entirely in upper case are most commonly initialisms. Mixed case strings were used to represent spelled words, since mixed case strings are normally spelled words rather than initialisms.

### 5.5.3 Using whole articles

The Python script was designed to extract whole articles containing lettered strings rather than isolated sentences in the corpus containing lettered strings or parts of sentences containing lettered strings. My motivation for this is that I want to retain the larger communicative context in which the lettered strings occur, that is, the newswire. I expect that the text prior to and following each lettered string can provide useful
information about the lettered string itself, how it is used in the context of the newswire that contained it, and how other lettered strings in that newswire might relate to it. For example, translations or transcriptions of some lettered strings occur in sentences other than the one containing the lettered string itself, and some lettered strings have two forms, a full form used when the lettered string is first introduced and an abbreviated form used after that (e.g. F-15 战斗机 ‘F-15 fighter jet’ vs. F-15). Use of the abbreviated form is contingent on the earlier use of the full form. Thus, the newswire as a complete discourse was necessary for interpreting lettered string tokens effectively and for testing the hypotheses regarding contextual support described in section 5.3.7.

5.6 Variables that were coded

Eight variables were coded for each lettered string in the sample described in 5.5.1. These eight variables enabled me to address the research questions and the hypotheses. They were:

1. Number of roman letters in the string: The total number of roman letters in the string was recorded: X 光 ‘X ray’ contains one roman letter, 200CC contains two, and *foreign agent* contains twelve. Numerals occurring with roman letters were not counted (e.g. in *H5N1*), since they are common symbols used in Chinese writing and occur in combinations with Chinese characters as well, rather than simply in combination with roman letters. I therefore treated them as being equivalent to Chinese *zi*. I wanted to count just the number of roman letters that are used in a string.
2. Roman letter string type: This variable describes whether the lettered string is \( zi \)-like. The variable has two levels, \textit{Has roman letter \( zi \) (RL \( zi \))} and \textit{has fully spelled (FS)}. \textit{Has RL \( zi \)} refers to strings that contain roman letters that could be considered roman letter \( zi \) in Chinese. RL \( zi \) are defined as letters with the following characteristics: 1) they can be pronounced independently of other letters around them with their letter name, 2) they act as ‘single letter units’ that have independent meanings as morphemes or symbols, as discussed in Chapter 4; and 3) they are ‘visually discrete’ (Hansell 1989), taking up one equidimensional square and not needing to be joined with other letters around them to be pronounced. These three criteria were not difficult to apply to roman letter strings. The only potentially difficult cases were words such as \textit{SARS}, which English speakers treat as acronyms but which Chinese speakers may or may not know to pronounce as acronyms.

I consider acronyms written in upper case letters to be RL \( zi \) strings because it is not incorrect to pronounce them letter by letter in the manner of initialisms. Whether they are pronounced as acronyms or initialisms depends on convention rather than rule. In contrast, I did not consider examples such as \textit{NIKE}, which was written in a newswire in all capital letters and which looks like an initialism, to be an RL \( zi \) string because its correct pronunciation in English is as a spelled word. Since Chinese speakers generally aim to approximate the English pronunciation of lettered words, they do not have the option of pronouncing \textit{NIKE} letter by letter as they do with \textit{SARS}, so I do not consider it an RL \( zi \) string.

The second level, \textit{has FS} refers to strings that contain roman letters that are part of a spelled out word, morpheme, or string of words. If a string were to contain both RL \( zi \)
and FS units, such as *ABC Company* in which *ABC* would be classified as RL zi and *Company* as FS, it would be coded as both *has RL zi* and *has FS*. No such instances occurred in the sample, however. All the strings were either all RL zi or all FS.

3. Initialism: This variable codes whether the roman letter string is an initialism. Based on the limited discussion of initialisms in morphology textbooks such as Bauer (1983), Katamba (1993), Aronoff and Fudeman (2005), Harley (2006), and others, initialisms are not considered to be an important category of words in English. By contrast, my initial examination of the corpus showed that a large number of lettered strings in the newswires do appear to be initialisms, however. I believe this is because initialisms are a clear case of the zi-like use of roman letters with respect to both orthography and morphology. Thus, I felt it would be useful to code whether a lettered string is an initialism.

4. Case: *Case* describes the case of the roman letter(s) used in the string. The variable has three levels. *Has upper case (UC)* refers to strings that contain one or more UC letters, while *has lower case (LC)* refers to those containing one or more LC letters. Strings containing a mixed combination of letters, that is, both UC and LC (e.g. personal names), were labeled as both *has UC* and as *has LC*.

5. Orthographic string type: This variable describes whether the lettered item is a hybrid string containing both roman letters and Chinese characters. Strings containing one or more roman letters and one or more Chinese characters were labeled as hybrid strings (e.g. *Mars 公司* ‘Mars Company’), while those that contained just roman letters were labeled as non-hybrid strings (e.g. *ESPN*).
6. Hybrid string type: Initial examination of the corpus showed that many hybrid strings were [M N] nominals (that is, compounds or phrases without the de particle) containing a modifier and a head noun. Of interest in these cases is the distribution of roman letter modifiers as compared with roman letter heads. The hybrid string type variable has three levels to capture this information, has roman letter modifier, has roman letter head, and has neither. The levels allowed me to code three types of strings: 1. those with a roman letter modifier but no roman letter head (i.e. those that have a Chinese zì head, such as A 组‘group A’), 2. those with a roman letter head but no roman letter modifier (i.e. those that have a Chinese zì modifier, such as 相声 TV ‘cross-talk TV’, a type of television program), and 3. those with neither a roman letter head nor a roman letter modifier, in other words, those that are hybrid strings but not [M N] nominals, such as B 小姐 ‘Miss B’.

One complication with the [M N] nominals was that certain examples had more than just two compounded components. In all these cases, the examples were right branching, with a Chinese zì head on the right. The examples were similar to the difference in English between book club and book club meeting notes, which is composed of three compounded words and has notes as the head. The structure of the compound is [[[book club]meeting]notes].

My interest in considering the roman letter modifier vs. roman letter head question in [M N] nominals is that I was interested in learning how roman letter-based components begin to participate in the Chinese compounding process. I wanted to know the first level of compounding that they participate in because speakers depend on an understanding of
roman letter elements at that level to engage in subsequent levels of compounding. For example, in IBM 公司产品 ‘IBM company products’, speakers need to understand IBM 公司 before they can understand IBM 公司产品. I therefore analyze IBM 公司产品 simply as IBM 公司, which has a roman letter modifier and a Chinese zi head.

Another reason I analyze roman letter compounds in this manner is that in my sample, even after two or more levels of compounding, the items still had the same structure as at the first level. The sample included examples such as IBM 公司产品 ‘IBM company products’ but none of the type IBM 公司 products ‘IBM company products’ ([[IBM 公司 ] products]], in which IBM modifies 公司 and in which IBM 公司 modifies the roman letter element products. Nor were there examples such as 相声 TV 节目 ‘cross-talk television program’ ([[相声 TV 节目]], in which 相声 modifies the head TV and then 相声 TV modifies the Chinese head 节目. Since the sample included no examples in which subsequent levels of compounding changed the head from a Chinese zi element to a roman letter-based one or vice versa, I considered only the first level of compounding in which the roman letter-based element was used.

7. Contextual support: In examining the corpus, I noticed that newswires provided Chinese zi-based ‘contextual support’ for lettered strings in the form of translations or transcriptions that help readers understand the lettered strings. Translations are either exact translations of the foreign term or near equivalents (e.g. dissident > 异议人士 yiyì rénshì lit. ‘different-opinion personage’). Transcriptions use Chinese zi to convey the pronunciation of the foreign term but not its meaning (e.g. Monet > 莫内 mònèi lit. ‘not-
inside’). I realized from the translations and transcriptions in the newswires that lettered strings are not necessarily presented simply as strings written in a foreign writing system that readers have to decipher with their own linguistic resources. Rather, the newswires provide Chinese zi-based assistance either by indicating the meaning of the lettered string or its pronunciation.

The variable contextual support has three levels: has translation, has transcription, and has neither. If a translation for a lettered string token occurred anywhere in the newswire, the token was labeled as having a translation. Similarly, if a transcription for a lettered string token occurred anywhere in the newswire, the token was labeled as having a transcription. If a particular lettered string occurred multiple times in a newswire, each token was labeled as having a translation or transcription when one was included in the newswire. The motivation for this form of labeling is that a translation or transcription is likely to occur only once in a newswire, and after it has been introduced, the reader is aware of it from there on.

8. Proper name: I found in examining the corpus that transcriptions appeared to occur primarily for proper names and that translations appeared to occur primarily for other types of items. For this reason, I coded tokens with an associated translation or transcription as being either a proper name or not proper name. I defined the notion of proper name narrowly as a personal name, place name, or organization name to make the tokens easier to code.
5.7 Sample coded data

Sample newswires from XIN and CNA are given in Appendix A. Coded tokens for the XIN newswire are in Appendix B and coded tokens for the CNA newswire are in Appendix C. A list of types of roman letter strings found in the samples from XIN and CNA in 1991, 1998, and 2005 are given in Appendix D.

Two sets of Excel spreadsheets were used to code the data in each set of newswires. The first spreadsheet contains all of the sentences in each newswire in the sample placed on separate lines. If a sentence contained a roman letter string, the string was given next to the sentence. Some sentences had multiple roman letter strings and are repeated on separate lines, once for each roman letter string.

This file is essentially a first pass at identifying lettered strings in the newswires. I used it to determine which sentences contained roman letter strings and then to look for the strings that were indicated. The sentence and newswire context was useful for determining what the words meant and for finding translations and transcriptions. It was not possible to automate the process of searching for roman letter strings beyond the simple measure of extracting the roman letter strings themselves because many roman letter strings were actually hybrid strings. (IBM 公司 ‘IBM company’ is an example.) I needed to check each roman letter string individually to determine whether any characters surrounding it formed a compound or a phrase with it.

The second Excel spreadsheet contains the coded data for the sample newswires. I went through the first spreadsheet row by row to extract the roman letter string and its surrounding characters, if relevant, from the first spreadsheet and then coded it for the
variables described in section 5.6 in the remaining cells of the corresponding row in the second spreadsheet. The appendices show the format of the Excel spreadsheet used to code the data. The columns provide the following information: 1. item number, 2. newswire DOC id, 3. article number (in my sample), 4. sentence number (0 = headline), 5. roman letter string, 6. number of roman letters, 7. RLzi (values: 0,1), 8. fully spelled (0,1), 9. initialism (0,1), 10. upper case (0,1), 11. lower case (0,1), 10. hybrid string (0,1), 11. roman letter modifier (0,1), 12. roman letter head (0,1), 13. translation (0,1), 14. transcription (0,1), and 15. proper name (0,1).

5.8 Data analysis

Analysis of the 687 tokens of lettered strings in XIN and CNA 1991, 1998, 2005 was conducted with R 2.7.0 and with Python. The results are presented in Chapter 6.
6.1 Introduction

This chapter presents the results of the corpus study conducted with data from the Chinese Gigaword Third Edition corpus. The entire corpus of newswire data from Xinhua (XIN) (PRC) and Central News Agency (CNA) (Taiwan) during the period 1991 to 2006 was used to test the hypotheses in sections 6.2, 6.6, and 6.7 below. The remaining hypothesis tests were conducted on XIN and CNA newswire data from 1991, 1998, and 2005. A random sample of 223 newswires yielding 687 tokens formed the data set that was analyzed; 345 tokens are from XIN and 342 from CNA. The statistical tests were run on R 2.7.0.

6.2 Increase in use of lettered strings from 1991 to 2006

In Chapter 5, I hypothesized that CNA may have a greater overall use of roman letter strings from 1991 to 2006 than XIN and that CNA may have a faster rate of growth in the use of roman letter strings than XIN during the same period. The rate of growth in the use of roman letter strings in XIN and CNA from 1991 to 2006 is shown in Figure 6.1. The
figure shows the number of roman letter strings in the newswires per 10,000 Chinese
characters during that period.

![Graph showing the number of Roman letter strings per 10K Chinese Characters]

**Figure 6.1 Rate of Increase in the Use of Roman Letter Strings in CNA and XIN**

The CNA data point for 2003 in Figure 6.1 is higher than those around it. The data
point is based on just a sample of newswires from that year rather than on the whole
population of newswires, since many newswires were missing in the LDC corpus for all
of the months of that year. The corpus data for 2003 contains only about 20% of the
number of newswires of the preceding years. This sample size was big enough to use,
however, yielding a sample proportion that could be compared with the population proportions in the other years for which the corpus data is complete. The CNA corpus data for 2004 and 2006 are also incomplete and also provided sample proportions rather than population proportions, unlike the other years. In 2004 the corpus contains only about half of the expected newswire files from January to August. In 2006 no newswires are provided in the corpus from August to November. The CNA data points for 2004 and 2006 are used as well because there is no significant evidence that they should be excluded. The sample proportions for 2003, 2004, and 2006 are close to the population proportions, and the larger sample size obtained by including them improves the fit of the regression line.

Figure 6.1 shows that during any given year from 1991 to 2006, the use of roman letter strings in CNA is greater than in XIN. The two sets of newswires have a similar use of lettered strings in 1991 and 1992 but diverge increasingly in subsequent years. The use of lettered strings increases in both CNA and XIN from 1991 to 2006, but the rate of increase is considerably faster for CNA than XIN during that time period. The data for each year show that the use of lettered strings in XIN is almost unchanged from 1991 to 1995, then increases slowly from 1996 to 1998, and finally speeds up considerably from 1999 to 2006.

A linear regression analysis of CNA revealed that the year was a highly significant predictor of the number of roman letter strings ($\beta = 1.99, p < .001$) and accounted for most of the variance in number of roman letter strings, $R^2 = .90, F(1, 14) = 136.2, p < .001$. Adding a quadratic component ($\text{year}^2$) to the CNA regression model did not have a
significant effect, so that term was not included. Rather, a linear regression model provided the best fit to the data. Adding a quadratic term was significant for XIN, however. A quadratic regression model provided the best fit to the XIN data. For XIN, number of roman letter strings was regressed on the year and the quadratic component, year^2. Both the year (β = -81.6, p = .00384) and the year^2 (β = 0.02, p = .00371) demonstrated significant effects on the number of roman letter strings. The two predictors also explained a highly significant proportion of variance in the number of roman letter strings, R^2 = .96, F(2, 13) = 163.4, p < .001. The quadratic regression coefficient for XIN captures the upward curve beginning in 1999 that indicates a faster rate of growth in roman letter strings from 1999 to 2006 in XIN than from 1991 to 1998.

In summary, the regression analysis for CNA and XIN shows that the rate of growth for roman letter strings over the sixteen-year period is best described as constant for CNA; in contrast, for XIN it changes slowly from 1991 to 1998 and then increases more rapidly in subsequent years until 2006.

6.3 Prevalence of roman letter zi strings

I proposed that strings of RL zi may be more common than fully spelled strings (FS) in the newswire corpus. To test this, I determined whether the proportion of roman letter strings composed of RL zi is greater than the proportion of roman letter strings composed of FS strings for both CNA and XIN. The data for RL zi strings and FS strings in XIN and CNA 1991, 1998, and 2005 are shown in Table 6.1.
The sample contains 687 tokens of lettered strings, 560 of which are RL \( zî \) strings and 127 of which are FS strings. The proportion of RL \( zî \) strings in XIN (0.97) is considerably greater than the proportion of FS strings (0.03), leading me to conclude that RL \( zî \) strings form the majority of lettered strings in XIN in the sample. Although the proportion of RL \( zî \) strings in CNA is about twice that of FS strings (0.66 vs. 0.34), it is about a third less than in XIN. Since the proportions of RL \( zî \) strings and FS strings in CNA are less obviously different than in XIN, a one-sample proportions test was used to test whether the proportion of FS strings in CNA is in fact in the minority. The test provides significant evidence that the proportion of FS strings in CNA is less than 50%, that is, in the minority \( (\chi^2 = 34.7, p < .001) \). The proportion is at most 0.38. These results indicate that the proportion of FS strings in CNA is greater than RL \( zî \) strings and lead me to conclude that RL \( zî \) strings form the majority of roman letter strings in the CNA newswires in 1991, 1998, and 2005.

A one-sample proportions test was also used to confirm whether the proportion of FS strings in both sections of the corpus, that is, in XIN and CNA together, is in the minority. The results provide significant evidence that the proportion of FS strings is less

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>RL ( zî ) strings</th>
<th>FS strings</th>
<th>Proportion of RL ( zî ) strings</th>
<th>Proportion of FS strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>XIN</td>
<td>345</td>
<td>334</td>
<td>11</td>
<td>0.97</td>
<td>0.03</td>
</tr>
<tr>
<td>CNA</td>
<td>342</td>
<td>226</td>
<td>116</td>
<td>0.66</td>
<td>0.34</td>
</tr>
<tr>
<td>Total</td>
<td>687</td>
<td>560</td>
<td>127</td>
<td>0.82</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 6.1 Proportion of RL \( zî \) Strings and FS Strings in XIN and CNA
than half ($\chi^2 = 271.7, p < .001$). It is at most 0.21, indicating that RL zì form the majority of lettered strings.

### 6.4 Comparison of fully spelled strings in XIN and CNA

In addition to finding the total proportion of RL zì strings and FS strings in XIN and CNA as shown in Table 6.1, I also found the proportion of FS strings in XIN and CNA for the three years to determine whether FS strings are more common in CNA than XIN in those years. The data are shown in Table 6.2.

<table>
<thead>
<tr>
<th></th>
<th>CNA Total Strings</th>
<th>CNA FS Strings</th>
<th>Prop. FS in CNA</th>
<th>XIN Total Strings</th>
<th>XIN FS Strings</th>
<th>Prop. FS in XIN</th>
</tr>
</thead>
<tbody>
<tr>
<td>1991</td>
<td>113</td>
<td>18</td>
<td>0.16</td>
<td>115</td>
<td>5</td>
<td>0.04</td>
</tr>
<tr>
<td>1998</td>
<td>117</td>
<td>33</td>
<td>0.28</td>
<td>116</td>
<td>5</td>
<td>0.04</td>
</tr>
<tr>
<td>2005</td>
<td>112</td>
<td>65</td>
<td>0.58</td>
<td>114</td>
<td>1</td>
<td>0.01</td>
</tr>
<tr>
<td>Total</td>
<td>342</td>
<td>116</td>
<td>0.34</td>
<td>345</td>
<td>11</td>
<td>0.03</td>
</tr>
</tbody>
</table>

Table 6.2 Fully Spelled Strings in XIN and CNA in 1991, 1998, 2005

Table 6.2 shows that the proportion of FS strings increases steadily in CNA from 1991 to 2005, so that by 2005, over half of the lettered strings in the sample are FS strings (58%). The proportion of FS strings in XIN remains low during the three years at less than 0.05. A two-sample proportions test was used to find the true difference in the proportion of FS strings in CNA and XIN in 1991 and in 1998. The test for 1991 provides significant

---

1 In this statistical test and in subsequent ones, the p-value represents the probability of a $\chi^2$-distributed random variable.
evidence that the difference between the proportions is at least 4% ($\chi^2 = 7.2, p = .004$); for 1998 the difference it is at least 16% ($\chi^2 = 22.6, p < .001$), leading me to conclude that the proportion of FS strings in CNA is greater than in XIN in both years. As for 2005, the proportion of FS strings in CNA and XIN in 2005 are obviously different (0.58 vs. 0.01), indicating that CNA had a higher proportion of FS strings in that year than XIN. In sum, CNA has a higher proportion of FS strings during each of the three years than XIN, and the gap increases over the three-year period. The proportion of FS strings in XIN remains low during each of the three years and does not appear to increase over the three years based on the sample data. The number is too small do a $\chi^2$ test in each year.

6.5 Most RL zì are initialisms

Although RL zì can have the variety of functions described in Chapter 4, I expected that initialisms would be the most common type of RL zì. The sample data from CNA and XIN show that of the 560 RL zì in the sample, 401, or 72%, are initialisms. A one-sample proportions test determines whether RL zì that could not be initialisms are in fact in the minority. The test provides significant evidence that the proportion of initialisms is greater than 50% ($\chi^2 = 103.7, p < .001$). It is at least 68%, which allows me to conclude that the majority of RL zì are in fact initialisms.

6.6 Initialisms vs. spelled words in CNA and XIN

6.6.1 Comparing initialisms and spelled words
In the previous sections I showed that the proportion of FS strings in total roman letter strings is greater in CNA than XIN in 1991, 1998, and 2005. Individual spelled words are the most common type of FS string, and probable initialisms are the most common type of RL zi string. Figure 6.2 compares the number of probable initialisms to the number of clear spelled words in CNA and XIN. The counts for CNA are shown for 1991 to 2005 and those for XIN are shown for 1991 to 2006.

Figure 6.2 Probable Initialisms vs. Spelled Words in CNA and XIN

Figure 6.2 excludes the CNA data for 2003 and 2006 since the number of newswires in the corpus for those years was much less than expected, leading to considerably lower
counts of both probable initialisms and FS words that would have had a significant effect on the shape of the regression curve. The CNA data for 2004 are also incomplete, creating noticeably lower data points for 2004 in the figure, but they are included to provide a better fit for the regression curves to 2005.

In both sets of newswires, probable initialisms outnumber spelled words in each year, which indicates that probable initialisms are more frequent than spelled words in both sets of newswires. In CNA the counts increase exponentially during the fifteen-year period but increase more slowly from 2002 to 2005. In contrast, FS words in CNA have low total counts and a slow rate of increase until 1997 (which appears flat due to the wide range of the scale), but after that year, the counts increase exponentially until they taper off from 2002 to 2005. In sum, the counts of both probable initialisms and FS words in CNA increase considerably after 1997.

In XIN the counts for probable initialisms are initially low but increase steadily from 1991 to 2006, with no sign of tapering off. In contrast, the counts for FS words are very low and barely increase from 1991 to 1998. The total counts increase and the rate of increase speeds up beginning in 1999 and continues until 2006.

6.6.2 Initialisms and spelled words in CNA

A quadratic regression model was fit to the CNA data to model the relationship between the year and the counts, with the counts as the dependent variable and the year and year^2 as the independent variables. The quadratic regression analysis for probable initialisms in CNA shows that the two predictors, year and year^2, account for most of the
variance in the counts \( R^2 = .94 \), which was highly significant, \( F(2, 11) = 108.4, p < .001 \). Both the year (\( \beta = 45.5, p = .0062 \)) and the year\(^2\) (\( \beta = -0.013, p = 0.0064 \)) demonstrated significant effects on the count. The negative year\(^2\) regression coefficient indicates that the counts increased more slowly than a purely exponential relationship. This can be seen in Figure 6.2 in the slowed increase in counts from 2002 to 2005.

A quadratic regression model was also fit to the data for FS words in CNA. The quadratic term, year\(^2\), was significant. The regression analysis indicates that year and year\(^2\) account for most of the variance in the counts \( R^2 = .96 \), which was highly significant, \( F(2, 11) = 174.6, p < .001 \). Both the year (\( \beta = 160.0, p < .001 \)) and the year\(^2\) (\( \beta = -0.04, p < .001 \)) significantly predicted the count. Similarly to the regression model for probable initialisms in CNA, the negative year\(^2\) regression coefficient indicates that the counts did not increase as quickly as a purely exponential relationship, with a slower rate of increase from 2002 to 2005.

6.6.3 Initialisms and spelled words in XIN

The counts for XIN probable initialisms and FS words in Figure 6.2 pattern differently from those in CNA. A linear regression model was fit to the data for probable initialisms. (Adding a quadratic term was not significant.) The regression analysis reveals that the year is a highly significant predictor of counts (\( \beta = 1886.1, p < .001 \)) and accounts for 88% of the variance in counts, \( R^2 = .88 \), which was significant, \( F(1, 14) = 111.6, p < .001 \). A quadratic regression model was fit to the data for FS words. The two predictors, year (\( \beta = -1.042 \times 10^5, p < .001 \)) and year\(^2\) (\( \beta = 26.13, p < .001 \)), significantly predicted
counts and also explained most of the variance in counts, $R^2 = .95$, $F(2, 13) = 157.1$, $p < .001$. The positive quadratic term in the regression model captures the faster rate of growth of FS words from 1999 to 2006.

### 6.7 Proportion of initialisms in CNA and XIN

In the previous section, I showed that probable initialisms are more numerous than spelled words in both CNA and XIN from 1991 to 2005. Figure 6.3 compares probable initialisms as a proportion of total roman letter strings in CNA and in XIN. CNA data is given for 1991 to 2005 and XIN data for 1991 to 2006.
Figure 6.3 Proportion of Probable Initialisms in CNA and XIN 1991–2005

Figure 6.3 excludes CNA data for the years 2003 and 2006. Since this figure is based on the same counts as Figure 6.2, the proportions for those years in the figure would have been significantly affected by the incomplete corpus data for those years in CNA and were therefore excluded.

Figure 6.3 shows that both sets of newswires had a high proportion of probable initialisms through the mid 1990s and a decline in the proportion of probable initialisms
after that time. The proportion of probable initialisms decreased quickly from 1997 to
2000 in CNA but stayed level from 2001 to 2005. In contrast, the proportion of probable
initialisms remained roughly the same in XIN from 1991 to 1998 and then dropped
somewhat more quickly from 1999 until 2006. Overall, the proportion of probable
initialisms in XIN was near 90% or above during the sixteen-year period, while in CNA it
dropped as low as 66%.

A linear regression analysis for CNA indicates that the year significantly predicted the
proportion of probable initialisms ($\beta = -0.03, p < .001$) and accounted for 88% of the
variance in the proportion of probable initialisms, $R^2 = .876, F(1, 12) = 92.84, p < .001$. A
quadratic term was not significant since the proportion of probable initialisms in CNA
does not appear to continue to fall from 2001 to 2006.

A quadratic regression model was fit to the XIN data. Proportion of probable
initialisms was regressed on year and year$^2$. Both year ($\beta = 1.812, p = .0180$) and year$^2$ ($\beta
= -4.55 \times 10^{-5}, p = .0177$) demonstrated significant effects on proportion of probable
initialisms. The two predictors also explained a significant proportion of variance in
proportion of probable initialisms, $R^2 = .8686, F(2, 13) = 50.58, p < .001$. The negative
regression coefficient year$^2$ models the faster rate of decrease for the proportion of
probable initialisms after 1998. Although the rate of decrease in the proportion of
probable initialisms did not fall quickly, the proportion of probable initialisms decreased
faster in later years than from 1991 to 1998.

A drop in the proportion of probable initialisms in Figure 6.3 implies a
correspondingly greater use of spelled words. The figure shows that the proportion of
probable initialisms in CNA decreased at a faster rate than in XIN from 1991 to 2006, which also means that the use of spelled words increased more overall and at a faster rate in CNA than in XIN during that period.

6.8 Length of roman letter strings

6.8.1 FS strings compared with RL ȥ%i strings

Examination of the corpus showed that fully spelled strings appeared to be longer on average than RL ȥ%i strings. I compared the average length of FS strings and RL ȥ%i strings in the sample to determine which type of string actually tends to be longer. The boxplot in Figure 6.4 shows the median lengths of FS strings and RL ȥ%i strings, as well as the range (dots at the top), the interquartile range (vertical dimension of the box), and the lower and upper adjacent values (the highest and lowest values not including the outliers). The data from the boxplot is given in numerical form in Table 6.3.
Figure 6.4 Length of Fully Spelled and RL zì Strings

Table 6.3 Summary Data for FS and RL zì Strings

<table>
<thead>
<tr>
<th></th>
<th>FS Strings</th>
<th>RL zì Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lower adjacent value</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Minimum</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>1st Quartile (25%)</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Median (50%)</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>3rd Quartile (75%)</td>
<td>13</td>
<td>3</td>
</tr>
<tr>
<td>Upper adjacent value</td>
<td>22</td>
<td>6</td>
</tr>
<tr>
<td>Maximum</td>
<td>61</td>
<td>7</td>
</tr>
</tbody>
</table>
Figure 6.4 shows that the median length of FS strings is 9 roman letters and that the median length of RL zi strings is 2 roman letters. The mean for FS strings is 10.7 roman letters ($SD = 8.40$) and 2.2 roman letters for RL zi strings ($SD = 1.15$). The difference between the means is at least 7 roman letters, $t(127) = 11.41, p < .001$. The t-test is not robust, however, since the plot for FS strings does not have a Normal distribution. A Wilcoxon rank sum test, which is used when there are outliers in the data, provides significant evidence that a difference exists between the medians ($W = 68821.5, p < 0.001$), leading me to conclude that FS strings are in fact longer than RL zi strings.

The boxplot also indicates that the length of FS strings is more variable than for RL zi strings. The height of the box (representing the middle 50% of the data) is greater for FS strings than for RL zi strings (7 letters vs. 2 letters), and the adjacent values are more widely spread apart for FS strings than for RL zi strings (2 and 22 letters for FS strings vs. 1 and 6 letters for RL zi strings).

Figure 6.4 also shows five outliers for FS strings and three for RL zi strings (all the same length). The outliers for FS strings are 24, 33, 38, 56, and 61 letters in length, and for RL zi strings all are 7 letters in length. The outliers for FS strings contain multiple words, which account for their longer length. The outliers for FS strings are DifferentPoliticalViewer (24 letters), CARIBBEAN COMMUNITY ANDCOMMON MARKET (sic.) (33 letters), C:\Program Files\Tommy\Smart Card\SmartSet.exe (33 letters), and two other long computer path names (56 and 61 letters respectively). The outlier for RL zi is the term TD-SCDMA (7 letters), repeated three times.
6.8.2 Preference for “short” RL strings

My second expectation with respect to the length of roman letter strings was that the most frequent length of roman letter strings would tend to be “short”, that is, closer to the average length of RL $zì$ strings than the average length of FS strings. Figures 6.5 and 6.6 are histograms of the lengths of roman letter strings in the XIN and CNA samples.

![Figure 6.5 Length of Roman Letter Strings in XIN](image)
Figure 6.6 Length of Roman Letter Strings in CNA

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>XIN</td>
<td>314</td>
<td>23</td>
<td>7</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CNA</td>
<td>201</td>
<td>54</td>
<td>31</td>
<td>23</td>
<td>13</td>
<td>9</td>
<td>6</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.4 Length of RL Strings and Counts in XIN and CNA

The histograms for XIN and CNA show that the most frequent roman letter strings are 1 to 3 roman letters in length, with a considerable drop-off after 3 letters. The length of RL strings in CNA is more variable than in XIN, however, as shown by the tokens of strings from 4 to 24 letters long. Although XIN has examples of strings from 4 to 9 letters long, there are considerably fewer of them than in CNA, and there are no examples of strings
longer than 9 letters except for the 24-letter outlier mentioned above. The histograms for both XIN and CNA confirm that the most frequent length of roman letter strings in the sample is indeed “short,” that is, closer to the average length of RLzi strings (2 roman letters) than fully spelled strings (11 roman letters).

6.9 Preferred case of roman letters

6.9.1 Upper case predominates

I observed that the use of upper case letters (UC) appeared to be most common in lettered strings, while mixed case (MC) and lower case (LC) appeared to be used less frequently. (Examples of these case combinations are PDF, Pdf, and pdf.) Table 6.5 shows the data for UC, MC, and LC in XIN and CNA 1991, 1998, and 2005. MC is determined by counting the number of tokens that contain both UC and LC.

<table>
<thead>
<tr>
<th>Case Combination</th>
<th># Tokens</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>only UC</td>
<td>580</td>
<td>0.84</td>
</tr>
<tr>
<td>MC (both UC &amp; LC)</td>
<td>76</td>
<td>0.11</td>
</tr>
<tr>
<td>only LC</td>
<td>31</td>
<td>0.05</td>
</tr>
<tr>
<td>Total</td>
<td>687</td>
<td>1.00</td>
</tr>
</tbody>
</table>

Table 6.5 Proportion of UC, MC, and LC in XIN and CNA

The proportions of UC, MC, and LC are noticeably different, with UC clearly being in the majority. MC is used in 11% of tokens, while LC is used in 5% of tokens. I conclude that all UC is the most commonly used case combination. The greater use of MC than LC can be attributed to proper names being the most frequent spelled word type.
6.9.2 Upper case writes RL zì strings

I have suggested that UC appears to be used to write primarily RL zì strings and that MC and LC are used to write primarily FS strings. The case combinations used for writing the two types of strings are shown in Table 6.6.

<table>
<thead>
<tr>
<th>Case</th>
<th>Total Strings</th>
<th># RL Zì Strings</th>
<th>Prop.</th>
<th># FS Strings</th>
<th>Prop.</th>
</tr>
</thead>
<tbody>
<tr>
<td>only UC</td>
<td>580</td>
<td>539</td>
<td>0.93</td>
<td>41</td>
<td>0.07</td>
</tr>
<tr>
<td>MC</td>
<td>76</td>
<td>0</td>
<td>0</td>
<td>76</td>
<td>1</td>
</tr>
<tr>
<td>only LC</td>
<td>31</td>
<td>21</td>
<td>0.68</td>
<td>10</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 6.6 Case Combinations and String Types in XIN and CNA

In the sample data UC is used to write mostly RL zì strings (93%) as expected. A one-sample proportions test confirms whether UC is actually used to write primarily RL zì strings. The test provides significant evidence that instances of using UC to write RL zì strings are in the majority and that the proportion of UC that are RL zì is at least 0.91 ($\chi^2 = 425.9, p < .001$). From this, I conclude that UC is used more frequently for writing RL zì strings rather than FS strings.

Mixed case is used to write FS strings exclusively in the sample (100%), as expected. Contrary to my expectations, however, LC is used to write both RL zì and FS strings. Further, RL zì strings are written with LC more frequently than FS strings in the sample (68%). A one-sample proportions test determines whether the proportion of LC used to write RL zì strings is actually in the majority. The test provides significant evidence that the proportion of LC used to write RL zì strings is at least 0.51 ($\chi^2 = 3.23, p = .0362$).
Thus, I infer that over half of roman letter strings written with LC are RL zi strings rather than FS strings.

### 6.9.3 Upper case writes initialisms

In the previous section, I showed that UC is used to write primarily RL zi, and in section 6.5, I demonstrated that most RL zi strings are initialisms. Thus, I have suggested that many UC strings may be initialisms. The data on case combinations used for writing initialisms are shown in Table 6.7.

<table>
<thead>
<tr>
<th>Case</th>
<th>Total Strings</th>
<th>Initialisms</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>only UC</td>
<td>580</td>
<td>401</td>
<td>0.69</td>
</tr>
<tr>
<td>only LC</td>
<td>31</td>
<td>1</td>
<td>0.03</td>
</tr>
<tr>
<td>Total</td>
<td>611</td>
<td>402</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Table 6.7 Case Combinations Used to Write Initialisms in XIN and CNA

The table shows that all but one example of an initialism in the sample is written in UC and that of the 580 UC strings in the sample, 401 are initialisms. A one-sample proportions test was used to show whether the majority of UC strings are initialisms. The results indicate that the proportion of UC strings that are initialisms is at least 66% ($\chi^2 = 84.21, p < .001$), allowing me to conclude that over half of UC strings are initialisms.

### 6.10 Characteristics of [M N] nominals

#### 6.10.1 Roman letter modifiers are preferred
I have suggested that in hybrid strings that are [M N] nominals (that is, compounds or ‘zero’ de phrases), the modifier, rather than the head, would tend to be expressed with roman letters. The data for lettered modifiers and heads in the XIN and CNA sample are shown in Tables 6.8 and 6.9. Hybrid strings are categorized into two types, strings that are [M N] nominals and strings that are not [M N] nominals. [M N] nominals are further divided into strings with a roman letter modifier and strings with a roman letter head.

<table>
<thead>
<tr>
<th>Hybrid String Type</th>
<th>Tokens</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>[M N] nominal</td>
<td>412</td>
<td>0.94</td>
</tr>
<tr>
<td>Not [M N] nominal</td>
<td>25</td>
<td>0.06</td>
</tr>
<tr>
<td>Total</td>
<td>437</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.8 Hybrid String Types in XIN and CNA

<table>
<thead>
<tr>
<th>[M N] Nominal Type</th>
<th>Tokens</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has RL modifier</td>
<td>382</td>
<td>0.93</td>
</tr>
<tr>
<td>Has RL head</td>
<td>30</td>
<td>0.07</td>
</tr>
<tr>
<td>Total</td>
<td>412</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.9 RL Modifiers and Heads in [M N] Nominals in XIN and CNA

Table 6.8 shows that the vast majority of hybrid strings are [M N] nominals (94%) and that only a small percentage is other kinds of hybrid strings (6%). Among the [M N] nominals, most have a roman letter modifier (93%) rather than a roman letter head (7%). A one-sample proportions test provides significant evidence that the proportion of [M N] nominals is in the majority ($\chi^2 = 299.03, p < .001$). The proportion of [M N] nominals
with a roman letter modifier is at least 0.9, which leads me to conclude that the majority of [M N] nominals in XIN and CNA have a roman letter modifier rather than a roman letter head.

6.10.2 Modifiers are composed of RL zi

In the previous section I showed that [M N] nominals tend to have roman letter modifiers rather than roman letter heads. Now I address whether the roman letter modifiers are likely to be RL zi or fully spelled. Table 6.10 shows the number of [M N] nominals in the sample that have modifiers formed with RL zi and the number of [M N] nominals those that have modifiers formed with FS units.

<table>
<thead>
<tr>
<th>Modifier Type in [M N] Nominals</th>
<th>Tokens</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL zi modifier</td>
<td>364</td>
<td>0.95</td>
</tr>
<tr>
<td>Fully spelled modifier</td>
<td>18</td>
<td>0.05</td>
</tr>
<tr>
<td>Total</td>
<td>382</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.10 Roman Letter Modifier Types in XIN and CNA

The data in Table 6.10 indicate that the majority of [M N] nominals in the XIN and CNA sample have RL zi modifiers (95%) rather than fully-spelled modifiers (5%). A one-sample proportions test shows that [M N] nominals with an RL zi modifier are in the majority and that the proportion of [M N] nominals with an RL zi modifier is at least 0.93
Thus, I infer that the majority of [M N] nominals in XIN and CNA have RL zi modifiers rather than fully spelled modifiers.

6.10.3 Heads are composed of RL zi

Just as the modifiers in [M N] nominals tend to be formed with RL zi, I have suggested that the roman letter heads in [M N] nominals tend to be formed with RL zi as well, rather than with fully spelled units. Table 6.11 gives the data for roman letter heads in [M N] nominals.

<table>
<thead>
<tr>
<th>Head Type in [M N] Nominals</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>RL zi head</td>
<td>30</td>
</tr>
<tr>
<td>Fully spelled head</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>30</td>
</tr>
</tbody>
</table>

Table 6.11 Roman letter heads in [M N] nominals in XIN and CNA

The XIN and CNA sample contains 30 tokens of [M N] nominals that have a roman letter head, as compared with 382 that have a roman letter modifier. All 30 of the [M N] nominals containing a roman letter head have one formed with RL zi rather than with a fully spelled unit. An exact binomial test was used to show whether [M N] nominals with an RL zi head are in the majority. (This statistical test is preferred when the sample size is small because other tests assume a Normal distribution, which is not satisfied here.) The test confirms that the proportion of [M N] nominals with an RL zi head is more than half (p < 0.001) and that it is at least 0.9. I conclude from the results that the majority of
roman letter heads in [M N] nominals in XIN and CNA in are formed with RL zi rather than fully spelled units.

6.11 Prevalence of RL zi in hybrid strings

In the previous two sections, I showed that RL zi tend to be used to form both modifiers and heads in [M N] nominals in XIN and CNA. I have suggested that hybrid strings in general may contain primarily RL zi rather than FS units. The data for hybrid strings are presented in Table 6.12.

<table>
<thead>
<tr>
<th>Hybrid String Type</th>
<th>Tokens</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Formed with RL zi</td>
<td>319</td>
<td>0.73</td>
</tr>
<tr>
<td>Formed with FS unit(s)</td>
<td>118</td>
<td>0.27</td>
</tr>
<tr>
<td>Total</td>
<td>437</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.12 RL Zi and FS Units in Hybrid Strings in XIN and CNA

Table 6.12 shows that the number of hybrid strings containing RL zi outnumbers those containing fully spelled units; 73% of hybrid strings in the sample contain RL zi, as compared with 27% that contain fully spelled units. A one-sample proportions test shows whether hybrid strings containing RL zi are actually in the majority. There is significant evidence that the proportion of hybrid strings containing RL zi is more than half (χ² = 91.53, p < .001), and that it is at least 0.69. This leads me to conclude that majority of hybrid strings contain a roman letter component composed of RL zi rather than fully spelled units.
6.12 Contextual support for initialisms

6.12.1 Translations for initialisms

I have proposed that when translations are provided for initialisms, they tend to be given for those initialisms that are more than one letter in length. Table 6.13 provides data for initialisms in the sample that have accompanying translations.

<table>
<thead>
<tr>
<th>Type of Initialism (Number of RLs)</th>
<th>Initialisms with a Translation</th>
<th>Tokens</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Roman Letter</td>
<td></td>
<td>22</td>
<td>0.19</td>
</tr>
<tr>
<td>&gt; 1 Roman Letter</td>
<td></td>
<td>96</td>
<td>0.81</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>118</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 6.13 Translations for Initialisms in XIN and CNA

Table 6.13 shows that of the 118 initialisms that have translations, less than a fifth contain just one letter (19%) and the rest contain more than one letter (81%). A one-sample proportions test indicates whether the proportion of one-letter initialisms that have a translation is less than half. There is significant evidence that the proportion is less than 0.5 ($\chi^2 = 45.16, p < .001$). At most 26% of one-letter initialisms in XIN and CNA have an accompanying translation. I conclude from these results that the majority of one-letter initialisms do not have an accompanying translation.
6.12.2 Transcriptions for initialisms

I observed that in XIN and CNA, initialisms appeared not to be given a transcription because the transcription would “duplicate” pronunciation information already inherent in the letter names associated with the letters in initialisms. The sample data from XIN and CNA contains only one initialism, \textit{MD-11}, with an associated transcription, 麦道 \textit{màidào} (MD < McDonnell-Douglas). \textit{Màidào} is not used to pronounce the letters \textit{M} and \textit{D}, however; it is used to indicate what \textit{MD} abbreviates.

A one-sample proportions test shows that the proportion of initialisms in XIN and CNA with accompanying transcriptions is at most 0.013, or about 1\% ($\chi^2 = 395.02, p < .001$), leading me to conclude in that in XIN and CNA, most initialisms do not have associated transcriptions.

6.13 Contextual support for fully spelled strings

6.13.1 Translations for fully spelled strings

Examination of the XIN and CNA sample shows that when translations are provided in newswires, they appear to be given for FS strings that are not proper names. Table 6.14 presents the relevant data.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Have translation</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>All FS Strings</td>
<td>127</td>
<td>62</td>
<td>0.49</td>
</tr>
<tr>
<td>FS Proper Names</td>
<td>76</td>
<td>14</td>
<td>0.18</td>
</tr>
</tbody>
</table>

Table 6.14 Translations for Fully Spelled Strings in XIN and CNA

145
Table 6.14 shows that the sample contains 127 FS strings, of which 62 have a translation (49%). Of the total FS strings, 76 are proper names, and 14 of these have a translation (18%). A 1-sample proportions test shows that the true proportion of proper name FS strings with a translation is less than half but may be as high as 28% ($\chi^2 = 29.07, p < 0.001$). This allows me to conclude that the majority of FS strings that are proper names do not have an accompanying translation.

6.13.2 Transcriptions for fully spelled strings

I observed that most transcriptions in the XIN and CNA sample appeared to be provided for proper names. Table 6.15 gives the data for proper names and transcriptions.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Have transcription</th>
<th>Proportion</th>
</tr>
</thead>
<tbody>
<tr>
<td>All FS Strings</td>
<td>127</td>
<td>53</td>
<td>0.42</td>
</tr>
<tr>
<td>FS Proper Names</td>
<td>75</td>
<td>51</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 6.15 Transcriptions for Fully Spelled Strings in XIN and CNA

Table 6.15 shows that of the 53 transcriptions for fully spelled strings in the sample, 51 were associated with proper names. (Two transcriptions were associated with proper names that are initialisms, as discussed above.) A 1-sample proportions test determines whether the proportion of FS strings with transcriptions is greater than 50%. The true proportion of proper name FS strings with a transcription is shown to be at least 58% ($\chi^2$).
= 9.01, \( p = 0.0013 \)), supporting the conclusion that transcriptions in XIN and CNA are usually associated with fully spelled strings that are proper names.

6.14 Discussion

6.14.1 Growth in the use of roman letter strings in CNA

The total number of roman letter strings in CNA and XIN was similar in 1991 and 1992 but diverged quickly after the early 1990s, with the use of roman letter strings in CNA increasing sharply by 2006. I suggest that there may be a sociopolitical explanation for the growing use of roman letter strings in CNA from the 1990s on, that is, a change in the leadership of Taiwan during the period under study and corresponding changes in Taiwan’s national identity. These influences may have contributed to the increasing use of the roman letter strings in CNA from 1991 to 2006.

Prior to the 1990s, Taiwan was under martial law from 1949 to 1987 and was governed by the Nationalist Party (KMT), which stressed Taiwan’s ties with the Chinese mainland and the promotion of Mandarin as the national language. Mandarin was considered to be the only politically legitimate language, and local language varieties such as the Southern Min dialect varieties variously known as Taiwan Amoy, Nan Min, Taiyu, and Taiwanese were suppressed. In 1988, Lee Teng-hui became the first native-born president of Taiwan. Lee implemented a drive toward ‘localization’ (Hsiau 1997) in which Taiwan’s local history and culture were promoted over the KMT’s longstanding focus on mainland Chinese roots. One aspect of this drive was the Taiyu language movement, which aimed to establish the legitimacy of the Taiwanese variety of Southern
Min Chinese spoken by the majority of native Taiwanese and to loosen the grip of the Mandarin-only language policy (Hsiau 1997). In 2000, Chen Shui-bian, another native-born Taiwanese, was elected to be Taiwan’s first non-KMT president. Chen continued to promote the Taiyu language movement and was reelected to a second term in office in 2004. Lee and Chen, Taiwan’s first native-born political leaders, held power during the time period that corresponds to the sixteen-year CNA corpus examined for this study. Both deemphasized the role of Mandarin as the sole legitimate language in Taiwan and paved the way for more openness in language use in Taiwan, including the growing use of English. The political changes in Taiwanese society during the time that Lee and Chen were in office are most likely reflected in the greatly increased use of roman letter strings in the CNA corpus.

The Taiyu language movement has also been tied to the establishment of a new Taiwanese social identity separate from that of mainland China that also contributes to increased use of English and, correspondingly, the increased use of roman letter writing. The popularity of 台湾英语 Táiwān Yīngyǔ ‘Taiwanese English’ has been an important aspect of Taiwan’s evolving identity, since English is considered to be an alternative to Mandarin. While the use of Taiyu represents a move toward ‘localization’, the use of English represents a corresponding move toward ‘internationalization’. Both strategies serve to distinguish Taiwanese society from that of mainland China. Kowal (2002:4) explains Taiwan’s motivation for embracing English as follows:

“The unpleasant implication of staying entirely within a Chinese-speaking commercial fold is that Taiwan becomes economically bound to China. The
alternative is to develop stronger economic ties with the rest of the world by accelerating the use of the international medium of English. In a time of economic uncertainty, the lure of financial gain through the use of Chinese as the main language of commerce serves as bait in a trap. It also paints Taiwan into not only a linguistic corner, but a political one as well.”

English provides Taiwan with a means of distancing itself from the Chinese mainland by giving Taiwanese people entry into the international community of English speakers (Kowal 2002), that is, the speakers of ‘world English’. As a member of this community, Taiwan is able to build trading relationships around the world, thereby maintaining an independent position in the international community economically and socially, if not politically.

The popularity of English as a means of establishing a new Taiwanese identity and Taiwan’s desire to strengthen its alignment with the international community and to maintain some distance from mainland China through the use of English are most likely also reflected in the increased use of roman letter strings in the CNA corpus. These factors may help to account for the dramatically increasing use of roman letter strings in the CNA corpus during the sixteen-year period under study. In particular, these changes may help to explain why the number of roman letter strings in CNA and XIN were roughly similar in the beginning years of the corpus and why their use increased markedly in CNA over the sixteen-year period. The time period covered by the corpus coincides with extensive changes in Taiwan’s political and social climate and with its acceptance of English as one of the local languages used in Taiwan.
6.14.2 Growth in the use of roman letter strings in XIN

The use of roman letter strings increased in XIN as well during the sixteen-year period, but at a slower pace than in CNA. From 1991 to 1998, the growth in use was flat, but picked up pace from 1999 to 2006. This two-part trend in the use of roman letter strings in the PRC most likely also reflects the effects of sociopolitical developments in the PRC, just as the rapid increase in CNA most likely reflects sociopolitical developments in Taiwan.

China’s role as a important player in the international community was strengthened during the sixteen-year period under study. Significant developments reflecting China’s increasing internationalization occurred during the years of the XIN corpus that show an upswing in the use of roman letter strings. Some examples of these events include the following: the former British colony of Hong Kong was returned to China in 1997, and the former Portuguese colony of Macau was returned in 1999; in July 2001, Beijing was awarded the 2008 Olympic Games, and in December 2001, China became a member of the World Trade Organization. The SARS epidemic broke out in China in late 2002 and spread to other countries by early 2003, requiring international cooperation to control. As these and other important events unfolded, China became an increasingly visible and influential member of the world community, and English continued to gain importance in Chinese society as a tool for pursuing the goals of modernization and internationalization (Adamson 2002). The nation’s growing focus on the role of English in China’s development is most likely reflected in the increased use of roman letter strings in XIN newswires from 1998 to 2006.
As for the comparatively slow rate of increase, I propose two possible reasons why the use of roman letter strings did not increase as quickly in XIN as in CNA. First, in contrast to the localization movement in Taiwan, mainland China maintained a strict language policy during the period under investigation, promoting Mandarin as the lingua franca of its large population of speakers of varieties of both Chinese and non-Chinese languages. The Chinese government fervently believes that widespread use of Mandarin as the national language maintains the unity of the Chinese nation and helps to give Chinese people a coherent national identity (Rohsenow 2001). Since China has many speakers of non-Mandarin varieties of Chinese, however, promoting Mandarin as a lingua franca nationwide has been a struggle, and the media is charged with the responsibility of assisting in this effort by serving as a model user of the standard language (Zhang 1999), a role that was confirmed anew in the National Language Law passed in 2001, which requires broadcast and print media to use the national language. The government’s focus on promoting Mandarin as the national, standard language and the media’s role in this effort may lead to much more conservative and limited use of English in official news writing than in Taiwan. Since the PRC media, including Xinhua, has the responsibility of using and promoting Standard Mandarin, it may have fewer opportunities to use English terms and phrases than Taiwan’s Central News Agency. Rather, my corpus study sample revealed that Xinhua news writing consisted primarily of Chinese text with a smattering of roman letter terms provided either because they are foreign words commonly expressed in roman letters in Chinese (DNA, APEC) or because they provide
supplemental information to readers (such as foreign names written in the original roman letters and given together with the transcribed Chinese version).

A second reason why the rate of increase is considerably slower in XIN than in CNA concerns the two societies’ overall level of familiarity with English. For reasons I discuss in the section on fully spelled strings, I believe the societal level of English familiarity is lower in mainland China than in Taiwan. This may also be reflected in the low number of roman letter strings in Xinhua newswires and a comparatively slower rate of growth in their use over the sixteen-year period under study. Generally speaking, fewer average readers of Xinhua newswires may be able to comprehend a wide variety of roman letter words and phrases than average readers of CNA newswires, which may contribute to a more limited and conservative use of roman letter strings in XIN newswires than in CNA newswires.

6.14.3 Numerous RL zi strings, fewer fully spelled strings

RL zi strings are considerably more frequent than FS strings in both the XIN and the CNA newswires. I suggest that this is because RL zi are most congruent with Chinese word formation patterns and with the Chinese writing system. Although FS strings increase in both CNA and XIN during the sixteen-year period, RL zi strings are the “default” roman letter units in the newswires of both news agencies. RL zi strings are the most easily assimilated type of roman letter unit that can be used in Chinese texts: the letters fit into the frames of Chinese writing, they can be pronounced with their individual letter names, they tend to be short as compared with FS strings, and they can be written
without the use of mixed case letters. (I discuss the length of RL ｚi strings and the use of upper, mixed, and lower case letters below.) I suggest that it is primarily for these reasons that both XIN and CNA newswires use a large number RL ｚi strings.

At the same time, the corpus data show that fully spelled strings are more common in CNA than XIN in each year sampled (1991, 1998, 2005) and that their use grew much more quickly in CNA over the course of those two intervals than in XIN. The more frequent use of FS strings in Taiwan may be facilitated by the Taiwan population’s generally greater familiarity with English as compared with the population of mainland China. A crucial factor that I believe contributes to Taiwanese individuals’ relatively greater facility with English is Taiwan’s focus since the 1960s on establishing strong international trade ties, which of course require communication in English. Evidence of this emphasis can potentially be seen, for example, in Taiwan’s memberships in international trade organizations, including the Asian Development Bank, the Asia-Pacific Economic Cooperation (APEC) Forum, and the World Trade Organization, among others. (APEC is one of the most frequent lettered words in both the CNA and the XIN corpus.) Taiwan has also enjoyed close political relations with the international community since the 1950s, and those ties also require the use of English to maintain. For example, Taiwan held the Chinese seat at the United Nations until 1971 and had diplomatic relations with the United States until 1979. Since 1979, the United States has continued to maintain quasi-diplomatic relations with Taiwan. I suggest that these ties have helped to foster people-to-people contacts with the English-speaking international
community, such as educational exchanges, overseas study, tourism, and the like, that have encouraged familiarity with English in Taiwan.

The result of Taiwan’s well-developed network of international trade and political relations is that even though Mandarin was promoted aggressively as the national language in Taiwan, Taiwanese people have most likely had to learn English to establish and grow these relationships. Throughout the decades after the end of World War II, the Taiwanese have used English to engage in foreign trade, manufacturing, international affairs, international education, and other activities that have sustained Taiwan’s position in the world community. In addition, between 1945 and 1970, there were U.S. military bases in Taiwan, and American soldiers had R&R there during the Vietnam War. The effect with regard to news writing may be that a larger number of readers of CNA newswires may have a sufficient level of fluency in English to deal with fully spelled strings than readers of XIN newswires. Since Taiwanese people have used English for a variety of international endeavors during the last several decades, Taiwan’s base level of societal familiarity with English is most likely higher than that of mainland China.

In contrast, Chinese citizens in the PRC have had considerably less exposure to English in China’s closed political climate from the 1950s until the end of the Cultural Revolution in 1976. Many people also had their schooling interrupted during the Cultural Revolution, and wide-scale teaching of English began only in the 1980s. For the most part, it is young individuals who are doing the learning rather than older people (Lam 2002). Generally speaking, individuals in their late forties and above have little familiarity English (Lam 2002), and an obvious generation gap exists between the young,
who have had exposure to English in school, and the old, who may have learned Russian as a foreign language. The urban-rural divide in education and exposure to foreign influences also contributes to the gap. Thus, it is my view that the base level of familiarity with English in China may be much less overall than in Taiwan. The result with regard to XIN newswires is that many readers may lack the fluency needed to deal with a wide variety of FS strings. During the time period under investigation, the average educated citizen that read Xinhua newswires may have been better equipped to comprehend a variety of RL zì strings than a variety of FS strings.

6.14.4 Initialisms vs. spelled words in CNA and XIN

The majority of RL zì in the XIN and CNA sections of the corpus are initialisms. Many of these are technical terms, and I propose that they can essentially be divided into two types. First, some are terms that are commonly used in newswires aimed primarily at readers knowledgeable about particular topics, and others are terms commonly used in newswires of interest to a more general audience as well. Examples of the former include military abbreviations, such as the names of military aircraft (e.g. F-15 (F < Fighter)), computer-related terms (e.g. IC (< integrated circuit)), sports terms (e.g. NBA), and terms pertaining to economics, trade, and business (e.g. GDP, GATT, BOT (< build, operate, transfer)). Examples of the latter include the names of well-known international organizations and companies (e.g. WTO, IBM), specialized terms that are relevant to everyday life (e.g. DVD, LCD), and technical terms that become topical as a result of current events (e.g. SARS, H5NI (avian influenza)). Second, many technical terms are
used infrequently, but they are preferred in the texts to the longer fully-spelled versions of the names or phrases they represent. The abbreviations are much shorter and may be supplemented with a translation in the text. Examples include \textit{CPI} (< consumer price index) and \textit{TGV} (Train à Grande Vitesse).

The use of initialisms representing technical terms in writing appears to allow the Chinese news agencies to import a wide variety of international terms while avoiding the longer, more cumbersome, and more difficult-to-understand fully-spelled versions. This gives the news agencies, and their readers, a chance to participate in using terms that are current in world English, while also limiting their selection to those that are most easily understandable to average speakers.

Figure 6.2 showing the proportion of probable initialisms vs. spelled words in XIN and CNA and Figure 6.3 showing the proportion of probable initialisms in XIN and CNA provide further evidence that RL \textit{zi} strings are the default type of roman letter strings in XIN and CNA rather than FS strings. The prevalence of probable initialisms in both XIN and CNA is an important manifestation of this phenomenon, since probable initialisms far outnumber FS words in both XIN and CNA during each year of the corpus. The difference between CNA and XIN lies primarily in their use of spelled words. While the use of spelled words grew quickly in CNA from 1998 to 2005, it grew much more slowly in XIN.

The possible differences in CNA and XIN readers’ familiarity with English proposed earlier may help to explain the relative popularity of spelled words in CNA and their limited use in XIN. This difference may be reflected in the use of spelled words in the
two parts of the corpus and may provide further evidence that average readers of CNA newswires were better able to handle fully spelled strings (spelled words in this case) than average readers of XIN newswires during the sixteen-year period under study.

Perhaps a better way to explain this is from the point of view of the news agencies themselves. The Central News Agency may have believed that there were a sufficient number of readers who would understand spelled words that such words were increasingly included in newswires. In contrast, Xinhua may have believed that there were a growing number of readers who would understand a variety of initialisms but many fewer readers who would understand a large number of spelled words. As a result, Xinhua may have increased the number of initialisms used in newswires but may have continued to limit the number of spelled words that were used.

Extrapolating away from the newswires to Chinese society in general, interest in English in mainland China appears to have been manifested primarily in the growth of initialisms during the sixteen-year period. And although the increase in the use of spelled words was flat until 1998, it also began to increase after that time, which may reflect the PRC’s gradually increasing societal familiarity with English. In contrast, in Taiwan, the use of both initialisms and spelled words increased dramatically after 1998. The push toward localization and internationalization described earlier is most likely reflected in this dramatic increase. Moreover, if Taiwan had a higher level of societal familiarity with English during the sixteen-year period as I have proposed, that familiarity would have facilitated the growth of both initialisms and spelled words.
6.14.5 Short length of roman letter strings

My data analysis indicates that RL \( zì \) strings tend to be shorter than FS strings and that the length of RL \( zì \) strings is less variable than the length of FS strings. An important reason for these findings is that FS strings may include multiple words written in roman letters, especially proper names and phrases, whereas RL \( zì \) strings tend to consist of initialisms, alphanumeric strings, and individual roman letters in hybrid words, all of which generally use a much smaller number of roman letters.

I have noted that there may be a general preference for short roman letter strings in the newswires. Figures 6.5 and 6.6 showed that most roman letter strings in both XIN and CNA have three or fewer letters, which provides support for my claim that short letter strings are preferred to longer ones. Another reason why roman letter strings appear to be short is that the count of their length does not include the Chinese \( zì \) that may be attached to them form a word. Thus, a word like POS 机 ‘POS machine’ (POS < point of sale) is formed with four \( zì \) rather than just the three RL \( zì \) that I counted. My study measured only the length of roman letter strings themselves and did not measure the total length of hybrid compounds or phrases. I was interested specifically in the length of the roman letter portion of hybrid words and phrases, but the total length of these types units could be considered as well. The fact that Chinese \( zì \) may be added on to the roman letters may be another motivation for keeping the roman letter portion short, since the Chinese \( zì \) add length to the roman letter portion.

Figure 6.4 showed that in addition to short RL strings, there are some longer roman letter strings in both XIN and CNA as well. I suggest that longer roman letter strings are
more likely to be *hapax legomena*, or roman letter strings appearing in only one newswire, such as websites, phrases, or proper names mentioned in the article and possibly accompanied by a translation or transcription. I consider such strings to be incongruous with Chinese grammar and writing in comparison with RL *zì* strings. Such hapaxes can be thought of as code switches, or insertions, that provide parenthetical information in newswires for those who can understand that information. Those who do not understand it simply skip over it and read just the translation or transcription if one is provided. Thus, even though there are some longer roman letter strings in the newswires, these strings may simply provide “optional,” supplemental information in the newswire, whereas shorter strings may have a better chance of being used more frequently.

### 6.14.6 Why upper case is preferred

In Table 6.5, I demonstrated that upper case letters are used more frequently than either mixed case or lower case letters. Upper case is used to write most RL *zì* strings (Table 6.6), of which there are many in the corpus. Moreover, upper case is used to write initialisms (Table 6.7), one of the main types of RL *zì* strings, which are also numerous in the corpus and are not often written with lower case or mixed case letters. The fact that initialisms and other types of RL *zì* are generally written with upper case letters contributes substantially to the high frequency of upper case letters in the newswires.

The sample also included 41 fully spelled strings (7%) that were written with all upper case letters (e.g. *AUSTRALIA ASIA AIRLINE*). These strings would generally be written with mixed case letters in English (*Australia Asia Airline*), since all upper case is used
primarily for emphasis. The presence of these unlikely upper case strings leads me to believe that the newswires reflect what may be less attention in Chinese writing to the functions of different letter case combinations than is common in English writing. These functions may be less meaningful in Chinese writing, in part because it does not differentiate upper and lower case when writing Chinese characters, and they may be leveled in favor of all capital letters.

Table 6.9 showed that in hybrid strings that are [M N] nominals tend to have modifiers expressed in roman letters rather than heads expressed in roman letters. Thus, modifiers formed with roman letters tend to be preferred to heads formed with roman letters in [M N] nominals, as I expected. This supports my view that by being in the modifier position, roman letter units play a secondary role in [M N] nominals. Conversely, by being in the head position, Chinese zi units play the primary role. The Chinese zi heads are established categories of nouns in Chinese, and the roman letter units simply modify those nouns. The whole item then becomes a new member of the ‘head family’ represented by the Chinese zi head.

Concepts expressed in roman letter units may also be too narrow or specialized in meaning to serve as heads frequently. For example, while WTO is a common initialism found in XIN and CNA, there may be fewer chances to use it to create an [M N] nominal of the form ‘__ WTO’, which uses it as a head, than one of the form ‘WTO __’, which uses it as a modifier. Similarly, single RL zi may also serve as modifiers more easily than heads. The sample includes multiple tokens such as A 组 ‘group A’, X 光 ‘X-ray’, and C 型 ‘type C’, in which the Chinese zi is the head, but none of the form ‘__ A’, ‘__ X’, ‘__
C’, etc. in which the RL zì is the head. This stands in contrast to Chinese morphemes such as 车 ‘vehicle’ that are commonly used to form compounds in which the morpheme is used as either the modifier or the head (e.g. 火车 ‘train’, lit. ‘fire-vehicle’ and 车牌 ‘license plate’ lit. ‘vehicle-plate’ respectively). Single letter morphemes generally take on only a small and restricted set of meanings, and they are usually not nouns, such that they are unlikely candidates to be used as heads. It is difficult to know what the head X in ‘光 X’ means unless X is acting as an abbreviation of a noun and the individual knows what it abbreviates, but one can figure out more easily what ‘X 光’ means based on the customary meanings of X in compounds.

6.14.7 Preferred modifiers and heads

Table 6.11 showed that in the few instances when [M N] nominals do have roman letter heads, those heads tend to be formed with RL zì units rather than fully spelled units, and Table 6.12 showed that RL zì are most common in hybrid strings overall. This leads me to believe that heads formed with RL zì are preferred to heads formed with fully spelled units and that RL zì are preferred for use in both [M N] nominals and in hybrid strings overall. I have argued that RL zì are most congruent with Chinese zì and are most easily adapted to their combinatorial properties, that is, to their status as what Chao (1968a) calls ‘start-free’, ‘start-bound’, ‘end-free’, and ‘end-bound’ morphemes. I suggest that RL zì are more easily paired with Chinese zì of any of these four types than fully spelled units. It can be expected that RL zì will serve as components of hybrid strings more frequently than fully spelled units.
6.14.8 Translations and transcriptions for RL strings

If translations are provided for initialisms in the corpus, they are most likely be provided for those that are two or more letters in length, as shown in Table 6.13. This finding suggests that 1-letter initialisms may tend to be treated simply as ordinary zi not requiring supplementary explanation in the form of a translation. In contrast, longer initialisms do tend to be given translations. They may be viewed as more opaque linguistic units requiring additional information in the form of a translation.

I proposed that when translations are provided in newswires, they tend to be given for items that are not proper names. This was confirmed by the data in Table 6.14. Most FS strings that were proper names did not have an accompanying translation. The reason for the lack of translations for proper names is that they may be difficult to convert into meaningful Chinese morphemes. For example, it is difficult to translate Michael into one or more Chinese morphemes that mean ‘Michael’, so a transcription is used in such cases.

My analysis overlooks one issue, however, which is not revealed in Table 6.14 but which impacts whether a translation is provided for a proper name. The decision to translate is not based simply on whether the item is a proper name but rather on whether one or more of the morphemes in that item are transparent to Mandarin speakers and can be translated easily. For example, 冰岛 bīngdǎo ‘Iceland’ lit. ‘ice-island’ is a proper name but lends itself to being translated, while Michael is also a proper name but is unlikely to be translated. My study did not consider the transparency of the morphemes in proper names and its relationship to translations, but transparency is certainly a factor that needs to be studied with regard to whether a roman letter item is translated or transcribed.
In Table 6.15 I showed that when transcriptions are given for fully spelled strings, they tend to be given for proper names. This relates to the question of whether the morphemes in a proper name are transparent enough to be translated. It may be that my sample included a relatively large number of proper names that have non-transparent morphemes. If this were the case, transcriptions may have been given for these proper names rather than translations.
CHAPTER 7

CONCLUDING REMARKS

7.1 Summary of findings

This dissertation has investigated lettered words in contemporary Mandarin in the PRC and Taiwan, many of which are either borrowed from English or based on English words. I examined initialisms such as WTO and also hybrid words such as 维生素 A ‘vitamin A’ that consist of both letters and Chinese characters. I also discussed the crucial issue of why initialisms are more common as lettered words in Mandarin than fully spelled words. The study focused on lettered words in written discourse because they are found often there. My analysis involves extensive examination of the written form of lettered words and the role of the script in reflecting morphological and semantic structure and the outcomes of processes of lexical borrowing.

I investigated four aspects of lettered words pertaining to their formation and use in written Chinese discourse. I first considered the characteristics of the Chinese lexicon and the writing system and their influence on the selection and assimilation of lettered words into Mandarin. I noted that the primary unit of metalinguistic awareness in Chinese is the zi, that is, the morpheme-syllable and the character with which it is written. Although
some Chinese words consist of just one zi, in modern Mandarin, many words consist of hierarchically organized strings of zi, which are written as visually discrete characters that occupy an imaginary equidimensional square. I pointed out that initialisms and hybrid words embody just these defining characteristics of Chinese zi. The letters can be pronounced individually with their letter names; they can represent morphemes or symbols with individual meanings; and each letter fills an equidimensional square. I argued that letters as used in initialisms and hybrid words overcome the mismatch in the morphophonemic writing system of English and the morphosyllabic writing system of Chinese since they are essentially used either as syllables or as morpheme-syllables in the same manner as Chinese zi. I showed that letters in initialisms and hybrid words are congruous with Chinese zi in terms of their pronunciation and combinatorial properties, unlike ordinary spelled words, in which letters are linked together to pronounce the word and which are not visually discrete, that is, not separated by white space. The congruence of letters in initialisms and hybrid words with Chinese zi makes them more amenable to integration into Chinese than their spelled counterparts. In them, Chinese speaker/readers to “find something understandable in terms of their own patterns” (Herskovits 1938, 5f.), which encourages acceptance.

The second issue I considered is the status of lettered words as products of language contact. I described the traditional borrowing processes in Chinese that rely on the use of one or more Chinese zi to ‘sinicize’ foreign words and concepts, including phonetic adaptation, semantic adaptation in the form of loan translation and semantic translation, and the simultaneous use of phonetic and semantic adaptation. I also noted that different
processes can be used to adapt different morphemes in foreign words and that explicative adjuncts can be added to specify the semantic category of the term. I suggested that lettered words represent an extension of these traditional processes. The letters are sinicized in terms of pronunciation, since they are normally pronounced with Mandarin syllables, but their written form is imported directly. The meanings of certain individual letters are also imported directly, and some lettered words are augmented with explicative adjuncts.

I introduced the notion of ‘single letter units’ in English, that is, letters used as morphemes or symbols with conventional meanings. I suggest that the meanings and functions of single letter units have been imported into Mandarin (A meaning ‘first, top, best’) so that roman letters used as single letter units do not need to be translated. Single letter units form a new set of zi in Mandarin, roman letter zi, that represent a point of congruence between the grammatical and writing systems of Chinese and English. Initialisms and most hybrid words contain RL zi and are readily integrated into Mandarin. In contrast, spelled words are not composed of RL zi and are not assimilated as easily. As a result, English compounds containing both single letter units and spelled components (e.g. BB gun) are adapted so that the single letter units are imported as RL zi and the spelled components are adapted with Chinese zi. I also noted that a recurring pattern in hybrid compounds is that they are endocentric and that the modifier tends to be a roman letter component and the head a Chinese zi component. I suggested that roman letter components are more easily integrated into Chinese in a secondary role as modifiers rather than in the primary role as heads.
The third issue I investigated is sociopolitical factors that may contribute to differences in the use of lettered words in the PRC and Taiwan. These factors may help to explain why some types of lettered words are more common in one Chinese society than the other. I suggested that Taiwan’s contemporary history as a major trading power contributed to a longer and more intense period of Mandarin-English language contact than in the PRC and resulted in a higher level of societal familiarity with English. In contrast, the PRC’s relative isolation from the international community until the late 1970s has led to a shorter period of Mandarin-English language contact and a correspondingly lower level of societal familiarity with English. I argued that this difference is reflected in the two societies’ acceptance of spelled words, which require a higher level of English fluency to comprehend and use actively. I suggested that Taiwan has a greater use of spelled words than the PRC because of Taiwanese society’s overall higher level of English familiarity.

The final issue I examined is the use of lettered words in Chinese newswires, a communicative context in which they commonly occur. I conducted a corpus study to identify patterns in the use of lettered words in newswires from the PRC and Taiwan and to examine the addition of translations and transcriptions, two types of ‘contextual support’ that are often provided alongside lettered words.

7.2 Lessons from the corpus study

The corpus study examined lettered words in newswires from the Xinhua news agency (XIN, PRC) and the Central News Agency (CNA, Taiwan) and confirmed that the
characteristics of lettered words that I posited in earlier chapters were also found in a large sample of lettered words. I also found trends in the use of lettered words in the two sets of newswires from 1991 to 2006 as well as differences in the types of lettered words used in XIN and CNA. The corpus study showed that the use of lettered words increased much faster in CNA than in XIN and that CNA contains many more fully spelled words than XIN. I suggested that both of these trends reflect Taiwan’s greater overall familiarity with English and greater facility in using a wider variety of English words. I also found, however, that the use of fully spelled strings increased slowly in XIN as well, which suggests a growing societal familiarity with English in the PRC. Despite the increased use of fully spelled strings over the sixteen-year period, the majority of roman letter strings in both XIN and CNA were still RL zì strings – initialisms in particular – rather than fully spelled strings. I suggested that this reflects a preference in Chinese for RL zì strings since they are most zì-like.

I compared the length of fully spelled strings with the length of RL zì strings and found that RL zì strings in the corpus are considerably shorter on average than fully spelled strings, which I argued indicates a preference in Chinese for short roman letter strings. I also looked at the use of three case combinations, all upper case, all lower case, and mixed case, finding that all upper case is by far the most common in the corpus. The high frequency of all upper case results from the large number of RL zì strings in the corpus, especially initialisms. I also examined the hybrid compounds in the corpus and found that they were endocentric, with a roman letter modifier and a Chinese zì head. Moreover, the roman letter modifier tended to be formed with RL zì rather than fully
spelled units, as predicted. Finally, I found that translations and transcriptions are sometimes provided for lettered words as a form of contextual support in newswires to help readers better understand lettered words. Translations were provided for initialisms longer than one letter and for fully spelled strings that were not proper names. Transcriptions were provided for fully spelled strings that were proper names, but not for initialisms. I suggested that initialisms are more likely to be translated than transcribed because speakers do not need pronunciation information for them since they pronounce them with their Chinese letter names. In contrast, proper names with opaque morphemes are likely to be transcribed rather than translated.

In summary, the corpus study revealed that there were many more lettered words overall in the Taiwan newswires than in the PRC newswires and that Taiwan newswires contained many more fully spelled strings. The study also showed, however, that lettered words formed with RL zi formed the core type of lettered word in both parts of the corpus. As fluency in English increases in the PRC and as the use of English in Taiwan continues to grow as part of Taiwan’s international orientation, the use of more fully spelled strings will most likely continue to increase in tandem with the continued use of words formed with RL zi.

7.3 Issues for future research

An important issue for future research is the pronunciation of lettered words in Mandarin. The assumption in Chinese society is that since many lettered words are borrowed from English, they are pronounced as in English. My observation is that the
situation is much more complex than this, with a variety of factors possibly affecting pronunciation. These factors need to be studied to identify patterns in the pronunciation of letter names, initialisms, acronyms, and other lettered words. Variation in pronunciation in different Chinese communities also needs to be examined.

Another important issue is how speakers’ familiarity with English, or lack thereof, affects their understanding and use of lettered words. Chinese individuals are confronted with lettered words in writing and speech in both the PRC and in Taiwan and need to deal with them whether or not they are familiar with English. It would be useful to know what strategies different speakers employ to handle these words, particularly individuals who are not fluent in English. This would be helpful for understanding aspects of variation in the pronunciation of lettered words and also how the meanings of lettered words are understood.

My study examined the use of lettered words in newswires, but news writing is not the only context in which they appear, of course. Further study is needed of the use of lettered words in other types of written discourse contexts, particularly in computer-mediated communication (CMC). It would be useful to know whether the same patterns that I have identified in newswires also apply to CMC. Since younger individuals familiar with English engage in CMC and since this form of communication is influenced by input methods for Chinese characters as well as by an emphasis on brevity, among other factors, I expect that the patterns may differ somewhat from those of the newswires. I predict, however, that the preference for RL zi is probably manifested in this form of writing as well.
An issue that I mentioned in my study but did not investigate directly is that certain lettered words have an equivalent formed with Chinese zi. ATM jī and tíkuǎnjī (‘ATM machine’) are examples. The lettered word and its Chinese zi equivalent may be used in complementary ways in speech and writing. Knowing these differences would help to clarify whether lettered words are used differently from their Chinese zi equivalents.

Finally, a comparison of lettered words in Chinese with those in other languages that use non-roman writing systems would also be illuminating. It would be particularly useful to know whether there are differences in how alphabetic writing systems assimilate lettered words and how non-alphabetic systems do so. A comparison of the use of lettered words in Chinese, Japanese, and Korean would help to determine whether there are similarities in their patterns of lettered words usage and formation.
APPENDIX A

SAMPLES OF XIN AND CNA NEWSWIRES
XINHUA NEWSWIRE

<DOC id="XIN_CMN_20051001.0017" type="story">
<HEADLINE>“我在空中见证开国大典”——一位飞行员的国庆回忆</HEADLINE>
<DATELINE>新华社济南10月1日电 (记者 邓卫华)</DATELINE>
<TEXT>
56年前的10月1日,年轻的飞行员王延洲驾机参加开国大典,亲眼见证一个新时代的到来;2005年国庆前夕,已经85岁高龄的王延洲老人回忆起56前的国庆,仍然兴奋不已。

“我们当时是代表中国人民解放军空军参加开国大典的阅兵仪式。参加开国大典的一共只有17架飞机,但意义却非比寻常,它象征着一个新军种的诞生。”

王延洲早年就读于黄埔军校,后在美国学习飞行,抗战中他曾击落过5架日本飞机,是至今仍在世的少数几位中国“王牌飞行员”之一。

开国大典是下午3点正式开始的,当毛泽东主席宣读中华人民共和国中央人民政府公告时,王延洲和他的队友们正在位于北京郊外的南苑机场待命。

“我刚刚打开飞机座舱里的收音机,就听到了毛主席宣布中华人民共和国成立的声音。56年过去了,这声音还常常盘旋在我的脑海里,就像刚刚听到一样。”

下午4点整,天安门城楼总指挥室负责人油江下达起飞令,17架飞机望见绿色信号弹呼啸升空,在北京上空盘旋待命。
17架飞机中，9架是P-51型战斗机，2架是蚊式战斗机，3架是C-46型运输机，1架是L-5型通讯联络机，最后2架是PT-19型初级教练机。17架飞机要形成一个纵队跟进队形，通过天安门上空，与地面的坦克队列相呼应。

“17架飞机5种机型，飞行速度相差很大。两种战斗机的时速是600公里，C-46型运输机的时速只有300公里，L-5型通讯联络机和PT-19型初级教练机的时速不足200公里；但上级要求，通过天安门时必须队列整齐，分秒不差，确实很有难度。”

经过反复摸索和精确计算，飞行队决定起飞的顺序按照先小后大，先慢后快，同时还专门选择了三个不同的航线。战斗机速度最快，从通县进入；运输机速度中等，从建国门和通县之间进入；其余飞机从建国门东侧进入。尽管难度很大，但我们经过多次合练后，可以保证开国大典时万无一失。”

下午4点35分，受阅分列式正式开始，9架领航的P-51型战斗机从通县拐弯进入航线，各分队保持规定的高度差和时间间隔，分别在900公尺、600公尺、450公尺的高度，由东向西依次进入，向天安门飞去。

“飞过天安门上空时，下面真是一片红旗的海洋。直到现在，我对这种鲜艳的红色都怀有一种特殊的感情。”

9架领航的战斗机飞过天安门后，油江再次下达命令：“9架P-51再通过天安门一次。”队长刑海帆按照原来预定的方案，加速由北京城西折向北郊上空，再由东向西第二次飞跃天安门上空。

174
“从时间上看,我们再次通过天安门时,正好尾随方
槐领队的L-5型和PT-19型飞机之后,配合得恰到
好处。后来,好多人以为开国大典受阅的飞机是26架,
其实后面9架是重复飞行的。”王延洲面带笑容,开心地
说。
</P>
</TEXT>
</DOC>
CENTRAL NEWS AGENCY NEWSWIRE

<DOC id="CNA_CMN_19910101.0002" type="story">
  <HEADLINE>
(專1)IBM在電腦晶片科技獲重大突破
  </HEADLINE>
  <DATELINE>(中央社譯稿)
  </DATELINE>
  <TEXT>
<p>
萬國商業機器公司(IBM)日前宣布
在基本電腦晶片科技方面獲致重大突破,可提升未來電
腦運算的速度。
</p>
<p>
去年十二月在舊金山一項科技會議上發布的這項電
腦晶片科技突破,亦有助於預防晶片的微小電路漏電,  
並減少電路受幅射影響造成錯誤。
</p>
<p>
IBM公司指出,這項進展雖不致於促成在近兩年  
內推出更快速的電腦,其重要性卻不容忽視,因為電腦  
廠商將揚棄現有的晶片設計,並開始考慮製造供本世紀  
末期使用的高性能個人電腦。
</p>
<p>
一位發言人說,目前該公司已運用此一科技製造電  
晶體或微小開關,用這些成品所生產電腦的運算速度比  
傳統式要快三倍。他說,這些新的設計製造也非常方便  
,品質與傳統的設計不相上下。
</p>
<p>
這些電晶體的製造,採用互補型金屬氧化半導體  
(CMOS)的科技。這種晶片備受歡迎,因為他們的運  
算速度較其他晶片為快、省電,同時放出的熱量也較其  
他晶片為少。該公司研究人員強調,此一新科技方法將  
不只限於用在互補型金屬氧化半導體上,未來也可用於  
其他方面。
</p>
</TEXT>
</DOC>
IBM公司宣佈的這項新科技，將進一步延長矽的壽命—矽是絕大多數電腦晶片所用的基本材料。晶片廠商喜愛用矽當材料，是因為其成本低廉，性能為業者所熟悉，然而廠商們擔心，在不久的將來，矽的功能將受到限制。而IBM公司的這項科技突破，除了減少對這方面的顧慮外，也降低了開發高成本的晶片材料半導體的迫切性。

亞利桑那州丹伯市從事市場研究的前瞻觀念公司一位職員指出，矽材料性能增進之速出乎任何人的想像。IBM這項開發技術當有助於確保矽到西元兩千年時，仍將用為製造晶片的材料。

IBM公司發現一種把矽加在一種不導電的物質上而能造成突破的方法，正是電腦科學家長久以來夢寐以求的產物。當今的矽電路大半加在矽材料上面，這種設計不但電流通電耗時久，而且也容易「誤入歧途」。

純矽質晶片產生的第一個問題是，外界的輻射可能與晶片本體互為作用而中斷了微妙的電流。IBM公司在電路與晶片的其他矽材料之間，加進一種絕緣物質(二氧化矽)以後，能有效保護電路免於中斷。此一科技新發展將裨益與國防有關和與太空有關的電子業，隨著電路體積日益縮小，性能益趨敏感，一般電子業勢將倚重此一新的科技。(尤淑雅譯自「亞洲華爾街日報」)
APPENDIX B

SAMPLE OF A CODED XIN NEWSWIRE
<table>
<thead>
<tr>
<th>Article_16</th>
<th>DOC ID: XIN_CMN_20051001.0017</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Headline</strong></td>
<td>“我在空中见证开国大典”——一位飞行员的国庆</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9 P-51</td>
</tr>
<tr>
<td>10</td>
<td>9 C-46</td>
</tr>
<tr>
<td>11</td>
<td>9 L-5</td>
</tr>
<tr>
<td>12</td>
<td>9 PT-19</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>12 C-46</td>
</tr>
<tr>
<td>16</td>
<td>12 L-5</td>
</tr>
<tr>
<td>17</td>
<td>12 PT-19</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>15 P-51</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>23</td>
<td>18 P-51</td>
</tr>
<tr>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>20 L-5</td>
</tr>
<tr>
<td>26</td>
<td>20 PT-19</td>
</tr>
<tr>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>9</td>
</tr>
<tr>
<td>12</td>
<td>9</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>11</td>
</tr>
<tr>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>16</td>
<td>12</td>
</tr>
<tr>
<td>17</td>
<td>12</td>
</tr>
<tr>
<td>18</td>
<td>13</td>
</tr>
<tr>
<td>19</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>15</td>
</tr>
<tr>
<td>21</td>
<td>16</td>
</tr>
<tr>
<td>22</td>
<td>17</td>
</tr>
<tr>
<td>23</td>
<td>18</td>
</tr>
<tr>
<td>24</td>
<td>19</td>
</tr>
<tr>
<td>25</td>
<td>20</td>
</tr>
<tr>
<td>26</td>
<td>20</td>
</tr>
<tr>
<td>27</td>
<td>21</td>
</tr>
<tr>
<td>28</td>
<td>22</td>
</tr>
<tr>
<td>Item #</td>
<td>Doc ID</td>
</tr>
<tr>
<td>-------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>186</td>
</tr>
<tr>
<td>2</td>
<td>186</td>
</tr>
<tr>
<td>3</td>
<td>186</td>
</tr>
<tr>
<td>4</td>
<td>186</td>
</tr>
<tr>
<td>5</td>
<td>186</td>
</tr>
<tr>
<td>6</td>
<td>186</td>
</tr>
<tr>
<td>7</td>
<td>186</td>
</tr>
<tr>
<td>8</td>
<td>186</td>
</tr>
<tr>
<td>9</td>
<td>186</td>
</tr>
<tr>
<td>10</td>
<td>186</td>
</tr>
<tr>
<td>11</td>
<td>186</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>0</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>10</td>
<td>9</td>
</tr>
<tr>
<td>11</td>
<td>10</td>
</tr>
<tr>
<td>Item</td>
<td>Sent.</td>
</tr>
<tr>
<td>------</td>
<td>-------</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Item #</td>
<td>Doc ID</td>
</tr>
<tr>
<td>--------</td>
<td>--------</td>
</tr>
<tr>
<td>1</td>
<td>CNA</td>
</tr>
<tr>
<td>2</td>
<td>CNA</td>
</tr>
<tr>
<td>3</td>
<td>CNA</td>
</tr>
<tr>
<td>4</td>
<td>CNA</td>
</tr>
<tr>
<td>5</td>
<td>CNA</td>
</tr>
<tr>
<td>6</td>
<td>CNA</td>
</tr>
</tbody>
</table>
APPENDIX D

TYPES OF ROMAN LETTER STRINGS
IN THE XIN AND CNA SAMPLES
XIN 1991
C1级亚军
C2级冠军和亚军
C3级亚军
卡拉OK机
BMX自行车
维生素A
维生素B
维生素C
世界一级方程式(简称F—1)赛车锦标赛
F—1赛车锦标赛
B—52轰炸机
B—52重型轰炸机
T恤衫
24K
HY—1涂料
CA型固体钽电容器
CA42型固体钽电容器
国际IC标准
CC5号样
"GNF"模型
(代表顾秉林的姓)
ACT作业者集团
"进步M—7"号货运飞船
"联盟TM—11"号飞船
"进步M—7"号
"联盟TM—11"号
X光片
三K党
《VDI—消息报》
"钢的KJT工艺研究"
DCJ32—1型测距经纬仪
Mars公司
M&M巧克力糖
540CC骨髓
SRC超低高度梁
10MSRC型钢
"V"字
农民唐XX
农民韩XX
美国交通运输部DOT标准
欧洲经济共同体ECE标准
WA全天候轮胎花纹
APPLE第四代机
a4
c2
c5
c6
d3
d7
e2
g4
g5
b线
d兵
c6兵
cde3对兵
《二硫化氨基甲酸盐萃取ICP—AES法测定土壤中的痕量砷》
KM—48炸弹
世界F-1赛车锦标赛
世界一级方程式(简称F-1)赛车锦标赛
本田V-12发动机
新型MP4/6万宝路麦拿仑本田赛车
F-1赛坛
F-1赛车赛
PVD光化学气相沉积设备
中国—新加坡国际受方付费直拨电话(IODC)业务开通仪式
"IODC"电话
"IODC"业务
"S"形
MS2401中外文打字机
ISO国际品质标准规范
A组
B组
"NO!NO!NO!"
美国DEC计算机公司
DEC公司
CT复查
"麻雀"空空导弹(AIM-7E)
F-4,F14,F-15型飞机
DNA
优质香料烟PK—873
PK—873
PK—873烟草
X,Y病毒

CNA 1991
F—型戰鬥機
澳亞航空(AUSTRALIA ASIA AIRLINE)
卡拉OK
美國經濟評議會(CONFERENCEBOARD)
KGB(國家保安委員會)工作人員
KGB工作人員
KTV
老人SOS電話服務中心
SSB國際緊急救難頻道
ER型客機
波音七四七—SP型客機
麥道MD—十一型客機
G—七層會議
萬國商業機器公司(IBM)
IBM公司
互補型金屬氧化半導體(CMOS)
BMW
I型轎車
BMW轎車
朋馳(BENZ)汽車
第七四五C次貨物列車
Q三—四九次柴油客車
S七0C型直昇機
UH—一H型戰鬥
T—三四型基礎教練機
B—一九〇〇型中型運輸機
S—二T型反潛機
C—一三〇型運輸機
AT—三高級教練機
F五E—F戰術戰鬥機
F—〇4戰術戰鬥機
IDF全天候高性能戰鬥機
U—六型定翼機
V—五〇S裝步營
CM二一裝步營
M四十八H勇虎型戰車
M—〇九自走砲兵營
M—一〇自走砲兵營
M—九飛彈
M—十一飛彈
ABC日報
NHK電視台
美洲國家稅務主管中心會議(CIAT)技術年會
CIAT
T五五戰車
日空航空(JAS)
貨幣供給額M2年成長率
THOMAS M. LIN
二〇五年
蘇聯幼岡斯克油氣生產公社(簡稱Y.P.A.)
西西伯利亞的奈伏幼岡斯克區(NEFTEYGANSK)
H型鐵軋鋼廠
百美B
TGV
ICE
「二十個感謝」(ThankTwenty)
DASH—八、三百型、五十二人座渦輪螺旋槳新客機
M—11
CF十八戰鬥機
C—三〇運輸機
「幻象F—E」戰機助陣
「F十六」戰機
關稅暨貿易總協定(GATT)
GATT
MIT (台灣製造的產品)
JOHN GARAMENDI
濫用治腎臟病藥物 EPO
EPO
美國貿易代表署 (USTR)
MD—11 客機
MD—12X 客機
麥克唐納道格拉斯 (MCDONALD—DOUGALAS)
MD—十一客機
MD—十二X 客機
CX 五一零次班機
MILD7 香菸
MILD7
三K 黨
ABC
CBS
ABC 總部
DC—九型客機
亞太經濟合作會議 (APEC)
秘魯生海鮮的吃法叫「CEBICHI」
C 型飛彈
輕型飛機 AN—2
KTV
(南京東路五段八號B 一)
WRC 電視台
編號 00 三四八四 T 二
船名 CAN THAN 的越南漁船
加勒比海共同體 (CARIBBEAN COMMUNITY AND COMMON MARKET)
蒙塞萊特 (MONTSERRAT)
「每月一書俱樂部」 「INHOUSE‘S READER」
F—七噴射戰鬥機
鮑凱利尼 A 大調第六號大提琴奏鳴曲
莫札特 A 大調小提琴奏鳴曲
「I LOVE YOU (我愛你)
U 字形
洛杉磯州大音樂系系主任史維恩 (JOHN SWAIN)
洛杉磯州大交響樂團的指揮巴克 (DAVID BUCK)
XIN 1998
F-15战斗机
F-16战斗机
GDP
D区
Dissent
Dissident
持不同政见者(DifferentPoliticalViewer)
DPV
DNA(脱氧核糖核酸)
DNA检测装置
蚊子DNA
AC米兰队
AB型血
200CC
B超
"Z"形步
A组
B组
空中客车A320型飞机
MVP
"BP"机
维生素C
甲B综合
全国足球甲B联赛
甲B
甲A
CVD
ISO9001合格证书
CT
卡拉OK厅
VCD视盘机
CI676班机
A—300大型客机
A—300型民航班机
F-15I远程战斗机
F-15I战斗机
L-188"厄勒克特拉"飞机
NBA
瑞典SKF公司
IC卡
DVD机
法兰克福股市DAX30种股票
巴黎CAC40种股票
E组
8mm摄录一体机
ISO14001环境管理体系的认证
丹麦AVK
APEC科技产业合作议程
美国"SCI"科学论文索引资料库
"BALL(球)"
国际商业机器公司(IBM)
"IBM电子商业会议博览会"
荷兰TVM车队
年末现金流通量(M0)
狭义货币(M1)
广义货币(M2)
A股
B股
H股
"N异丙烯酰胺"物质
"闪电-M"火箭

CNA 1998
PLAYBOY
「十四K」
中美WTO入會協議
世界貿易組織(WTO)
HIV 愛滋病毒
SLOT1架構
OCKET7
PII世代
440BX晶片組
440EX晶片組
PII350MHZ
100MHZ外頻
PII333MHZ
60MHZ
440LX晶片組
CELERON266MHZ
128KB
MENDIOCINO微處理器
 SOCKET7系列產品
SLOT1微處理器
AMD集團
HP
IBM
DRAM(一六M,六四M)
邏輯IC(繪圖晶片)代工經驗
DRAM市場
DRAM
日本東京TAKANAWA王子飯店
亞太一A衛星
亞太IIR型衛星
C-頻段轉發器
KU-頻段轉發器
HBO
ESPN
無本金交割遠期外匯交易 (NDF)
外國專業投資機構 (QFII)
三月貨幣供給額M1B年增率
貨幣供給額M1B變動率
印吉利克 (INCIRLIK)
F十五
F十六
F———
石油輸出國家組織(OPEC)減產協議
OPEC
巨幅釋迦牟尼佛像大唐卡 (Thang--ka)
H型鋼
KTT證券金融公司
印尼GARAM牌大麻香菸
毒品MDMA (俗稱快樂丸)
FORMOSA@MAIL.MOJ.GOV.TW
電子郵件(E-MAIL)
專屬網站網頁(HOMEPAGE)
台灣地區航空公司代表協會(BOAR)
「ASIAN FORUM日本會議」
VISION(展望)
BOT甄審辦法草案明定案
BOT案
台灣向美國傾銷DRAM案
XSspeed超高速網際網路服務
全美第一大動態隨機存取記憶體(DRAM)製造商
DRAM價格
英國前外長侯艾 (Lord Geoffrey Howe)
佳士得公司(CHRISTIE’S)
莫內(MONET)
梵谷(VAN GOGH)
高更(GAUGUIN)
雷諾爾(RENOIR)
克爾貝(COURBET)
戴伽斯(DEGAS)
莫里梭(MORISOT)
DNA比對
DNA基因比對
腸病毒Echo型
「駐台北匈牙利貿易辦事處」(Hungarian Trade Office)
匈牙利首任處長邬以敬(Gyorgy Ujlaky)
「AB檔案」
ISO9000認證
「外國代理人」(foreign agent)
IC產品
電腦IC記憶體
XIN 2005
A组
B组
X射线
DNA鉴定
I型糖尿病
UT斯达康
GDP
IT提升改造传统产业论坛会
3G产业
第三代移动通信(3G)产业
3G发展
3G政策
3G市场
TD-SCDMA
WCDMA
CDMA2000
TD-SCDMA产业化
3G技术网络
世界超级跑车锦标赛(FIA GT)
F1标准
GT赛事
BPR GT耐力赛
马萨拉蒂MC12赛车
N-GT组别
GT1组别
JMB
GT2组别
金融街(CBD)
4X200米自由泳接力
中国居民消费价格指数(CPI)
CPI
H5N1型禽流感病毒
高等经济商业学院(ESSEC)
ESSEC
P-51型战斗机
C-46型运输机
L-5型通讯联络机
PT-19型初级教练机
P-51
CBA
4X100米混合泳接力
F组
V类
劣V类
IT业者
IT界
IT界人士
IT界厂商
IPV6是互联网
IPV4网址空间资源
IPV6产业化
IPV6
IPV6产品
IPV6全球论坛
IT设备制造业
IBM个人电脑业务
NBA
NBA头号新人
NBA克里夫兰骑士队
THANK YOU
WTA(女子网球协会)年终总决赛
IT
LCD(液晶显示器)
ATP巡回赛
HDW公司

CNA 2005
俄羅斯路克石油公司(Lukoil Overseas Kumkol B.V.)
美國聯邦準備理事會(Fed)
Fed
多明尼加總統費南德茲(Leonel Fernandez Reyna)
副總統艾布奎克(Rafael Albuquerque)
參議院議長包蒂斯塔(Andrez Bautista)
眾議院議長巴奇可(Alfredo Pacheco)
外交部長莫拉萊斯(Carlos MoralesTroncoso)
外销暨投資推廣部部長馬蒂內茲(EddyMartinez)
農業部部長羅梅洛(Amilcar Romero)
「杜華德,桑傑士,梅亞大十字銀質勳章(Orden alMerito Duarte, Sanchez y Mella en Grado de GranCruz Placa de Plata)
深灣(Bahia Honda)
耐吉 (NIKE)
銳跑 (Reebok)
阿迪達斯 (Adidas)
印度SAHARA集團
聯合國糧食計畫署(WFP)
嚴重急性呼吸道症候群(SARS) 疫情
WFP
WFP後勤補給主管
負責中東,中亞,東歐區域主管阿布杜拉(Amir Abdulla)
MI-26 直昇機
MI-26
MI-8 運輸直昇機
MI-8
胸部X光篩檢
KTV
連鎖 KTV
正子斷層造影檢查(PET/CT)
美國前駐中國大使尚慕杰(James Sasser)
「自由基(free agent)」
台灣和蒙古事務主任韋德寧(Dennis Wilder)
美國國務院主管全球事務的國務次卿杜布林斯基(Paula J.Dobriansky)
美中關係全國委員會會長歐倫斯(Stephen A.Orlins)
喬治華盛頓大學艾略特國際事務學院院長職務的何漢理(Harry Harding)
法國世界報(Le Monde)
里昂德力(E.Leandri)
台灣芳珂FancI公司
棉籽糖 (raffinos se)
凱蒂 (Katie)
夏馨姊姊茱麗 (Julie Williams)
OPEC
貝爾斯登(Bear Stearns Cos.)
摩根士丹利 (Morgan Stanley)
全球貨幣研究部門主管Stephen Jen
歐洲央行 (ECB)決策官員
ECB理事Axel Weber
ECB
七大工業國 (G7)財金首長華府峰會
經濟專家David Malpass
油國組織 (OPEC)各成員國
摩根士丹利駐倫敦貨幣研究主管Jen
OPEC國家
LED
IC設計
英國廣播公司 (BBC)
渥太華卡爾頓大學的史密斯(ErinConway-Smith)
歐(Ceridwyn Au)
多倫多城市新聞(Metro News)
卡瀚(Saleem Khan)
魁北克省布其頓(Birchton)自由報
布里(Charles Bury)
多倫多國家郵報的里昂(Melissa Leong)
加拿大記者協會會長史奇尼德瑞特(PaulSchneidereit)
Tamiflu
H5N1禽流感
可有效對抗禽流感的藥物Relenza
飛利浦 (LPL)
TFT-LCD面板產業
吉娜戴維斯(Geena Davis)
「三軍統帥」(Commander-in-Chief)
唐納蘇德蘭(Donald Sutherland)
佛萊迪普瑞茲(Freddie Prinze)
零分念做「Love」
而蕃茄的英文字,除了眾人熟知的Tomato之外,
也有Love Apple的美名
ABC三級
B級
C級
A
C
C:\windows\regedit.exe

202
c:\winnt\regedit.exe
\HKEY_LOCAL_MACHINE\SOFTWARE\Microsoft\Windows\CurrentVersion\Run
mssmart
msupdate
smartCard
C:\windows\system32\smartup.exe
smarth.exe
C:\Program Files\Tommy\Smart Card\SmartSet.exe
ISO-9001品質管理系統驗證
ISO-14001環境管理系統驗證
BIBLIOGRAPHY


Chan, Mimi and Helen Kwok. 1982. A Study of Lexical Borrowing from English in Hong Kong Chinese. Hong Kong: University of Hong Kong Centre of Asian Studies.


Cheung, Samuel [张洪年]. 1972. 香港粤语语法的研究 [Cantonese as Spoken in Hong Kong]. Hong Kong: The Chinese University of Hong Kong.


[中国文字改革委员会研究推广处编, 1959, 《普通话三千常用词表》, 初稿, 北京:文字改革出版社.]