Why are Chinese characters so damn hard?

An investigation into character confusion

Michał Kosek

Thesis submitted for the degree of
Master in Linguistics
60 credits

Department of Linguistics and Scandinavian Studies
Faculty of Humanities

UNIVERSITY OF OSLO

Autumn 2016

Why are Chinese characters so damn hard?

An investigation into character confusion

Michał Kosek

Why are Chinese characters so damn hard?

http://www.duo.uio.no/

Printed: Reprosentralen, University of Oslo

Abstract

This thesis looks into the problem of learning Chinese characters for foreign language learners and focuses on learning approaches that stress recognising characters without writing them by hand, which are becoming popular due to the widespread use of computer-based input methods. Mistaking characters for each other has been identified as an important problem that learners need to overcome. The empirical investigations in the thesis include a self-observation diary study and connectionist simulations of learning Chinese characters.

The diary study collected over 1,500 pairs of characters that had been mistaken for one another in the process of learning. The analysis of these cases revealed an interplay of various factors that led to character confusion: graphical, semantic and phonetic similarity, as well as association caused by frequent co-occurrence of given characters in some words. A more detailed analysis distinguished character components that have a semantic or phonetic value in modern Chinese. It showed how the presence of similar components may contribute to character confusion, and found more complex cases of relationships between the value of the components of the target character and the actual pronunciation and meaning of the character it was confused with.

The connectionist simulation of character acquisition presented in this thesis is based on the DISLEX model, which consists of two self-organising maps and aims to provide a neurobiologically plausible account of word learning. An evaluation of the first version of the model showed that the pairs of confused characters collected in the diary study were represented significantly closer to each other than the average. Nevertheless, the model had major flaws, which were addressed in the second version. It included a more sophisticated representation of the semantic, phonetic and graphemic features of the characters. The second model showed a significant improvement over the first one.

The model accounted for character confusion by representing the approximate pronunciation of the characters, the approximate pronunciation indicated by their phonetic components, frequently recurring graphical components and the semantic classification of the characters (as indicated by the hypernyms). These results give an indication of what a psychologically plausible representation of Chinese characters may look like. Experiments with more learners are required to assess the scope of applicability of these findings and the predictive value of the model.

Contents

Chapter 1

Introduction

1.1 Goal

Why Chinese is so damn hard (為什麼中文這麼TM難？) is the title of a famous essay by David Moser (1991). When such a question is asked by a scholar of Chinese who has already spent several years studying the language full-time, it deserves some attention. The first thing usually associated with the difficulty of Chinese is the characters, and indeed, the majority of reasons brought up by Moser are related to the writing system. This thesis investigates the problems related to learning Chinese characters.

Characters are basic components of the Chinese writing system; learning the characters is a time-consuming task both for first- and second language learners of Chinese. Lack of adequate reading skills is a common problem even among those second language learners who have been learning Chinese for several years, and it is often caused by the fact that they cannot recognise enough characters. This thesis looks at issues related to recognising characters and focuses on one of the difficulties that L2 learners usually face: the problem of confusing one character for another.

The goal of this thesis is to provide data that can increase our understanding of the process of learning Chinese characters and find ways to identify which characters are likely to be confused by second-language learners. A typical Chinese character represents a morpheme and is therefore associated with a graphical form, sound and meaning. We will investigate the degree to which this confusion may be attributed to similarities between different characters with regard to these three aspects.

1.2 Structure of the thesis

The remainder of this chapter provides the motivation for this thesis by discussing the notion of Chinese as a difficult language and the role of characters in its difficulty. Chapter 2 presents important facts about the Chinese writing system, its evolution and the structure of the Chinese characters. One of the most difficult aspects of learning Chinese is learning Chinese characters. We will therefore investigate the number of characters required to achieve adequate comprehension of texts aimed at native

speakers. Chapter 3 summarises models of reading and the bilingual mental lexicon. Most importantly, it discusses the advantages and disadvantages of different models of reading, particularly connectionist models. This is important background information for the study presented in subsequent chapters.

Chapter 4 begins with a summary of a pilot study of character recognition that motivates further research presented in this thesis. Next, it discusses some of the character teaching approaches with a particular focus on character-centred, recognition-based methods that make use of semantic and phonetic components of the characters. It argues that an important problem that learners may face is confusing characters for one another. This phenomenon may be particularly likely to occur when learning to recognise characters without learning to write them by hand. Such recognition-based learning methods are increasingly popular, due to the widespread use of technology that enables character input without handwriting. The last section of the chapter poses research questions that may be answered using data about characters that have been confused by the learner. The study presented in later chapters addresses these questions.

Chapter 5 begins with a formal definition of what it means for one character to be confused with another, and argues for a self-observation-based diary study as a method of data gathering. It then discusses the patterns of confusion that have been discovered in the gathered data, and argues that one approach to forming a hypothesis about which characters are likely to be confused is to construct a computational model of character learning. The specific decisions made in constructing the model are presented in chapter 6. The initial representation did not perform well, and therefore the model had to be improved. The results of running the model are then analysed. In particular, the analysis of how the model was improved may give indications as to what kind of representation is more psychologically plausible. Chapter 7 concludes the thesis, summarising its main points and suggesting future research directions.

1.3 Motivation

1.3.1 Difficulty of mastering Mandarin for Western learners

It is hard to compare the difficulty of learning different languages, even for learners with identical L1. There are many factors that need to be taken into account, including motivation, learning methods and aims and goals of the learners. Let us look at expected learning times for different languages provided by the Foreign Service Institute, an American institution that offers language training to diplomats[1].

Even though FSI's method of estimation is not explicitly stated, the target proficiency level of language teaching is well-defined: professional working proficiency, or S3/R3 in the ILR scale (which roughly corresponds

[1] http://web.archive.org/web/20071014005901/http://www.nvtc.gov/lotw/months/november/learningExpectations.html

to the C1 level in the Common European Framework of Reference for Languages). Moreover, it is likely that the estimates are based on learners who are highly motivated, have similar goals and are taught using similar methods[2]. This group of learners is more uniform that general population of foreign language learners. Therefore, relative differences between the estimates for different languages should tell us something about the relative difficulty of these languages for native English speakers.

According to FSI, obtaining professional working proficiency in Mandarin Chinese in reading and writing by an English speaker requires about 2200 hours of training. For comparison, the time to attain the same level in most Germanic and Romance languages is estimated at 600 hours. More interestingly, the time to learn Indonesian or Swahili, languages just as unrelated to English as Chinese, is about 900 hours. Also languages that are typologically or genetically similar to Mandarin are much easier. Vietnamese, like Mandarin, is a very analytic language and has been under cultural influence of Chinese for a long time. Still its expected learning time is two times shorter than that of Mandarin (though Vietnamese is said to be a somewhat more difficult than other languages in the 1100-hour group). The same estimate applies to Burmese, which together with Mandarin belongs to the Sino-Tibetan language family. As we can see, these differences cannot be explained by cultural and linguistic distance between languages.

1.3.2 Role of Chinese characters in the difficulty of Mandarin

The group of languages that require 2200 hours of instruction at FSI includes Mandarin, Cantonese, Japanese and Korean, that is, it all the languages from the FSI list that use or recently used Chinese characters to a significant degree. Arabic is the only language in the most difficult group that has no relation to the Chinese writing system whatsoever.

About 2000 Chinese characters are taught in education systems in both North and South Korea, but their usage is currently infrequent in the Korean language, and therefore they are not essential to learn for Korean L2 speakers. As of 2016, the use of Chinese characters in modern Korean is minimal. However, the description of Korean on the FSI website claims that "[t]he use of Hanja [Chinese characters] is still common in South Korea", which suggests that the learning time estimates come from the 1990s or earlier, when Chinese characters were an important part of the Korean as a second language curriculum.

Japanese, the other language that uses Chinese characters in its writing system to a large degree, has an official list of so-called Joyo kanji – 2136 characters that one needs to know in order to use the written language fluently. They are used in the context of teaching Japanese writing for both first- and second-language speakers. Japanese mass media generally

[2]Quote from the FSI website: "It must be kept in mind that that students at FSI are almost 40 years old, are native speakers of English, and have a good aptitude for formal language study, plus knowledge of several other foreign languages. They study in small classes of no more than 6. Their schedule calls for 25 hours of class per week with 3-4 hours per day of directed self-study."

annotate non-Joyo characters with pronunciation, which clearly indicates that there is a significant number of native speakers who cannot read such characters. It is hard to compare the difficulty of learning the Chinese and the Japanese writing systems, as they function in very different ways. It is clear, however, that one of the aspects of the difficulty of Chinese is the sheer number of characters that need to be learnt. There is no official Chinese list corresponding to Joyo kanji, but the number of such important characters is definitely higher in this language. The problem of estimating the number of characters required for text comprehension will be discussed in section 2.5.

Even though applying the FSI estimates to the whole population of language learners is not straightforward, it is clear that Chinese will pose a much greater challenge to typical learners from Western countries than most other languages, such as Indonesian or Vietnamese. Moser (1991), in the above-mentioned essay, lists 9 reasons for why Chinese is hard to learn for English speakers. 5 of them are, broadly speaking, related to the writing system, 3 reasons are related to the lexicon, and one has to do with the fact that Chinese is a tonal language. There is no reason to think that each of Moser's points is equally important or that the list is exhaustive. We can note that there are many tonal languages (including the above-mentioned Vietnamese and Burmese) and vocabulary learning problems are frequent, especially when one learns a language from a different language family. On the other hand, the challenges caused by the writing system are quite unique to languages that use Chinese characters.

Chinese characters outnumber graphical symbols in every other writing system that is in use in the modern world. This comes from the fact that they are, to a large degree, logographic: while other writing systems associate individual symbols with particular sounds or their sequences, logographic systems also associate them with meaning. Phoneme inventory and phonotactics naturally restrict the number of recurring sound patterns in every language, but meaning is not subject to such a limitation. If we combine that with the fact that mapping between the graphical form of the character, its pronunciation and its meaning is far from straightforward, we can see why Chinese characters are a likely reason for why Chinese takes more time to learn than most other languages.

1.3.3 Importance of Chinese characters

Chinese characters are likely to be the main obstacle to learning Mandarin. But is it necessary to learn them? It is possible to achieve a conversational level in the language without too much contact with the characters. Some heritage speakers may even have near-native spoken fluency without ability to read. There are, however, many aspects of Chinese that an advanced language user needs to know, which are very hard or perhaps even impossible to learn without learning the characters. There are many thousands of *chengyu* 成語 (chéng yǔ), four-letter idioms that come from classical Chinese phrases, and their structure makes sense only when analysed character by character. Many words in the formal register are predominantly used in writing, and only rarely in speaking, so it is hard to learn them without reading. A

lot of language content in Chinese, from commercials to jokes, depends on character homophony or polysemy. The characters in proper names are often chosen in a way that conveys some additional meaning. This all makes learning Chinese characters inevitable for anyone who wants to achieve an upper-intermediate or advanced command of the language.

Chapter 2

Chinese writing system

This thesis deals with the issue of learning Chinese characters in the context of learning Standard Chinese (which is a form of Mandarin) as a second language. This chapter provides basic information about Chinese characters and their relation to modern spoken Mandarin. However, in order to get a better understanding of the structure and function of the Chinese characters, some perspective is needed. To this end, a brief look at other Sinitic languages that use or used this script will be beneficial, in particular Old Chinese, which was spoken at the time when the Chinese writing system emerged.

2.1 Relation of the Chinese writing system to the Chinese languages

Even though the focus of this thesis is the characters, some information about the phonology of the underlying language needs to be provided. As mentioned above, there is no single underlying language: the Chinese writing system was and is used to write several different languages. Let us first look at modern Mandarin; other languages will be shortly mentioned in the next section.

Each Chinese character, apart from very few exceptions, represents a syllable. The structure of a syllable is very limited in Mandarin: the largest possible form is CGVX, where C is a consonant, G is a glide, V is a vowel, X is a coda. Moreover, a syllable needs to have one of four tones (numbered from 1 to 4), or be in the fifth, neutral tone. In *Pinyin* 拼音 (pīnyīn), the standard phonetic transcription of Mandarin, tones are indicated by a diacritic mark over the vowel and its absence indicates the neutral tone, e.g. *ā* (1st tone, tone contour[1]: 55), *á* (2nd tone, tone contour: 35), *ǎ* (3rd tone, tone contour: 21 or 214), *à* (4th tone, tone contour: 51), *a* (neutral tone). The phonological system has been analysed in various ways, with different numbers of consonants, glides and vowels. All the analyses

[1]The tone contour notation uses numbers from 1 to 5, where 1 signifies the lowest pitch level and 5 signifies the highest pitch level. In contrast, the Mandarin tone numbers are arbitrary.

generally agree, however, that there are only about 400 possible syllables, if we disregard the tone. If we take the tone into account, we will find about 1200 syllables that are actually used in the language (i.e. syllables that are actual pronunciations of at least one character).

The Chinese writing system is logographic to a large degree: most characters do not only represent a string of phonemes, but also a particular meaning; in other words, they are morphemes. That is, for a pair of homophonic morphemes we can usually expect them to be written with two different characters. Due to the restrictive phonotactics, this situation is extremely frequent in Mandarin: For example, among the 6000 most frequent characters, over 40 are pronounced *yì* and over 30 are pronounced *xī* (with the same phonemes and tone). This situation does not usually produce problems with communication, because most of these morphemes are bound and restricted to specific words. Even though many of the most frequent words are monosyllabic, the vast majority of words in Chinese is bisyllabic. Longer words are infrequent (Da 2005), apart from a set of *chengyu* 成語 (chéng yǔ), four-character idioms taken directly from Classical Chinese.

The words in written Chinese are not separated by spaces. There is no graphical marking of word boundaries at all, and morphology, syntax, semantics and pragmatics all contribute to correct interpretation of segmentation of sentences into words. Such interpretation may sometimes be ambiguous.

While the vast majority of characters are morphemes, it is not the case for all of them. Some characters are associated only with a particular, multi-character word, and therefore we cannot say they have a meaning by themselves. For example, the individual characters in the word 蝴蝶 (hú dié) 'butterfly' are not morphemes, because they do not appear in any other word, nor alone. In other cases characters may represent morphemes, but not in a particular context. This is often the case with loanwords. For example, the characters in the word 挪威 (nuó wēi) 'Norway' do have meanings: 挪 (nuó) 'move, shift' and 威 (wēi) 'impressive strength, might, power'. In the context of the name of the country, however, they have been chosen for their sound, not meaning, and therefore 挪威 (nuó wēi) needs to be treated as one indivisible unit. On the other hand, even here the meanings of the characters are not completely random – there is a strong tendency to transcribe country names and other such proper nouns with characters that have a positive meaning.

2.2 Structure of Chinese characters

This section presents a very basic analysis of the building blocks of the characters, in order to present a general idea of what they consist of. The following sections will provide a deeper discussion about possible ways of systematising the characters and understanding their structure.

On the most basic level, each character consists of strokes needed to write it with a brush or pen, that is, individual lines and dots. On a higher level, we can find out that most of the characters consist of recurring components

in different configurations. We can group them into simple characters, that contain only one component, e.g. 馬 (mǎ) 'horse', and complex characters, that contain more components, e.g. 媽 (mā) 'mother'. This distinction is unrelated to the distinction between simplified and traditional characters discussed in the next section.

Simple characters, such as 馬 (mǎ) 'horse' or 木 (mù) 'wood, tree', have only one component and make up the oldest character group. In modern Chinese writing, we can think of most simple characters as arbitrary symbols representing morphemes with particular meanings and pronunciations. Complex characters, conversely, consist of several components. Some of character components may be free, that is, have the ability to function alone as independent characters. Other components are bound, and can occur only as a part of a character.

There are different ways of systematising character components, but it is clear that some of the components play a semantic role, and some indicate pronunciation. For example, 媽 (mā) 'mother' has two components: 女 (nǚ) 'woman' and 馬 (mǎ) 'horse', and the former is a semantic component (indicates that meaning of the whole character is somehow related to women), while the latter is a phonetic component. It indicates that the pronunciation is similar to *mǎ*. In some cases the pronunciation is identical, but often it is not – here 媽 (mā) and 馬 (mǎ) have different tones.

The classification of components is character-specific – the same component can be semantic within one character, and phonetic in another. For example, the character 沐 (mù) 'wash' can be decomposed into the semantic component 氵 'water', and the phonetic component 木 (mù) 'wood, tree'. On the other hand, in the character 杏 (xìng) 'apricot', 木 (mù) 'wood, tree' is clearly a semantic component.

There are components that have no other role; they cannot stand alone, and therefore have no associated pronunciation, e.g. 氵 'water' or 艹 'plant' (although in the case of 氵, we may treat is as an orthographic variant of the character 水 (shuǐ) 'water'). Other components, such as 木 (mù) 'wood, tree', can stand alone as simple characters. Complex characters can serve as components, too. For example, the character 睬 (cǎi) 'pay attention, take notice' can be decomposed into the semantic component 目 (mù) 'eye', and the phonetic component 采 (cǎi) 'pick'. This phonetic component is itself a character that can stand alone, which, according to Harbaugh (1998), consists of two semantic components: 爪 (zhǎo) 'claw' and 木 (mù) 'wood, tree', which both relate semantically to the action of picking. Such a combination of semantic components is less transparent than in the case of 氵'water' as a semantic component of 沐 (mù) 'wash'. Sometimes the same component may play both a semantic and a phonetic role. For example, the character 娶 (qǔ) 'take a wife, marry a woman' can be decomposed into the semantic component 女 (nǚ) 'woman', and the component 取 (qǔ) 'to take, to fetch'. The latter component has both a

semantic and a phonetic role in this case.

The character 采 (cǎi) 'pick', when used as a component of 睬 (cǎi) 'pay attention, take notice' is a "black box". That is, it is there to indicate pronunciation, and its internal structure is not relevant in the analysis of 睬 (cǎi); the only thing that matters in this context is the fact that it is pronounced *cǎi*. In other words, even though 木 (mù) is a component of 采 (cǎi), there is no reason to regard it as a component of 睬 (cǎi), unless we are doing a purely graphical analysis. Another example of this very frequent situation is 张 (zhāng) 'stretch, extend' being used as a phonetic component of 涨 (zhǎng) 'rise, go up'; the structure of the former is not relevant in the analysis of the latter.

When a standalone character is used as phonetic component, it often indicates its own pronunciation. This is, however, not a fixed rule. There are many phonetic components that do reliably indicate pronunciation, but it is different from their own. Consider characters 都 (dōu /dū) 'all/capital', 赌 (dǔ) 'gamble', 堵 (dǔ) 'block'. The component 者 clearly plays a phonetic role here and indicates a pronunciation similar to *du*, even though the character 者 (zhě) is pronounced *zhě*.

The classification of components as semantic or phonetic can be done in different ways, depending on one's aims. We are concerned with learning characters, therefore some deconstructions will be more important than others for us. Hundreds of characters that have meaning related to water contain the 氵'water' component and we can certainly consider it pedagogically important. The combination of 爪 (zhǎo) 'claw' and 木 (mù) 'wood, tree' may serve as a mnemonic aid to remember 采 (cǎi) 'pick', but on the other hand, for some learners this character is simple enough, and can be remembered as one entity, and decomposition creates an unnecessary burden. It should also be noted that some pedagogical approaches use incorrect etymologies as an aid for remembering complex characters (a well-known example is Heisig & Richardson, 2015). This approach can be useful if the actual etymologies are not interesting enough to make a vivid association.

2.3 Development of Chinese characters and writing styles

There are several non-Sinitic languages that use Chinese characters: they make up an important part of the Japanese writing system, they are still sometimes used to write Korean and some minority languages in China, such as Zhuang, and they were formerly used to write Vietnamese. These languages have had, however, relatively little impact on the current Standard Chinese writing system.

Among the Sinitic languages, currently only Mandarin and Cantonese have standardised writing systems that use Chinese characters. In practice, other Chinese dialects are often written with the characters, too. However, they contain many morphemes that are not associated with any particular

Figure 2.1: Examples of the character 鳥 'bird' in the oracle bone script (left), the bronze script (middle) and the seal script (right). The last character is the standardised small seal found in Shuowen Jiezi. Source: `http://www.chineseetymology.org/`

character. Different strategies are used in such cases. Depending on the situation, they may involve using arbitrary characters with a given pronunciation, creating new characters, or using Latin letters to write these morphemes.

The language that had the largest impact on the modern Chinese writing system is Old Chinese, which was spoken in third century BC when the script was standardised in the newly unified China under the First Emperor Qin Shi Huang 秦始皇 (qín shǐ huáng). Understanding of the structure and function of the Chinese characters and the way they work requires recognition of the fact that they make up a system that was designed to write Old Chinese. After the standardisation, it remained largely unchanged and was used to write subsequent languages that eventually evolved into modern Mandarin and other Chinese dialects. Let us first take a short look at the history of the characters.

Chinese characters were written in many different ways, and several writing styles remain in use. The most important forms of Chinese characters in the ancient writing period are illustrated in Figure 2.1 and include (in chronological order): oracle-bone inscriptions 甲骨文 (jiǎ gǔ wén), bronze inscriptions 金文 (jīn wén) and seal script 篆書 (zhuàn shū), further (somewhat fuzzily) divided into large seal script 大篆 (dà zhuàn) and small seal script 小篆 (xiǎozhuàn). The small seal script was the first standardised version of Chinese characters: it was imposed as the only writing standard in newly unified China after the Warring States Period by the First Emperor.

In the same period, an important simplification of the characters has been popularised. This simplified way of writing characters, the clerical script 隸書 (lì shū), can be traced back to the Warring States Period (Qiu 2000). However, as the unification of China involved writing many documents, this informal way of writing characters became popular among lower-level state officials. This change had enormous importance and long-lasting effects. The clerical script is the earliest script that is quite easy to understand by anyone who can read modern Chinese characters. Yin (2006, p. 3) lists four features of the simplification, also referred to as the *clerical change* 隸變 (lì biàn): "1) the curved strokes in the seal script became somewhat straighter, 2) the overall number of strokes was reduced, 3) some different components were merged into one, and 4) some components were modified and simplified". The subsequent dynasty, Han, saw the development of the clerical script

that resulted in development of several scripts: the cursive script 草書 (cǎo shū), the semi-cursive script 行書 (xíng shū) and the regular script 楷書 (kǎi shū). The cursive script was meant for fast writing, but is not generally understandable nowadays.

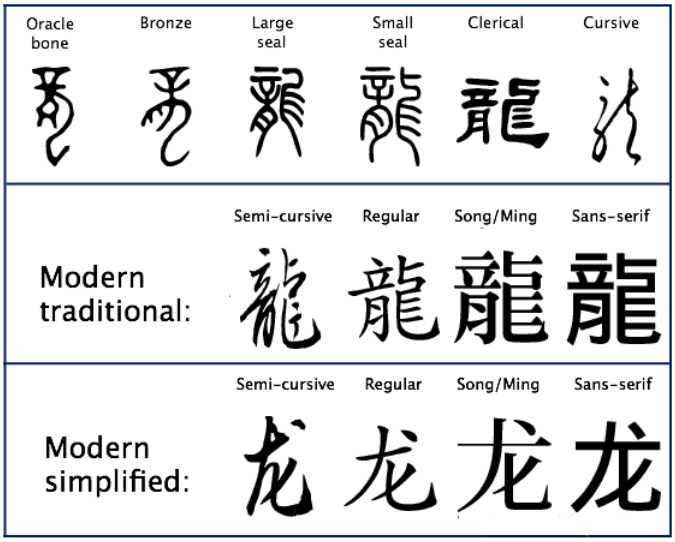

Figure 2.2: The character 龍 'dragon' in several old and modern writing styles. Based on `https://zh.wikipedia.org/wiki/漢字#/media/File:Hanzi2.png`

The regular script has had the strongest influence on Chinese writing and remains one of the most popular typefaces used in modern printing. Other frequently used typefaces, sans-serif 黑體 (hēi tǐ) and Song/Ming 宋體 (sòng tǐ) /明體 (míng tǐ) typefaces are basically variants of the regular script and all are readable by adult native Chinese. There are, however, some differences between them and it is possible for a learner to recognise a character in the typeface that they are most exposed to, but not in the others. Figure 2.2 provides an overview of the writing styles discussed in this section.

Over the years, people wrote some characters in various ways. It led into the development of *variant characters* 異體字 (yì tǐ zì), that is, characters with different shapes, but the same meaning and pronunciation. Different variants may have different geographical distribution. For example, the character 裏 (lǐ) 'inside' came to be written as 裡 (lǐ) and now the former variant is predominantly used in Hong Kong, and the latter – in Taiwan. In some cases, several variants may be used interchangeably. For instance, the character 臺 (tái) has a variant 台 (tái), and in Taiwan, the name *Taiwan* can be written as either 臺灣 (tái wān) or 台湾 (tái wān), the former being slightly more formal. Note that in some other contexts 台 (tái) may also be an independent character, not a variant character: Taishan, a city in Guangdong province, PRC is written as 台山 (tái shān) and never as 臺山 (tái shān). That is, words written with 臺 (tái) may also be written with 台 (tái) (which is, in such cases, just an alternative way of writing 臺 (tái)), but words

that originally were written with 台 (tái) cannot be written with 臺 (tái) (because, in such cases, 台 (tái) is a distinct character, and 臺 (tái) is not a variant of 台 (tái)).

The traditional/simplified distinction stems from another major simplification of the characters, which took place in People's Republic of China (PRC) in the 1950s. There are many characters with different variants, and the simplification scheme chose an official variant for each of these. Moreover, many characters were simplified in a way that resembles some aspects of the *clerical change*, namely points 2, 3 and 4. The simplification is not regular, that is, simplification of many characters is ad-hoc, and not guided by any universal set of traditional-to-simplified conversion rules. The document that provides details of all the simplifications is the *Chinese Character Simplification Scheme* 漢字簡化方案 (hàn zì jiǎn huà fāng àn). The characters that have not undergone simplification, and have the same structure since the introduction of the clerical script, e.g. 光 (guāng), are called *inherited characters* 傳承字 (chuánchéng zì). The ones that have undergone simplification are called *simplified characters* 簡體字 (jiǎn tǐ zì). The original forms of the simplified characters are called *traditional characters* 繁體字 (fán tǐ zì). For example, 马 (mǎ) is a simplified version of the traditional form 馬 (mǎ) 'horse'. The terms *simplified characters* and *traditional characters* are extended to mean whole characters sets, with inherited characters belonging to both sets.

Apart from PRC, the simplified characters are also used in Singapore and Malaysia, while the traditional ones are prevalent in the Taiwan, Hong Kong and Macau. The following is a non-exhaustive list of different relations between traditional and simplified characters.

- **One-to-one**: There is only one traditional and one simplified character, and they have identical pronunciations and meanings.

- **Many-variants-to-one**: In the cases of characters with several variants, there is usually a many-to-one mapping between traditional and simplified characters. Sometimes one of the variants is chosen. As we have seen above, texts in traditional characters may use the character 臺 (tái) or its variant 台 (tái), and some names contain the non-variant character 台 (tái). In the simplified character texts they are always written as 台 (tái). In some other cases, the simplified character is not a variant of the traditional ones. We have seen that 裡 (lǐ) is a variant of 裏 (lǐ). However, they both are simplified as 里 (lǐ), which is not a variant of either of them.

- **Many-characters-to-one**: There is also a significant number of cases several different traditional characters, which different meanings and possibly different pronunciation, are merged into one simplified equivalent. Sometimes the simplified equivalent is one of the traditional characters. The traditional character 里 (lǐ) means 'li, a unit of length', but it was merged with 裏 (lǐ) 'inside' and its variant 裡 (lǐ), and in

simplified character texts, all these three characters are always written 里 (lǐ). In other cases, the simplified character may have different shape than any of the traditional equivalents. For example, the traditional characters 發 (fā) 'to send' and 髮 (fà) 'hair' both have the same simplified equivalent: 发, which is pronounced *fā* when it means 'to send' and *fà* when it means 'hair'.

- **One-character-to-many**: Finally, there are some opposite cases, where more than one simplified character has the same traditional equivalent. It may happen because the simplification scheme is applied to some characters only when they represent a particular pronunciation and meaning. For example, 徵 is generally pronounced *zhēng* and simplified as 征. However, when 徵 means 'fourth note in the traditional Chinese pentatonic scale', it is pronounced *zhǐ*. In this case, 徵 (zhǐ) has no simplified equivalent. The same applies to the character 乾, which is simplified as 干 when it is pronounced *gān* and means 'dry', and is unchanged in the simplified script when it is pronounced *qián* and is used to create words such as 乾坤 (qiánkūn) 'cosmos'.

  Another reason for one-to-many equivalence between traditional and simplified characters is related to regional differences in the use of the former. For example, in Taiwan, the traditional character 著 has many different meanings and pronunciations: *zhe* 'ongoing action marker', *zháo* 'catch', *zhuó* 'send', *zhāo* 'chess move; trick'. In other places in China, including both Hong Kong and the mainland, the variant character 着 is and was used instead, even before the simplification. However, 著 may also mean *zhù* 'write; show; marked', and in this case, the form 著 is used in simplified writing, too. The next section provides another example of one-to-many equivalence caused by regional differences, further complicated by diachronic changes.

2.4 Classification and organisation of Chinese characters

Even though this thesis is concerned with the characters used in modern Chinese, we cannot assume a completely synchronic perspective and look at the characters without any consideration of their etymology. Some of the etymological information continues to influence modern readers' understanding of the structure of the characters. For example, there are two characters: 月 (yuè) 'moon' and 肉 (ròu) 'meat'. However, 肉 is not used as a component, and when serving as a component, 月 may mean either 'moon' or 'meat'. This is a result of an orthographic change that took place during the formation of the clerical script. Therefore, the component 月 has two different functions in modern characters. Moreover, most learners of Chinese script get explicit knowledge about this distinction at some point. A completely synchronic explanation of the two functions would have to mirror

the diachronic explanation. However, it can be argued that a diachronic explanation is sufficient in such situations.

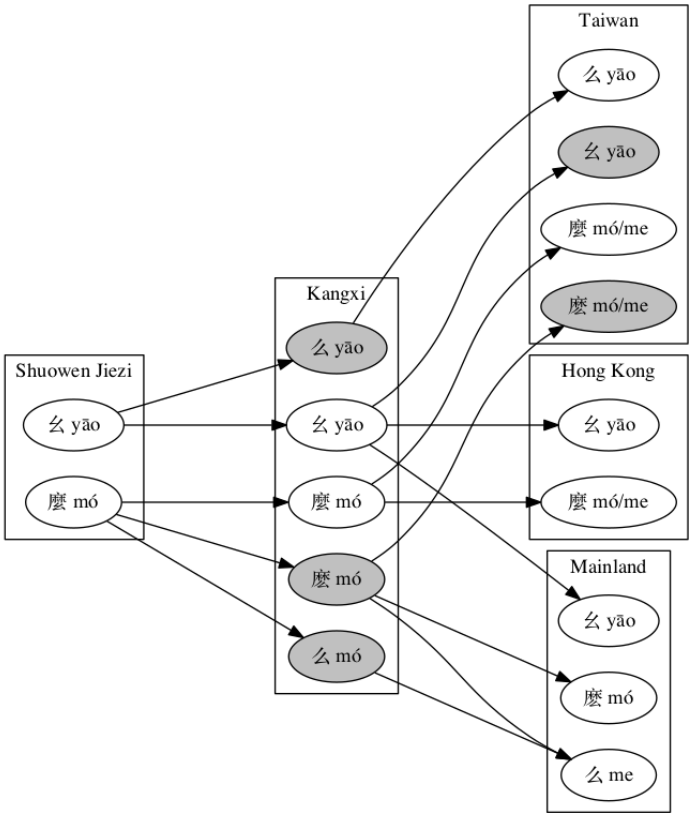

Figure 2.3: The evolution of 幺麼. Grey nodes represent variant characters. Based on historical character information from Shuowen Jiezi and the Kangxi dictionary (both reprinted in Sturgeon, 2011). Modern data about Hong Kong are based on the Longman Advanced Chinese Dictionary (2003), data about Taiwan are based on MOEDict (2015) and data about Mainland China are taken from the Table of General Standard Chinese Characters (State Council 2013)

Consider differences between the traditional and simplified script. It is hard to provide a synchronic explanation of the correspondence between the words 怎麼 ‘how’ and 么麼 ‘petty’ (as written in the traditional script in Taiwan), and their simplified equivalents 怎么 and 么麼 . The character 么 is in no way a simplification of 么 , in fact, the character 么 has not been changed in the simplification process. Moreover, 么 is actually a simplification of 麼 and not 麼[2]. The relation between these characters becomes understandable only after we learn that the earliest forms were 么

[2]Cf. *Table of General Standard Chinese Characters* (2013), an official character simplification document

and 麼 (mó), and different variants have been standardised in different places. Figure 2.3 explains their evolution in more detail. Note that even here we give a simplified picture of the situation: we only consider prescriptive data from the dictionaries and ignore several other variants of these characters. Still, this level of detail is enough to show the origins of the complicated relation between 麼, 麼, 幺 and 么 in different scripts nowadays.

2.4.1 Organisation of Chinese characters in dictionaries

Looking at how characters were organised at different stages in history may provide some understanding of the conceptual structure that modern Chinese speakers have, which is what second language learners need to acquire. Chinese had no concept of word before the beginning of Western influence in the 19th century. That means that characters were seen as the basic units of the written language.

The organisation of dictionaries at different periods, described by Wang Li (2010), may indicate how the perception of the system of Chinese characters evolved. We should note that the early dictionaries had a rather limited purpose: explaining characters in the classical works from older periods, that were not understandable anymore. Erya 爾雅 (ěr yǎ), the oldest surviving dictionary from 3rd century BC, was organised semantically into 19 thematic groups, such as 釋器 (shì qì) 'explaining utensils' or 釋鳥 (shì niǎo) 'explaining birds', with each group containing characters with a related meaning.

A groundbreaking change, both in the structure, as well as in contents, is found in Shuowen Jiezi 說文解字 (shuōwén jiě zì). It is a basis of much of later research on character structure, and some scholars even today continue to use it as the main source of information. This dictionary was created between 100 and 121 CE by Xu Shen 許慎 (xǔ shèn), a scholar of the so-called Old Text school that supported study of classical works in their original version in the seal script. He believed that systematic study of characters will allow greater understanding of the classics. Therefore, it is a dictionary that explains the structure of seal script characters. The explanations, however, are written in the standard script of the period, the clerical script. The definitions were provided primarily as an aid in understanding how the characters formed, since a supposed original meaning of the character is provided, not necessarily the one that was found in the classics (Bottéro & Harbsmeier 2008). The most important thing about this dictionary is its unprecedented focus on character structure. It decomposed characters and interpreted their components: they were marked with 从 (cóng), which indicated a semantic component, or as 聲 (shēng), which indicated a phonetic component. It was possible for one component to be marked both as 从 (cóng) and 聲 (shēng). The whole dictionary was organised according to selected semantic components. It is something very different from the organisation of Erya: while Erya was organised according to actual meaning, Shuowen was organised according to one of the graphical components of the characters, which Xu Shen regarded as having a semantic value. Moreover, Shuowen popularised *liu shu* 六書 (liù shū),

a theory of six principles of character formation, that remained uncontested until the 20th century.

Shuowen Jiezi was divided into sections, one for each of the selected semantic components. These components are therefore called 部首 (literally 'section headings'; in English they are referred to as *radicals* or *bushou*). The system of radicals has been the organising principle of the vast majority of later Chinese dictionaries up to the present day, but the radical list has been modified with time. Shuowen had 540 *bushous*, later dictionaries often removed the ones that were rarely used. The set of 214 radicals used in modern Chinese dictionaries was introduced in the Zihui 字彙 dictionary (published in 1615), but they were greatly popularised by the Kangxi 康熙 dictionary (published in 1716) and are commonly known as the Kangxi radicals.

This origin of the radicals often leads to the misunderstanding that they are the same thing as the semantic components. However, this was not true even in Xu Shen's time: there are a lot of entries with more than one component marked as semantic, but only one of them is chosen as the radical. Moreover, in Shuowen there is at least one case of a phonetic component being used as a radical: the character 鳧 'wild duck' is listed under the radical 几 , which is presented as having only a phonetic role, and not under the semantic component 鳥 'bird'. This example is exceptional, and might be considered a mistake. Nevertheless, further changes distorted the system even more. For example, 舅 'maternal uncle' had the radical 男 'male' in Shuowen. However, in a later period 男 was removed from the radical list, and in the current system 舅 has the radical 臼 'mortar', which clearly plays a phonetic, and not a semantic role. Later additions also include radicals that have no meaning at all and are strokes rather than components, e.g. 丿 . Apart from that, we currently have access to earlier stages of Chinese writing than Xu Shen had, and can find cases where the etymology in Shuowen is wrong and the radicals are not actually semantic.

As we can see, radicals are arbitrarily chosen character components and their only purpose is organising written dictionaries. Therefore, we cannot equate them with semantic components. We can only speak of general tendencies. For example, most radicals play some semantic role in the character (although that role is often not clear), and they often tend to be placed on the top or on the left-hand side of other components. Different components exhibit different tendencies: e.g. the grass radical 艹 is placed on the top (as in 花 'flower'), but the heart radical 忄/心 is placed on the left side or on the bottom (as in 悅 'pleased' or 想 'think, miss'). Sproat (2000) wrote a set of rules for proper placing components, and found that it works with 88% accuracy on 2588 frequent characters, which means that the remaining 12% had to be specified manually as exceptions. So in general, there is no unambiguous way of finding out which element of an unknown character is a radical and for this reason they are not even suited well

to their primary role: facilitating dictionary look-up. It is therefore not surprising that their use is diminishing as more people start using electronic dictionaries, which allow unknown characters to be written directly.

The third major group of dictionaries organised characters according to their pronunciation. It was the method employed in rhyming dictionaries, such as Qieyun 切韻 (qiè yùn). It was created in 601 CE and recorded the language which is now called Early Middle Chinese[3] and, according to Pulleyblank (1991), is the earliest stage of spoken Chinese that can be systematically reconstructed. This dictionary shows us how the phonology of Chinese was perceived by the Chinese themselves before the contact with Western linguistics in the late 19th century. The pronunciation was indicated in terms of other characters, using the so-called *fanqie* 反切 (fǎn qiè) method. An entry in this dictionary consists of four characters: *headword onset rhyme* 反, and has the following interpretation: the first character should be pronounced with the onset of the second character and the rhyme of the third character. The character 反 marks the end of an entry. For example, "東德紅反" indicates that the headword 東 should be pronounced with the onset of 德 [tək] and the rhyme of 紅 [ɣuŋ], that is, as [tuŋ][4]. In Chinese linguistics, syllables have never been analysed in terms of phonemes; the onset and rhyme (called *initial* and *final*, respectively) were the lowest level of phonological analysis.

We can conclude that there were three general ways of organising characters in dictionaries: according to their meaning, according to their graphical form and according to their pronunciation. The characters were grouped by their meanings in the earliest surviving dictionary, Erya. Grouping according to pronunciation was introduced latest, and was used in rhyme dictionaries. Shuowen Jiezi, which was primarily an etymological dictionary, introduced grouping by arbitrary graphical parts of characters, radicals. Despite the flaws of this system, it is the one that has been used most widely.

The three aspects that were used for organising characters (graphical form, meaning and pronunciation) are all important from the perspective of the learner. In the next subsections we will look deeper at the graphical form of Chinese characters and its complicated relationship with both pronunciation and meaning.

2.4.2 Six categories of Chinese characters (*liu shu*)

Even though *liu shu*, the theory popularised by Xu Shen, until recently remained uncontested as a theory of dividing of Chinese characters into six categories, there was no agreement as to how individual characters should be classified. Even though the principles of *liu shu* were described in Shuowen Jiezi, the dictionary itself did not classify each headword. The definitions of the categories were ambiguous and left much room for disagreement by later scholars. There are the following categories in *liu shu*: 指事 (zhǐ shì) (simple ideographs), 象形 (xiàng xíng) (pictographs), 會議 (huì yì) (compound ideographs), 形

[3]There is no agreement on what variety of Early Middle Chinese is recorded in Qieyun.

[4]Reconstructed pronunciations after Pulleyblank (1991)

shēng
聲 (semantic-phonetic compounds), zhuǎn zhù
轉注 (*zhuanzhu*, discussed below) and jiǎ jiè
假借 (phonetic loans).

Simple ideographs and pictographs are what we called *simple characters* earlier in this chapter, that is, the ones without any subcomponents. The name zhǐ shì
指事 means 'indicate things' and is generally used to describe characters that show abstract entities such as shàng
上 'above' and xià
下 'below'. The name xiàng xíng
象形 means 'resemble form' and is used to describe characters that are direct depictions of an entity, such as mù
木 'wood, tree'. Qiu (2000) points out that the distinction between the two is often blurry and there is no agreement between scholars about exact boundaries of these two categories.

Semantic-phonetic compounds make up the largest category. Getting sound information from semantic-phonetic compounds is not straightforward. Chen Zhiqun (2009) notes that in only about 20% of 7000 most popular semantic-phonetic compounds have exactly the same pronunciation as their phonetic indicator. However, according to Zhao (2005), characters were only loaned to represent words that had identical pronunciation at that time. On the other hand, Qiu (2000) allows that a frequent graph with a similar pronunciation might have been preferred to a rare graph with exactly the same pronunciation. He also mentions two other reasons for the mismatch between pronunciation of characters and their phonetic indicators: identically pronounced words may diverge over time due to diachronic changes and identical words in one dialect may have two different pronunciations in another one. Therefore, when a character is used as a component in semantically unrelated words, we can expect the reason to lie in phonology – at some point in time, in some dialect, they were pronounced at least similarly, if not identically.

Some phonetic loans are used only in their loaned meaning. For example, the character 不 was originally a pictograph of a calyx of a flower. However, it was borrowed as a word for 'not' (which presumably had identical pronunciation at that time), and after some time is stopped being used as a character for 'calyx'. It is, however, not necessary for the original usage to cease. The character huā
花 'flower' has huà
化 as the phonetic component, and the 艹 'grass' radical as the semantic component. It is, however, also used in modern Mandarin as the word 'to spend', as in huā qián
花錢 'spend money'. As Qiu (2000) points out, even though it looks the same as the character huā
花 'flower', we cannot analyse it in terms of semantic and phonetic components, we should rather say that huā
花 'to spend' as a whole is a phonetic loan from huā
花 'flower', borrowed only for its sound.

Zhuanzhu is the most unclear category of *liu shu*. Qiu (2000) writes: "Of all the names assigned to the six principles of writing [liu shu], the surface meaning of the term *zhuanzhu* is the murkiest. The description of the *zhuanzhu* given in the *Shuowen*'s postface is also insufficiently clear". Qiu goes on to list 9 interpretations of this category that have been proposed throughout history, which vary widely. According to some, only a handful characters would belong to this category, according to others, the vast

majority of characters would be categorised as *zhuanzhu*. He concludes that Chinese characters can be described without referring to this category at all.

Compound ideographs have an internal structure, but all the components are used to indicate meaning. Several rules of formation of compound ideographs have been proposed. For example, the meaning of a character may be formed from the common attribute of referents of its components. According to Chen Zhiqun (2009), this is the traditional interpretation of the character 明 míng 'bright', composed of 日 rì 'sun' and 月 yuè 'moon', which share brightness as a common attribute. There is also another subtype, compound ideographs that contain components that can be read as if it was a phrase. A typical example of such a character is 歪 wāi 'crooked'. Clearly, it is a combination of 不 bù 'not' and 正 zhèng 'straight'. Note that the modern pronunciation of the character has nothing to do with the pronunciation of its components.

However, Chen Zhiqun argues that most characters formed according to these rules are relatively late creations, created or reinterpreted to fit the already existing *liu shu* theory. She argues that a large part of traditionally defined compound ideographs are actually *complex pictographs*, which will be explained in more detail in the next subsection.

2.4.3 Three categories (*san shu*) and three stages of development of Chinese characters

In the 20th century scholars became more open about the deficiencies of *liu shu*. Tang Lan (1979, quoted by Qiu 2000, p. 163) wrote: "What do the six principles tell us? First, there were never any clear-cut definitions; each person could come up with his own interpretations. Second, when the six principles were used to classify characters, it usually was impossible to determine which category each character should be placed in. In the light of these two points alone, we should neither place all our faith in the six principles nor fail to seek other explanations." There were several attempts to provide a better classification system. At least three scholars (Tang Lan 1979, Chen Mengjia 1988, Qiu Xigui 2000) created each their own *san shu* 三書 sān shū , systems of three categories. Here we will look at the latest *san shu* system, proposed by Qiu (2000), which divides characters into *semantographs* 表意字 biǎo yì zì , *phonograms* (sic) 形聲字 xíngshēng zì and *loangraphs* 假借子 jiǎ jiè zi .

The categories of loangraphs and phonograms are more or less equivalents of phonetic loans and semantic-phonetic compounds from the *liu shu* theory, while the rest is generally classified as semantographs. To see why these three categories are much more natural we need to look at the process of formation of Chinese characters, which was better understood after the excavation of oracle bone inscriptions in the 20th century. Again, at least three scholars suggested that Chinese writing was created in three distinct stages (Chen Mengjia 1988, Boltz 1994, Chen Zhiqun 2009). Their descriptions of these stages are, however, quite different. We shall look at

Chen Zhiqun's proposal, who argues that it is an improvement over the other two. Chen Zhiqun's three stages are: the pictographic stage, the multivalent stage and the determinative stage.

In the pictographic stage, the character depicted the referent directly. It was not limited to concrete nouns, such as 足[5] 'foot', 人 'person', 女 'woman', 子 'child' or 木 'tree'. Words describing actions or states got pictographs that showed a prototypical situation illustrating a given action or state. For instance, the character for 出 'go out' showed a foot stepping out of a pit, the character 从 'to follow' showed two people, one after another, the character 'to give birth' showed 子 'child' under 女 'woman' and the character 上 'above' showed a short line above a longer one. Even though they would traditionally be put into three different classes: pictographs ('foot', 'person', 'woman', 'child', 'tree'), compound ideographs ('go out', 'follow', 'give birth') and simple ideographs ('above'), Chen Zhiqun (2009, p. 262) points out that they all can be called pictographs, because "they were invented the same way: as simple depictions of the best exemplar [of an object, state or action], with a distinctive feature highlighted". Chen Zhiqun also points out that the depiction was schematic, and some of the features were highlighted either because they were highly relevant to the described object, state or action, or in order to differentiate the character from other characters with similar shape.

Even though we may formally divide a character such as 从 'follow' into two components (人 'person' + 人 'person'), semantically it is indivisible: it is a depiction of a person following another, and therefore of the verb 从 'to follow'. In other ways, its meaning is not compositional: it does not come from the meaning 人 'person' combined with another meaning 人 'person'.

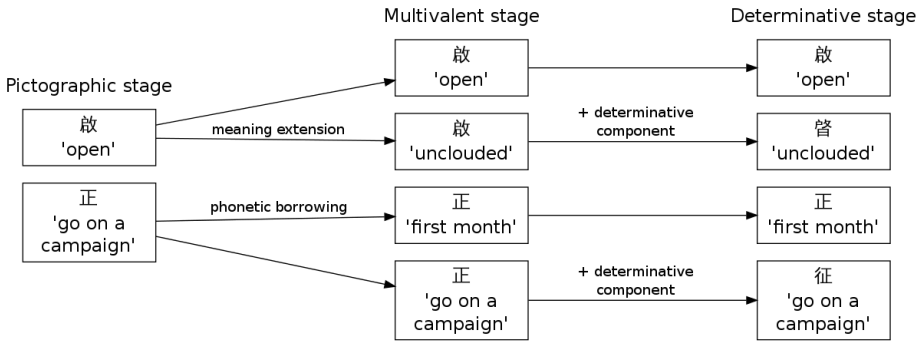

Figure 2.4: Examples illustrating the three stages of development of Chinese characters, according to Chen Zhiqun (2009). Stages occurred independently for every character, so the forms of characters grouped in each of the stages did not necessarily occur simultaneously.

[5]For the sake of simplicity, the characters in this sections are presented in their modern form, even though we talk about their development in the past. This should not be a problem, since we are concerned with the arrangement of character components, not their exact shape. The characters in this section are not glossed with their modern pronunciation, since it is irrelevant from the historical perspective.

The development of characters through the three stages presented in Figure 2.4 uses two examples provided by Chen Zhiqun (2009, pp. 278, 289). In the pictographic stage, the character 啟 represented its meaning, 'open': it showed a door 户 opened by a hand 又. The character 正 depicted a foot 止 marching towards a destination and meant 'to go on a campaign'.

In the second, multivalent stage, two different processes took place: meaning extension and phonetic borrowing. The character 啟 'open' can serve as an example of meaning extension: according to Chen Zhiqun (2009, p. 278) it started to be used to signify 'unclouded' as an extension of 'opening (the cloud to reveal the Sun)'. Phonetic borrowing took place in the case of the character 正, which originally depicted a foot 止 marching towards a destination and meant 'to go on a campaign'. It was borrowed to write the word for 'first month', which had the same pronunciation; thus, it became a loangraph. Meaning extension and phonetic borrowing are superficially similar, as they both caused appearance of characters with multiple meanings. The difference between the two is, however, important: the former led to the creation of polysemes, while the latter led to the creation of homographs.

In the third, determinative stage, writers started marking differences between some of the polysemes and homographs created in the multivalent stage. For example, the 日 'sun' component was added to the character 啟, to mark the meaning 'unclouded', as opposed to 'open'. The 彳'road' component was added to 正 to form 征 and mark the original meaning 'go on a campaign', as opposed to 'first month' (which became the only meaning of the original 正). This led to creation of compound ideographs: the new character 暩 'unclouded' got two components (啟 'open' and 日 'sun'), and its meaning became compositional: it signifies that it is one of the extended meanings of 'open', the one that is related to (opening the clouds to reveal) the sun. Similarly, the new 征 'go on a campaign' has two components: 彳'road' and the original sign 正 'go on a campaign'. Phonograms were also created in the third stage: it happened when a semantic component was added to a loangraph, that is, to a character whose form was not related to its meaning and which was borrowed only for its sound value.

Chen Zhiqun also notes that some characters underwent only two stages, without the multivalent stage: even if no meaning extension nor phonetic borrowing took place, a character was sometimes given an additional determinative component, by way of analogy to other characters in its semantic field. For example, the characters 屰 'go to meet' and 𠦂 'encounter, meet' were given the determinative 辶 'go', and formed 逆 and 遘, which have the same meanings. Unlike in the previously discussed cases, the additional component was not needed to differentiate meaning, but it was added under the principle of analogy to other semantically related characters with this component. The original versions of the characters (屰 and 𠦂) simply ceased to be used. It should be noted, however, that 𠦂 is has continued to serve as a phonetic component in other characters.

Most characters discussed in this section belong to Qiu Xigui's category of semantographs. According to Chen Zhiqun, a large part of this category consists of pictographs: *simple pictographs* such as 人 'person',

or *complex pictographs* such as 从 'follow', which are semantically non-compositional, even though they do consist of two or more graphical components. Pictographs were written as prototypical examples of an item or situation in question. They had their features highlighted, especially those that made it possible to distinguish it from similar characters. Other semantographs can be called *compound ideographs*, because they have been constructed out of existing characters, and their meaning is a function of meaning of the components.

Going back to the *san shu* classification introduced at the beginning of this subsection, we can now link each of its three categories with different character histories. Loangraphs are characters that were borrowed for their sound in the multivalent stage and remained unchanged in the determinative stage. Phonograms are characters that were borrowed for their sound in the multivalent stage and got a determinative component in the determinative stage. Finally, semantographs are characters that either weren't changed in the multivalent stage or underwent meaning extension. Semantographs that did not change in the determinative stage are pictographs (simple or complex, depending on the number of graphical components in the original character). Semantographs that did get a new component in the determinative stage are compound ideographs.

2.4.4 Later construction and reinterpretation

The aim of the three-stage theory is to account for the development of the earliest Chinese characters from oracle-bone inscriptions to their present form. The categorisation based on this theory can be used to classify the majority of modern characters. There are characters that were directly created as semantic-phonetic compounds, by way of analogy to existing phonograms, but they do not pose categorisation problems: we can interpret them as being already in the determinative stage. However, there are characters created in later periods that consist of semantic components and were created or reinterpreted according to different principles than the ones of the three-stage theory.

Components that were used for their meaning, rather than depicting something directly, became more common around the Warring States period (Outlier Dictionary 2016), which overlaps with the periods of the bronze inscriptions and the seal script. This led to an invention: some of the characters created in the later periods contained only meaning components. The above-mentioned 歪 (wāi) 'crooked' is one such later creation that does not conform to the three-stage theory. Its two parts, 不 (bù) 'not' and 正 (zhèng) 'straight' (which is another, currently the most common meaning of 正), are meaning components.

The process of reinterpretation can be illustrated by the story behind the character 明 (míng) 'bright', as presented by Chen Zhiqun (2009). The character 明 was originally a pictograph that meant 'early morning' and depicted something that was typical for the morning: the sun 日 and the moon 月 together on the sky. The meaning 'bright' was represented by the

pictograph 朙, depicting the moon 月 shining through a window 囧. As the result of script reforms, 明 started to be used as a variant of 朙. In the Tang dynasty period (618–907) 明 began to be used as the standard form, and the likely reason is that 明 conformed best to the *liu shu* theory – 'bright' was understood as the shared property of the sun and the moon. That is, while originally it was a semantically non-compositional pictograph depicting 'early morning', 明 was reinterpreted as a compound that can be semantically decomposed into 日 'sun' and 月 'moon'.

2.4.5 Decomposition of modern Chinese characters

This short overview of the history of Chinese characters presented above shows that dividing components into those with semantic function and those with phonetic function is not sufficient. Indeed, there is an important difference between the components of 从 (cóng) 'follow', and the components of 歪 (wāi) 'crooked', even though both of them are often classified as semantographs. 从 (cóng) 'follow' consists of two components 人 (rén) 'person'. However, they are not used there for their meaning, but for their form: each of them is a pictograph of a person, and only an *image* of two persons next to each other leads us to the meaning 'follow'. The case of 歪 (wāi) 'crooked' is clearly different. As we saw above, 不 (bù) is actually a pictograph of a calyx of a flower and 正 (zhèng) is a pictograph of a foot with a line above. Unlike in 从 (cóng), the pictographs of the components are not relevant for understanding 歪 (wāi). Only *meanings* of the components are clearly relevant: 'not' and 'straight'. As we can see, the two cases are different. The upcoming Outlier Dictionary of Chinese Characters[6] is probably the first dictionary of Chinese characters that makes this distinction.

For the purpose of this thesis we are interested in a synchronic analysis that can be psychologically plausible. However, it is unlikely there is a single analysis that would describe how all learners conceptualise components when they recognise any particular character. It stems from the fact that the structure of most characters is not as clear as in the cases of 歪 (wāi) and 从 (cóng). Let us look at 征 (zhēng) 'go on a campaign' and 正 (zhèng) 'straight' as examples.

A historically accurate analysis (based on Chen Zhiqun's theory) would see 征 (zhēng) as a compound ideograph with 正 (zhèng) used for its form (as it depicts a foot 止 marching towards a destination, represented by 一) and with 彳 used for its meaning 'road', and reinforcing the meaning of the whole compound. On the other hand, the modern character 正 (zhèng) is a loangraph, and as such, does not have any internal structure – it was borrowed in the past only for its sound value. However, we cannot ignore the fact that pronunciations of 正 (zhèng) and 征 (zhēng) have always been related. Given that 彳 is related to the character's meaning, and 正 is related to its pronunciation, it looks like a typical phonogram. This is how both Xu Shen and Qiu Xigui (2000) analyse it; the latter explicitly states that the fact that most people will view such

[6] http://www.outlier-linguistics.com/

yíng	zhǎn	pán
盈	盞	盤
‘full of’	‘small cup’	‘dish, tray’
shèng	pén	hé
盛	盆	盒
‘contain’	‘basin, pot’	‘box, case’
jiān	jiàn	jiàn
監	艦	鑑
‘supervise’	‘warship’	‘scrutinise’
làn	lán	lán
濫	藍	籃
‘overflow’	‘blue’	‘basket’

Table 2.1: Some of the characters containing the component 皿

characters as phonograms is more important than the actual etymology. Xu Shen, on the other hand, probably wanted to be historically accurate, but was hindered by insufficient data about character evolution available at the time of writing Shuowen Jiezi. This is probably also the reason for why 止 is indicated as the only semantic component in Xu Shen's description of the character 正 (zhèng).

Historical accuracy is even less important for authors of various character teaching materials. Such materials contain information that is valuable for research, as they are likely to influence how learners conceptualise the characters. Ann (1982), too, presents 征 (zhēng) as a phonogram. However, his presentation of 正 (zhèng) is different: it is not described as a loangraph, but rather as a compound, made up of 一 (yī) ‘one’ and 止 (zhǐ) ‘stop’. Its meaning ‘straight, right’ is said to derive from ‘stop at one, unite at one’, which can be considered a folk etymology. Even though both Ann and Chen decompose 正 into 止 and 一, the former claims that 正 combines meanings of these elements (‘stop at one’), while the former claims that 正 is a combination of their forms (depicting a foot marching towards a destination).

Some learning materials decompose characters in a way that cannot even be called folk etymology, as it is clear that its only purpose is serving as a memorisation aid. For example, Heisig & Richardson (2015) associate 彳, 一 and 止 with ‘queue’, ‘one’ and ‘footprint’, while Matthews & Matthews (2007) associate them with ‘step forward’, ‘unicorn’ and ‘stop’, respectively. They both go on to present stories that combine these elements in a way that can serve as a mnemonic for real meaning and pronunciation of the characters.

We can see that it is impossible to predict and model different ways a learner may analyse a character and conceptualise its components. We can note, however, that memorisation aids are used only to commit characters into long-term memory, and after several repetition of the character, its mnemonic can be safely forgotten. Therefore, their influence on how characters are processed in the long run is probably not big. On the other hand, there are patterns that, even if they are not taught explicitly, get entrenched every time a character is processed. For example, consider Table 2.1, which lists some frequent characters containing the component 皿 (compiled from Matthews 2004 and Ann 1982). We can see that meanings of the characters in the first two rows are semantically related to concepts such as 'contain' and 'container'. The characters in the last two rows, which have 監 (jiān) as their component, exhibit phonetic similarity: they are all either pronounced *jian* or *lan* (with various tones). In characters such as 盞 (zhǎn) the component 皿 'container' is used for its meaning, while in characters such as 艦 (jiàn) the component 監 (jiān) is used for its sound. Note that the words 監 (jiān) 'supervise' and 鑑 (jiàn) 'scrutinise' are related: the character 鑑 (jiàn) was given a semantic component in the determinative stage to mark a specific meaning of the polysemous word 監 (jiān) . However, just like in the case of 征 (zhēng) 'go on a campaign', which was discussed above, the clear phonetic pattern is likely to cause reinterpretation of this character as a phonogram.

There is no reason to assume that components can only have a single role. In the case of 監 (jiān) and 鑑 (jiàn) we have only two characters with the same component and meaning, so the possible pattern is not very entrenched in the learner's mind. But in the case of 籃 (lán) 'basket' (jiān) we have two clear patterns. The character shares the component 監 with similarly pronounced characters, 濫 (làn) and 藍 (lán) . At the same time, it shares the component 皿 with other characters semantically related to containers, such as 盤 (pán) 'dish, tray' and 盒 (hé) 'box, case'. Despite that 皿 is graphically a part of 監, both patterns are relevant and should be taken into account.

Semantic and phonetic patterns, such as the ones described above, are quite frequent: Guder-Manitius (1998) identified 122 components that are shared by semantically related characters and 683 components that give some indication about how the character is pronounced. Therefore, it is likely that learners will take advantage of such patterns, consciously or unconsciously, regardless of whether they were explicitly taught to notice them.

But how about the character 監 (jiān) itself? Etymologically, 皿 in 監 depicts a container with water, while the top of the character depicts a person using the container as a mirror to inspect his or her face (Outlier Chinese Dictionary 2016). The component 皿 (mǐn) is not used for either meaning or sound here, it is used for its form: it is an iconic depiction of a container with water. Alone, it has nothing to do with the meaning of 監 (jiān) ; it only functions as a part of the scene with a person inspecting their face, and only the scene taken

as a whole is related to the meaning of the character. As we can see, the use of components for their form is idiosyncratic, and while the etymology may help to memorise the character, for our purposes it is indistinguishable from other, historically inaccurate mnemonics. Moreover, perhaps apart from few simple cases, such as 从 (cóng) 'follow', the use of components for their form is unlikely to be noticed by the learner, unless taught explicitly.

We can conclude that use of components for their form is pedagogically relevant, as it can both help to remember the character and provide information about the development of Chinese script, which gives a better understanding of how it works as a system. It is, however, hard to include processing of form components in a general model of reading acquisition of modern Chinese, as the way they are interpreted may vary a lot between learners. On the other hand, the sound components and the meaning components exhibit patterns that are likely to be relevant for any learner.

2.5 Number of characters required for text comprehension

A better understanding of the problem of learning written Chinese requires us to find out the number of characters that a successful learner needs to acquire. As mentioned in the introduction, Chinese does not have a single official character list, such as Joyo kanji for Japanese, that would provide a clear answer. Therefore, we will try to make an estimate based on several such lists and a corpus character frequency analysis.

2.5.1 Official character lists and requirements

Let us first find out the number of characters that Chinese natives from different parts of the Chinese society are expected to know. 6500 of the characters in the *List of Standard Common Characters* published by the authorities of the People's Republic of China is said to "satisfy needs related to publishing news, printing and editing"[7]. That is an exhaustive list of characters one is expected to find in a modern Chinese text, apart from some infrequent proper names. However, only highly educated native Chinese speakers are expected to know so many characters. Character requirements for Chinese pupils in primary and secondary education set a much lower limit. A corpus analysis by Xing, Shu & Li (2004) revealed 3306 different characters in primary school textbooks (grade 1 to 6). According to the 2011 standard for the curriculum of language and literature classes in China[8], students should recognise about 3000 characters at the end of the primary school (6th grade), and by the last grade of secondary education (9th grade) the students are required to recognise about 3500 characters. We can regard this as a high estimate for the required number of characters for L2 learners

[7] http://www.gov.cn/zwgk/2013-08/19/content_2469793.htm

[8] 義務教育語文課程標準 (Standard for the Curriculum of Language and Literature Classes in Compulsory Education), 2011 edition, Beijing Normal University Publishing Group, http://mat1.gtimg.com/edu/pdf/edu/xkb2011/20120130155433177.pdf

New HSK level	Words (cumulative)	Characters (cumulative)
1	150	174
2	300	347
3	600	617
4	1200	1064
5	2500	1685
6	5000	2663

Table 2.2

Old HSK level	Words (cumulative)	Characters (cumulative)
Basic	1033	800
Elementary	3052	1603
Intermediate	5257	2194
Advanced	8840	2865

Table 2.3

of Chinese – we can certainly expect people who finished secondary school to be fully literate. The lower bound is the official definition of literacy in PRC, defined as knowledge of 1500 characters for peasants and 2000 characters for city dwellers. As we shall see, this should be regarded as an absolute minimum, very unlikely to be sufficient. We need to note that first language speakers who learn to read very often already know the words they are reading, they just do not know their written forms. This is very different from the situation of second language learners, who have a much smaller vocabulary and often do not know the words they are trying to read and therefore have much less room for context-based guessing.

There are also official word list for second language learners of Chinese – words that people taking official Chinese language exams are expected to know at different levels. They are called the HSK lists, after the names of the exams (漢語水平考試 *Hànyǔ Shuǐpíng Kǎoshì* 'Chinese Proficiency Test'). The exams have been thoroughly changed in 2010, and there are two sets of such lists – before and after the reform. It is important to consider both lists, since the older ones contain significantly more words and characters, which reflects the fact that the highest level of the new HSK is significantly lower than the highest level of the old HSK. Tables 2.2 and 2.3 show word and character statistics for different levels.

We can see that the old HSK lists for levels up to "intermediate" contain 5257 words that consist of 2194 individual characters, the lists for levels up to "advanced" contain 8840 words that consist of 2865 characters. The lists for the highest of the new levels, level 6, contain 5000 words that consist of 2663 characters.

Word-text coverage	No. characters (Chinese Internet Corpus)	No. characters (Sinica Corpus)
80%	1588	1864
95%	3663	3458
98%	4433	

Table 2.4

2.5.2 Language corpora

Another way to estimate the number of characters that one needs to learn is to look at language corpora. We can look at them in light of Hu & Nation's (2000, cited in Koda 2005) investigation of the amount of vocabulary that is required for text comprehension. They compared the relation of vocabulary coverage (the percentage of known words in the text) to the comprehension of a text among English as L2 learners. "At 95[%] coverage [...], some participants comprehended the text, but most did not. At the 80% level [...] none of the sample apprehended the text meaning. [...] the researchers speculated that adequate comprehension requires roughly 98% text-word coverage" (Koda 2005, p. 58). There are reasons to believe that in the case of reading Chinese as L2 the required text-word coverage is even higher. Study performed by Hayden (2005) shows that reading Chinese causes a cognitive load that even for learners at the advanced level is significantly higher than for native readers.

We can expect that if we consider character coverage instead of word coverage, the coverage requirements for successful comprehension will be just as high, and possibly even higher. As mentioned above, Chinese words very often consist of more than one character. Even if one knows all characters in a word, one may fail to understand the word. On the other hand, one sometimes may not recognise some characters, and still be able to understand the word. We have seen examples of characters that appear in only one word, such as in 蝴蝶 (hú dié) 'butterfly'. It is therefore clear that recognising only one of these characters is enough to be sure which word is written. Such words are, however, a small minority, and should be treated as exceptions. In the great majority of cases, each character is a morpheme. It is also theoretically possible to learn to recognise shapes of whole words, without learning to recognise individual characters, but this would mean ignoring all the information that comes from decomposition, so it is unlikely to be a good method. These issues will be discussed further in connection with the character-based approach to learning, now we will conclude that recognising individual characters is usually an important prerequisite to recognising words. Under this assumption, we can look at language corpora and extract the most frequent words that make up 98% of all the texts and count the number of different characters that appear in these words. This will produce an estimate of the number of characters that one needs to know in order to read with adequate comprehension.

Two language corpora were used to investigate this: the Chinese Internet Corpus, containing texts downloaded from Internet, with 90 million words, compiled at University of Leeds (Sharoff 2006), and Taiwan-based Academia Sinica Balanced Corpus of Modern Chinese (Huang & Chen 1992). The results are summarised in Table 2.4.

Since the Chinese Internet Corpus contains texts that have been automatically downloaded from Internet, it is not clear if it is balanced, and it is hard to assess to what degree it reflects what a typical learner reads. The Sinica Corpus is a better source, as it was designed to balance different genres. Texts in the corpus are divided into 6 categories: philosophy (8%), science (8%), society (38%), art (5%), life (28%) and literature (13%). From this corpus it is only possible to extract character statistics for 80% and 95% word-text coverage. We can see that the numbers for both corpora are quite similar. One can, however, argue that the numbers we obtained are too high for estimates of characters that need to be learnt. The composition of the Sinica Corpus lets us assume that many of its texts may be hard to understand to many native Chinese speakers.

It must be noted that knowing these characters is a necessary condition (given the assumptions about the representativeness of the texts in the corpus, the required text-word coverage and the need to know all the characters in each word), but is not a sufficient one. Meanings of many multi-character words, although usually related to the meanings of the individual characters they consist of, often cannot be guessed, and need to be learnt by heart. Nevertheless, meanings of individual characters may serve as cues that facilitate remembering words.

We also need to account for the fact that both simplified and traditional characters are in use nowadays, and most educated native speakers of Chinese can read both. If we assume that a successful learner should have this ability too, the number of characters that need to be learnt will rise significantly. There are 2236 characters in the official, non-exhaustive simplification tables that contain simplified-traditional character pairs[9]. However, many of them are very rare, and unlikely to be needed in practice. A better estimate was found by cross-checking these tables with characters from HSK lists. Simplified characters that appear on these lists into were automatically converted into traditional characters, which showed that 1037 of them have traditional equivalents. As discussed in section 2.3, there is no one-to-one equivalence between traditional and simplified characters, so it is not an exact result. Nevertheless, it is a reasonable estimate. We can conclude that a learner that has learned enough characters in one variant (simplified or traditional), needs to learn about one thousand more characters to be able to read both variants.

[9] 簡化字總表 (List of Simplified Characters), Oct. 1986, `http://zh.wikisource.org/zh/簡化字總表`

CEFR Level	Overall Reading Comprehension
C2	Can understand and interpret critically virtually all forms of the written language including abstract, structurally complex, or highly colloquial literary and non-literary writings. Can understand a wide range of long and complex texts, appreciating subtle distinctions of style and implicit as well as explicit meaning.
C1	Can understand in detail lengthy, complex texts, whether or not they relate to his/her own area of speciality, provided he/she can reread difficult sections.
B2	Can read with a large degree of independence, adapting style and speed of reading to different texts and purposes, and using appropriate reference sources selectively. Has a broad active reading vocabulary, but may experience some difficulty with low frequency idioms.
B1	Can read straightforward factual texts on subjects related to his/her field and interest with a satisfactory level of comprehension.
A2	Can understand short, simple texts on familiar matters of a concrete type which consist of high frequency everyday or job-related language. Can understand short, simple texts containing the highest frequency vocabulary, including a proportion of shared international vocabulary items.
A1	Can understand very short, simple texts a single phrase at a time, picking up familiar names, words and basic phrases and rereading as required.

Table 2.5

2.5.3 Correspondence to the Common European Framework of Reference for Languages (CEFR)

We have seen a rather wide range of estimates for the number of characters that a second language learner of Chinese must know in order to have adequate comprehension of texts aimed at native speakers. Table 2.5 contains definitions of reading comprehension levels defined by CEFR (Council of Europe 2001), that will let us match them with the findings about the number of characters.

The lowest estimate we have considered is about 2000, the literacy threshold for urban residents in China. Since a person who barely passes the literacy threshold can at best be considered semi-literate, the reading comprehension of such a person is probably somewhere in the A2–B1 level range. The highest estimate at about 4400, based on text corpora, shows the number of characters that are probably enough for reading texts aimed at educated native speakers, that include philosophy, science and art. This corresponds to reading comprehension at least at the C1 level.

Achieving C1 reading ability is a very ambitious goal for most learners. Even the highest level of the new HSK test (HSK6) does not test reading

ability at that level. The correspondence of the HSK levels to CEFR is the subject of a controversy: according to the official statements, HSK6 (which assumes knowledge of 2663 characters and 5000 words that use them) is at the C2 level. However, all the estimates we have made so far make it clear that somebody who can recognise only 2663 characters has no way to read at the C2 level, which requires reading skills of an educated native speaker. German Association of Chinese Teachers concludes that the actual level of HSK6 is B2[10]. Our corpus study suggests that even this estimate may be too high, and more characters need to be learnt to achieve a B2 reading ability. We saw that the ability to recognise 3458 most frequent characters gives only 95% word-text coverage of the Sinica corpus, at which only a minority of readers can comprehend the text. This suggests problems larger than "some difficulty with low frequency idioms" mentioned in the B2 level description. However, the size of different text genres in the corpus plays a big role here and a different distribution of genres might have led to a higher word-text coverage and lower character recognition estimates.

Based on all these data, we can conclude that a useful target for a learner who wants to understand majority of non-specialist texts aimed at native speakers, and wants to have at least a B2 reading ability, most likely lies somewhere between the number of different characters on the HSK lists (about 2600 for the new, and about 2800 for the old exam) and the number of characters native Chinese should know after finishing secondary education, which also gives a minimal useful word-text coverage for a wide range of Chinese texts (about 3500).

[10] http://www.fachverband-chinesisch.de/sites/default/files/FaCh2010_ErklaerungHSK.pdf

Chapter 3

Psycholinguistic models of reading

The previous chapter showed that learners who want to read Chinese need to be able to recognise a very high number of graphical symbols (characters). Importantly, the structure of the symbols cannot be described by a small set of rules that would allow reliable decoding of sound and meaning associated with any given character. In order to better understand the process of reading Chinese, we will now turn to reading models. Firstly, we need a general overview of the field of second language reading, particularly word recognition. There has been very little research in the area of second language reading of logographic languages. Therefore, we will cover some general background information about the modelling of reading and some historically important models predominantly concerned with the first-language reading of English. Next, we will discuss differences between two major modern approaches to modelling reading: rule-based (represented by the Dual-Route Cascaded model) and connectionist. Since the connectionist approach is more suitable for modelling the process of learning to read, we will consider two connectionist frameworks: the Parallel Distributed Processing framework and self-organising maps. Next, we will look at the Lexical Constituency Model, which was built to model reading Chinese, and see how it differs from models of reading of alphabetic writing systems. Finally, the last section of this chapter shortly introduces the Modified Hierarchical Model, which may provide insights into how lexemes are linked into semantic information in the mental lexicon of L2 learners.

3.1 Second language reading

Koda (2005) describes processing of written words as a component process that consists of two operations: "obtaining a word's meaning and extracting its sound" Koda (2005, p. 31). Since reading involves analysing written characters, we have three processing components: orthographic processing, phonological processing and semantic processing.

Experiments show that words are processed in different ways during reading different languages, depending on the writing system. Frost, Katz

& Bentin (1987) compared word naming speed in Hebrew, English and Serbo-Croatian, and found that the naming speed was most affected by word frequency (slower for infrequent words) in the case of Hebrew, and least affected in the case of Serbo-Croatian. This can be explained by the orthographic depth hypothesis (Katz & Frost 1992), which assumes that in more transparent ortographies with direct grapheme-phoneme equivalences the phonological information is obtained directly from the graphic form, while in less transparent ortographies the word must be identified first, and only then the phonological information is obtained.

In the case of Chinese, the character is the smallest unit that needs to be identified to obtain phonological information. However, experiments suggest that for the native Chinese, the graphic form of whole words may be directly associated with meaning (Zhou & Marslen-Wilson 2000). As for second language readers, even those at advanced level, reading causes a higher cognitive load than for native speakers (Hayden 2005). The processing is therefore less automatic, and such readers are more likely to read character by character, obtaining phonological information of each character, and then grouping them into words and obtaining their meaning. We can therefore make a hypothesis that in the case of second-language readers who are used to other, non-logographic scripts, the semantic processing depends more on phonological processing than in the case of native Chinese speakers.

3.2 Reading-related variables and their effects

Before we discuss reading models, let us look at different variables that influence reading, listed by Cortese & Balota (2012). These variables were measured in behavioural studies, and one of the aims of different models is to account for effects introduced by these variables.

- **Frequency effect**: more frequent words are more quickly recognised. This is a very robust effect, confirmed by many different studies. A related variable is **familiarity** or **subjective frequency**. Word frequencies are usually taken from language corpora, but it does not account for the fact that different people encounter some words with different frequency. When conducting an experiment that uses word frequency as a variable, one may ask about perceived frequency in a questionnaire. Such a measure is obviously subjective, so its pros and cons need to be weighted.

- **Age of acquisition**: the degree of importance of this variable is controversial. Some studies claimed it is related to word recognition performance (e.g. Brown & Watson 1987). The problem, however, is that early learned words tend to be the frequent ones, and additionally the ones with higher imageability, and it is hard to separate these variables. Zevin & Seidenberg (2004) suggest that it is the **cumulative frequency** rather than the age of acquisition that has an effect on word naming. In other words, it is not the age at

which the word has been acquired that is important, but the number of times that the word was used during one's lifetime.

- **Orthographic length**: naming long words requires more time than naming short words. According to most studies this effect applies mostly to low-frequency words and non-words. In the case of Chinese characters, one could hypothesise that the number of distinct components in a character is analogous to the number of letters in a word in an alphabetic language.

- **Regularity and consistency**. These two concepts are related to two types of models described below, dual-route and connectionist, respectively. Dual-route models contain grapheme-to-phoneme conversion rules that can derive pronunciation of English words such as *hint* /hɪnt/ and *mint* /mɪnt/. The word *pint* /paint/ is irregular, because it cannot be pronounced using such general rules. In the connectionist models, however, there are no explicit rules, but the pronunciation is derived through analogy. In this case, words ending in *-int* is usually pronounced /ɪnt/, the word *pint* is therefore inconsistent. Many studies, e.g. Cortese & Simpson (2000), Jared (1997, 2002), have shown that consistency has a stronger influence than regularity on latencies.

  The notions of regularity and consistency can be transferred into the context of the Chinese writing system. For example, we can note that a vast majority of characters with the component 比 are pronounced *bi* (their tone may vary, but this is irrelevant for this and following examples in this paragraph). There is, however, an important exception: the character 昆 is pronounced *kun*. Moreover, there are several characters where 昆 *kun* is a component, e.g. 鯤, 崑, which are pronounced *kun*, too. We can therefore say that such characters are consistent, but not regular (unless we include rules that specifically deal with 昆 *kun*).

- **Feedback consistency**: the probability that a word pronounced in a given way is spelled in a given manner. For example, the English ending /-əʊn/ can be written as *-one* or *-oan*, it is therefore feedback inconsistent. The effects of feedback consistency are controversial; for example, Balota et al. (2004) found such effects both in lexical decision and naming, while e.g. Peereman, Content & Bonin (1998) found that there are no such effects when we account for word familiarity. The Chinese writing system is generally feedback inconsistent – there are usually many different ways of writing the same syllable; we can, however, observe different degrees of such inconsistency.

- **Neighbourhood size**: it can refer to the **orthographic neighbourhood size**, that is, the number of words that can be made from the target word by changing only one letter, or the **phonological neighbourhood size** – the number of words that can be made from the target word by changing only one phoneme. For example, the words

warship and *worship* are both orthographic and phonological neighbours, and *word* has a phonological neighbourhood that includes *work* and *ward*, while *lord* is only its orthographic neighbour. The effects of the neighbourhood are complex, and depend on the actual task and the number of neighbours.

The concept of phonological neighbourhood applies to Chinese just like to any other language. It should be noted that in the cases where a phonetic component of a character does not indicate the exact pronunciation, the actual pronunciation is often in its phonological neighbourhood, e.g. characters with the component 工 *gong* often have this pronunciation, but some are pronounced *hong* or *kong*. In the case of orthographic neighbourhood, the nearest equivalent would be the set of characters that differ in only one component.

3.3 Sequential bottom-up information processing models

Early models of reading were based on the view of the mind as a direct functional equivalent of a symbol processing machine, such as a computer. This view was predominant in cognitive science and related disciplines before the 1980s. Despite the clear inspiration by the computers, there was not enough computational power in that era to make computer simulations, so these models are not specific enough to make concrete calculations and predictions. However, even though these models are not in use anymore, they introduced several important concepts and influenced later models, and therefore they are worth mentioning. My description of the non-computational models follows Tracey & Morrow (2012).

The Information-Processing Model (Atkinson & Shiffrin 1968), which provides an overview of the information processing in the mind, consists of several components: *sensory memory*, *short-term (working) memory* and *long-term memory*, all controlled by executive control processes. As information from the senses reaches the sensory memory, it is processed by perception and arrives in the short-term memory. The contents of the short-term memory may be saved by the articulatory loop into the long-term memory, and retrieved later. The information is stored in the long-term memory in abstract interconnected structures known as schemata.

The above assumptions led to the development of several reading models. One of them was Gough's (1972) model, which divides reading into several stages: first the visual system stores the input from the eyes as an iconic image, which is then analysed by the scanner, which may recognise the image as a character. It is then decoded into a phonemic representation, which contains abstract representations of sounds. When a whole word is decoded into such a representation, it is looked up in the lexicon, in order to extract its meaning.

Another early model, the Automatic Information-Processing Model (AIPM), has been proposed by LaBerge & Samuels (1974). Its components include four types of memory: visual, phonological, episodic and semantic,

which are used during reading to store, respectively, images, sounds, context and meaning. This model describes also two types of attention: *external attention*, which is determined by behaviour, such as eye movements, and *internal attention*, which, as a state of the reader's mind and as such, is not directly observable. The latter is a key component of the model. An important aspect of attention is that its capacity is limited, more precisely, there is a limit to the amount of information that can be processed in a given time. Moreover, learning and repetition lead to automaticity: some tasks can be performed without the use of attention, which can be used for processing other information at the same time.

Similarly to Gough's model, AIPM can distinguish different phases of reading that include decoding and extracting meaning. Samuels (1994) describes the lower performance of beginner readers as a result of necessity to switch attention between the processes decoding and comprehending. For advanced readers, these processes have been automatised and therefore their performance can be higher.

3.4 Top-down and interactive models

In the two above-mentioned models, the information flow is one-way, bottom-up: from the sensory input all the way to the place where the meaning is extracted. There are many phenomena that cannot be described with such models. The knowledge of the wider context often aids lower-level processing, for example, understanding the meaning of a word may help extracting its pronunciation, and understanding the context of a sentence may be helpful to determine the meaning of an individual word. This led to the development of models that stressed sequential top-down processing, such as Goodman's (1967) model. It views reading as a "psycholinguistic guessing game", where the source of meaning is not in the text, but in the human mind that generates hypotheses about what comes next, reading is only used to confirm the predictions (Koda 2005).

Since the early 1980s there has been a growing understanding that both directions of processing are essential: background knowledge and context influence text comprehension, but reading involves eye fixations on almost every content word, and word recognition is crucial to understanding (Koda 2005). This is further confirmed by the study mentioned in subsection 2.5.2, which has shown that successful text understanding requires about 98% of the words in the text to be already known. In Rumelhart's (1994) Interactive Model such considerations were taken into account: this model allows non-sequential processing, with the information flowing both in the "bottom-up" and the "top-down" direction, as the readers use both information extracted from the text and their prior knowledge. Stanovich's (1980) interactive-compensatory model additionally accounted for the fact that the relative amount of different types of processing may depend on the reader: it may be mostly top-down for a poor reader with a lot of background knowledge and bottom-up for a good reader with little background knowledge.

3.5 Modern reading models

Even in the early days of reading models, not everyone postulated sequential processing. Cortese & Balota (2012) trace the roots of modern models back to Selfridge's (1959) pandemonium model of visual perception and Morton's (1969) logogen model. Selfridge postulated that written text is first independently analysed in terms of basic components, such as horizontal and vertical lines, and the patterns of the recognised features are used to identify letters. This idea was further developed in Morton's model: the logogens are word recognition devices that may have different levels of activation depending on their frequency: the more frequent a word is, the less further activation is needed to reach the recognition threshold. This allows to account for the frequency effect that was mentioned above.

The biggest problem with all the models discussed so far is that they are non-computational. They describe how various processes are supposed to work and their relation to each other, but are hard to falsify, as many of their aspects are left unspecified. Modern models, on the other hand, can be used to run actual computer simulations. This lets them make specific predictions, and if the predictions fail, they can be either adapted to new data or replaced by more robust ones.

The first computational model in the field of reading was McClelland & Rumelhart's (1981) Interactive Activation (IA) model. It was one of the pioneering works in the field of *connectionism* in cognitive science. The basic building block of connectonist models are units that work according to principles similar to that of neurons in the brain. Their inputs and outputs are connected to other units, and their activation depends on their inputs. If inputs are strong enough to make the signal exceed the unit's internal threshold value, it becomes activated and sends the activation signal to its outputs. Connectionism attaches an important role to the connections between units: they not only pass the information, but also vary their strength depending on the signal; frequent signals get transmitted with increasing strength and speed, while infrequent signals are transmitted slowly and may be ultimately inhibited. Moreover, there is a large number of connections between units, and they may form various patterns of strengthening and inhibition. These connections, even though they may turn weaker or stronger, operate according to the same basic connectionist principles.

The IA model already had the most important aspects of the modern models: it assumed processing at several levels of abstractions (visual features, letters, words, context), assumed that the processing is parallel and interactive, with top-down and bottom-up processes working simultaneously. The implemented model, however, only simulated relations between features, letters and words. McClelland, Rumelhart, Group, et al. (1986) and Seidenberg & McClelland (1989) introduced a more more general framework, called Parallel Distributed Processing (PDP). Its outline is presented in Figure 3.1. It is a model family rather than a single model. For example, the 1989 paper presented experimental results from a model that contained only phonology and orthography layers. It was followed by many others,

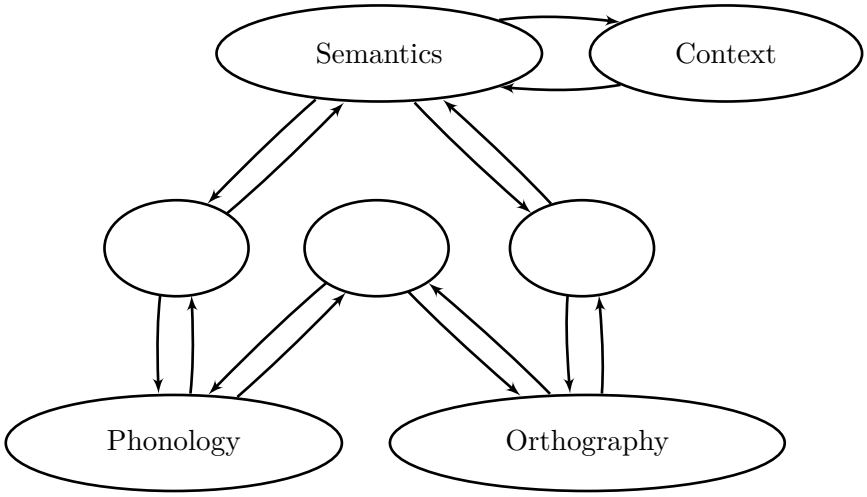

Figure 3.1: Parallel Distributed Processing framework described by Seidenberg & McClelland (1989)

e.g. Bullinaria (1996), Plaut et al. (1996), Plaut (1997), Zorzi, Houghton & Butterworth (1998), Harm & Seidenberg (2004). Each of them concentrated on a different aspect of reading and implemented a different subset of the structure presented in Figure 3.1.

As we can see, the PDP framework allows a bidirectional, interactive flow of information. The processing units that pass signals to each other include an orthographic processor, a phonological processor, a context processor and a meaning processor. The processors are connected to each other through hidden layers – additional sets of units that increase the number of parameters available to model, that with proper learning can make the processing more fine-grained.

PDP is not the only modern connectionist approach to modelling reading. An important feature of PDP is *supervised learning*: the model is given an item (e.g. a word) to recognise, and receives explicit feedback whether the answer was correct or not. An alternative approach uses self-organising maps (SOM), which use *unsupervised learning*: they do not receive any feedback and adapt their structure based on generalisations made from data. Both types of learning are important: people who learn to read depend on feedback to some degree, but only some part of their mistakes is actually corrected by a teacher or a peer. Most of the time they simply take the information they have and generalise them to new cases. The DISLEX model, which we will discuss later, is based on SOM and uses unsupervised learning, and therefore can be argued to be more psychologically plausible in this respect than PDP.

According to the connectionist models, the reading process is uniform and involves different activation patterns of units and their connections. Not all modern reading models, however, are based on these assumptions. A prominent example is the Dual-Route Cascaded model, presented in the following section.

3.6 Comparison of PDP and DRC models

The Dual-Route Cascaded model (Coltheart & Rastle 1994) assumes that there is a "lexical route" for processing well-known words and a "sublexical route" for previously unseen words and for the ones that have not been completely internalised. The lexical route in DRC is implemented as a connectionist model similar to McClelland & Rumelhart's (1981) IA. The sublexical route, however, is assumed to be functionally equivalent to a rule-based, sequential process that converts orthographic representation to its phonological equivalent.

There are two main arguments supporting the DRC architecture. Firstly, it is consistent with data from some types of dyslexia. Cortese & Balota (2012) point out the differences between people with surface dyslexia and those with phonological dyslexia. The former have problems with reading out loud irregular words, especially relatively infrequent ones, but no problems with reading words with regular pronunciation, even the ones they see for the first time. The latter display an opposite pattern: if they know a word, they read it without problems, no matter how irregular it may be, but cannot read unknown words. This double dissociation is much easier to account for in a double-route architecture, such as DRC. Another argument in favour of DRC is that it accounts for a rather wide variety of phenomena that can be measured during reading experiments, such as the above-mentioned frequency effect. The 2001 version of DRC is based on data from studies of over 20 phenomena.

Seidenberg (2012) agrees with the general idea behind computational models that can make specific simulations and predictions that may be confirmed or disconfirmed by behavioural studies. However, he strongly argues against the DRC model. He provides two types of arguments. On the one hand, he argues that the data presented in favour of DRC are not so strong. On the other hand, he describes the philosophy behind PDP, and argues why the PDP family as a more promising tool for further development of theories of reading.

Seidenberg (2012) notes that the impairments supporting the DRC model are rarely observed. Supporters of DRC choose them as the most informative. It is possible that they are just extreme cases in a wide range of impairments, caused by different degrees of recovery and influenced by individual differences. In most cases there is no clear indication of the double dissociation that is one of the main reasons for supporting DRC. Sandak et al. (2012), who conducted research on neurobiological bases of reading, present similar arguments. The original neurobiological models were shaped by studies of patients with dyslexia, and lead to favouring DRC in their line of research. Currently, neuroimaging studies of healthy subjects are seen as more informative. They do indicate that there are different patterns of activation for the activities associated with the two routes of DRC. There are, however, cooperative and competitive interactions between the two subsystems. This is inconsistent with the basic assumption of DRC that the two pathways are independent. Sandak et al. conclude that PDP is a more viable way forward for modelling neurobiological phenomena in reading.

Seidenberg also argues that not all the phenomena that DRC claims to cover are modelled properly and that DRC suffers from the problem of overfitting to specific studies. For example, it accounts for the interaction between frequency and regularity measured by Paap & Noel (1991), but fails when tested on stimuli from Seidenberg, Waters, et al. (1984) and Taraban & McClelland (1987).

Seidenberg downplays the importance of comparing how well models fit into data from particular studies. Unlike DRC, which is a single model for many phenomena, there are many different PDP models, each aimed at modelling a particular subset of the phenomena. Seidenberg does not see this as a problem, arguing that the models are just tools for formulating a theory of reading, and the high number of models is only a sign that different researchers concentrated their attention on different aspects of reading. His main argument in favour of PDP is that it is an architectural framework that assumes mechanisms that are neurobiologically plausible (as they are functionally modelled after neurons) and that they account not only for observed performance of a skilled reader, but also provide a model of learning.

It can be argued that Seidenberg, the co-author of PDP, does not give a fair assessment of DRC. However, even if we disregard all other arguments, the difference related to learning is fundamental. DRC contains at its core a set of phonological processing rules and no psychologically plausible mechanism of their acquisition, with many parameters set manually. On the other hand, PDP, as all connectionist models, is intrinsically based on learning – the parameters of the model are the weights of the connections between units, and they are re-adjusted in response to the learning items. DRC may turn out to be usable in some applications that require modelling learners who already fully acquired reading skills. In this thesis, however, we are concerned with the acquisition of reading. In these circumstances, it is clear that we should concentrate on the connectionist models.

Even though learning is an integral part of PDP, this framework has primarily been meant to model skilled reading. Therefore, it is not obvious if it learns in the same way as actual learners. Nation et al. (2012) discuss that problem with regard to one of the fullest implementations of PDP by Harm & Seidenberg (2004), and list three issues. The first one is the lack of pre-training on orthography. Children usually learn to recognise individual letters before they read words that contain them. Powell, Plaut & Funnell (2006) did a similar pre-training on a PDP model, and this led to results that were more similar to the data obtained from learning children. The second problem is related to the fact that the models are taught through supervised learning: they are given a training item to recognise, and they get positive or negative feedback that indicates whether their answer was correct. Based on the feedback, the parameters of the models are changed. As mentioned in the previous section, unsupervised learning is more common in practice and therefore more realistic. The third issue with Harm & Seidenberg's model is that it requires a very large training set: the same examples need to be repeated thousands of times. When people learn to read, however, they often need just a few exposures to a new word to learn it.

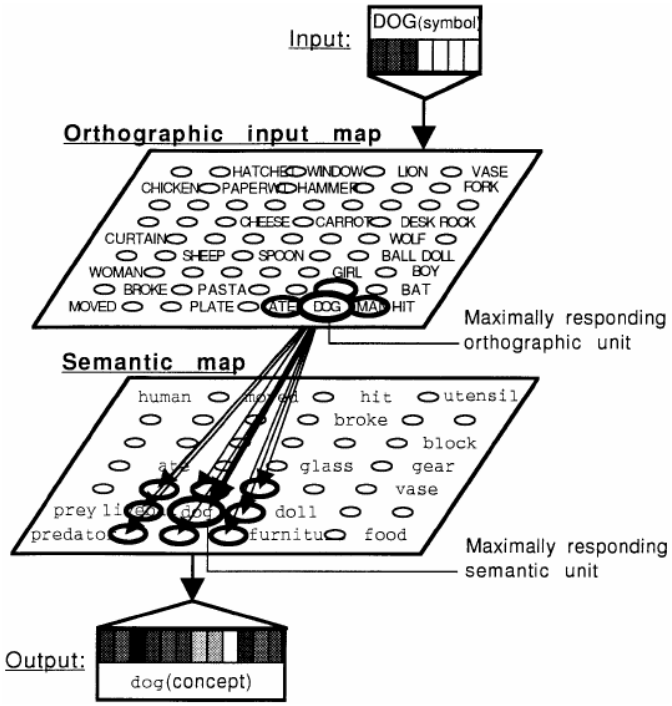

Figure 3.2: Mapping from the string DOG to the concept of dog in the DISLEX model (Miikkulainen 1997)

3.7 Self-organising maps and the DISLEX model

Self-organising maps (SOM), introduced by Kohonen (1989), can address some of the issues with PDP that were mentioned in the previous section. Most importantly, SOM uses an unsupervised learning algorithm, which finds structure in the input data without relying on any explicit error-correcting feedback. Another important feature of SOM is their neurological plausibility: "Not much is known about the structures underlying higher functions such as the lexicon. However, the perceptual mechanisms are very well understood, and they appear to be organized around topological maps. For example, nearby regions in the mammalian primary visual cortex respond to nearby regions in the retina" (Miikkulainen 1997, p. 334). Moreover, this model has been used to model both mono- and bilingual language acquisition (e.g. Miikkulainen 1997, Miikkulainen & Kiran 2009, Grasemann et al. 2011).

Self-organising maps take multidimensional data and organise them in a two-dimensional space of units that can be activated. In the learning process, input items are mapped into activation patterns in the output space. The input items are drawn at random, and at the beginning the activation is random as well. With time, however, similar input items begin to activate output units at similar locations.

Input data are represented with n-dimensional vectors, which may be simply represented as lists of n real numbers from 0 to 1. The crucial part

of creating a SOM model is creating an appropriate representation of input items: choosing the features (which are to become the dimensions of the space) and mapping them into numeric values. For example, a very simple model of decoding letters created by Miikkulainen (1997) maps each English letter to a number representing the relative number of black pixels in the letter's image. That is, the letter *M*, which has the most pixels, was mapped to 1.0, and other letters were assigned accordingly smaller numbers, e.g. *S*, which has roughly two times fewer pixels, was mapped to 0.518519. The model was in a way like a person with a very significant sight impairment, who sees completely blurred letters, and can tell them apart only by their relative difference in darkness. Even such a simple model was actually able to differentiate written words, and group them according to their graphical similarity. For example, the model located the words BOY and BAT next to each other, while CHICKEN was placed on the other end of the map.

Learning to read does not just consist of telling letters and words apart, but also involves mapping the graphical form into pronunciation and meaning. Therefore, a reading model should contain two or more SOM, linked by associative connections. Let us look at the DISLEX model (Miikkulainen 1997), which was created to test how introducing noise to the connections may cause dyslexic and aphasic effects. As mentioned above, SOM are meant to be neurologically plausible. This makes them also suitable for simulating learning processes of healthy individuals.

DISLEX contains several maps, but only two of them are of our interest now: the orthographic input map and the semantic map. The vectors for the semantic output were generated automatically with the so-called FGREP algorithm (Miikkulainen 1993), based on what roles they could take in different semantic frames. Each of the maps is organised during learning, according to the principles presented above. Additionally, units from the orthographic input map are fully interconnected with units from the semantic map. At first, the weight of all connections are equal, but they are modified in the process of so-called Hebbian learning: if two units are activated together, the weight of their connection is strengthened (Hebb 1949). For example, the unit representing the string of letters DOG in the orthographic input map is activated together with the unit from the semantic map that represents the concept 'dog'. Figure 3.2 presents the model after learning. It receives an input vector with numeric values of darkness of the letters D, O and G. We can see that several orthographic units are activated, but DOG is the maximally responding unit. It is linked to the concept 'dog', which gets the highest activation, but again, several related concepts are activated too. The closeness of units on the semantic map represents similarity of meaning, the same way as closeness on the orthographic map represents orthographic similarity. DISLEX makes it easy to account for word confusion: words that are likely to be confused will be represented as adjacent units on the orthographic and/or semantic map. Moreover, each wrongly activated unit for a given input is a sign of confusing one word for another.

3.8 The Lexical Constituency Model: a monolingual Chinese reading model

The reading models described in the previous sections were created with English, or languages with alphabetical writing systems in mind. It is important to look at a reading model that specifically takes the Chinese writing system into account: the Lexical Constituency Model, described by Perfetti & Liu (2006).

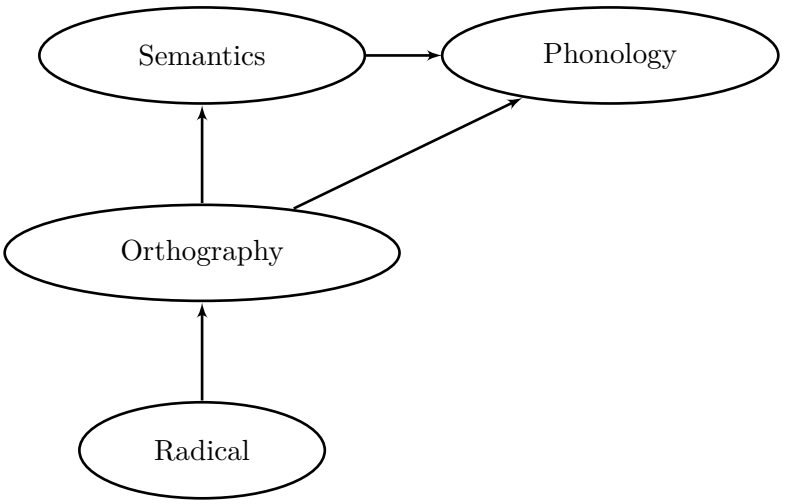

Figure 3.3: The Lexical Constituency Model described by Perfetti & Liu (2006)

Even though in the later models the processing order is not fixed, the general assumption of most models is that pronunciations are activated at an early stage, and take an active role in extracting meaning. In the case of Chinese such an assumption is far from obvious, as this writing system makes accessing phonological information more difficult, and a direct connection between the graphical form and meaning is possible. However, Perfetti & Liu (2006) cite studies that suggest that phonological information is a part of the word recognition process. Moreover, they argue that while the phonological information is not as reliable in Chinese as in alphabetic scripts, its validity is still high enough to be useful as an aid in word recognition.

The Lexical Constituency Model, as presented by Perfetti & Liu (2006), focuses on modelling recognition of a restricted set of 204 characters, but – as the authors suggest – "can be easily expanded within its general design principles". The model, shown in Figure 3.3, specifies 144 character components and four possible spatial relationships: "left-right, top-down, close outside-inside, or open outside-inside". This is called the radical level. They are combined in a different way to produce the characters, whose representations are at the orthographic level. It is in turn linked to the phonological level, which represents the pronunciation of the character in Pinyin, the standard romanisation scheme used in PRC. Then, meaning is represented by 204 items, each representing the meaning of an individual

character. Each character is linked to its basic meaning, and related meanings are related to each other.

An important feature of this model is that it specifically deals with the internal structure of the characters. In the case of alphabetic systems we can expect the letters to be treated as indivisible units, but in the case of Chinese characters their structure may be important for their processing. However, Perfetti & Liu (2006) admit that they did not take another important aspect of the characters into account – their different functions. As described in previous sections, the character components may have a semantic or a phonological function, and we may expect that depending on the type, the processing may be different.

3.9 The Modified Hierarchical Model of the mental lexicon

Problems with reading are not limited to extracting phonological and semantic information that was discussed above. Languages split the perceived reality into concepts, and very often there is no one-to-one equivalence between concepts from different languages. The lack of proper conceptual structure may cause problems with reading. We will therefore now turn into models of the bilingual mental lexicon, and see what it has to say about the acquisition of concepts in different languages.

To analyse how second language learners process a language, one needs to make assumptions about the general structure of the mental lexicon. We are interested in a bilingual model that can account for the learner's L1 and its influence over L2. We will look at Pavlenko's Modified Hierarchical Model (MHM), which was designed to address problems with three other models (Revised Hierarchical Model, Distributed Feature Model and Shared Asymetrical Model), at the same time retaining their strengths (Pavlenko 2009).

Figure 3.4 presents the overall structure of MHM. The upper part is concerned with the lexical knowledge: the knowledge of the words in either L1 or L2. The bottom part represents the conceptual knowledge. Concepts may be independent of words. For example, in many languages, *tongue* and *language* are expressed with the same words, even though they are clearly separate concepts for speakers of these languages, and are likely to be similar to the concepts of *tongue* and *language* that English speakers have. In other cases, speakers of different languages may have different boundaries between concepts. For example, even though the Chinese word 碗 (wǎn) is generally translated as 'bowl' and 盘 (pán) is usually translated as 'plate', there are objects that Chinese speakers are likely to call 盘 (pán), while native speakers of English would rather call them 'bowl'. Some concepts may be completely missing from the conceptual store. For example, 上火 (shàng huǒ) is a Chinese word that describes symptoms associated with a particular condition described by Chinese medicine. It is an everyday word in Chinese, but it refers to a concept that English speakers usually do not have.

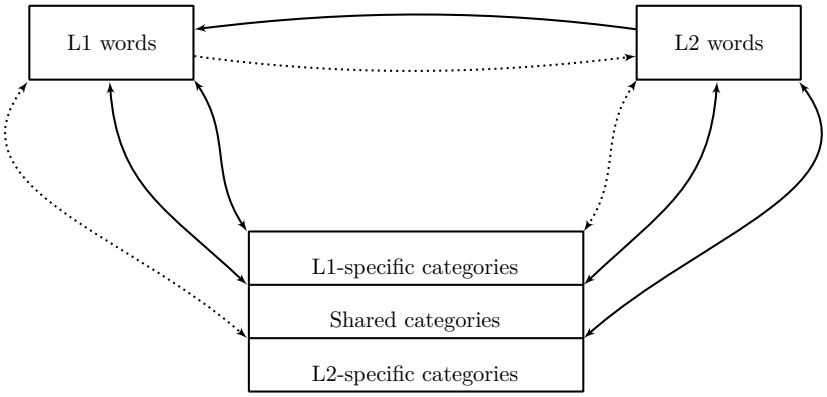

Figure 3.4: The Modified Hierarchical Model described by Pavlenko (2009)

The links between L1 words and L2 words show us the possibility of lexical transfer. Jarvis (2009) divides it into lexemic and lemmatic transfer. The lexemic transfer is concerned with phonological and graphemic properties of the word in two languages. Conversely, the lemmatic transfer is concerned with syntactic and semantic properties of words. Only the latter form of transfer is likely to take place between unrelated languages (Jarvis 2009). The lemmatic transfer may have an influence on perceived lexical relations of the words. For example, the words 短 (duǎn) 'short (of length)' and 矮 (ǎi) 'short (of stature)' have clearly distinct meanings, but are likely to be associated with the English lemma *short* and regarded as synonymous. This kind of mediation may also have an influence on the perceived category of the word and its hypernyms.

According to MHM, we can observe both the kind of lexical transfer described above and the conceptual transfer, where L2 words might access concepts without the mediation of L1, yet the conceptual structure still uses L1 categories. Only after significant exposure to the target language, L2-specific categories can start to emerge.

We can see that L1 words and concepts play an important role before L2-specific categories are built. In the context of this thesis, the assumptions of MHM will let us create a semantic representation of the learner's knowledge of L2 words. This representation will be a part of the model presented in the next chapters.

Chapter 4

Problem statement

This chapter begins with a summary of a pilot character recognition test that provides some information about the character knowledge of learners at different levels. It will motivate more research into character learning techniques. We will then consider several approaches to learning Chinese characters and take a closer look at one specific group: recognition-based approaches, which focus on learning to recognise characters and not writing them by hand. It is argued that given the recent widespread use of computer technology, the importance of recognition-based approaches is growing, as there are increasingly more learners who do not feel any need to learn to write by hand. However, these approaches are usually associated with individual learners, and there is little research about their consequences. We will look at how characters have been systematised by proponents of approaches that can be considered to be based primarily on recognition, and look at the important role of character components. We will then illustrate the problem of confusing characters with one another and argue that this is something that learners using recognition-based approaches may typically face. The chapter ends with questions about characters that are likely to be confused, reasons for the confusion, and more generally, about how the learner combines information about meaning, pronunciation and components of the character.

4.1 Pilot study of character recognition

In order to localise a possible plateau among learners of Chinese, a cross-sectional pilot study was performed among users of a website with an intelligent tutoring system for learning Chinese created by the author of this thesis (Kosek 2014). The users were asked to fill in a questionnaire and take a character recognition test. The questionnaire gathered information about how long one has been learning Chinese, the self-assessment of one's Chinese proficiency, and the number of days in a week one usually has some contact with written Chinese. The goal of the character recognition test was to find an objective indicator of learners' level. The estimation of the number of characters that they can recognise was done as follows (Kosek 2014):

The test was built using character frequency list in Modern Mandarin compiled by Da (2004). The characters were divided into groups, depending on ranking in the frequency list: characters with ranks 1-190, 191-375, 376-750, 751-1500, 1501-3000 and 3001-6000 (with each interval being roughly twice as long as the previous one). 6000 is, as noted above, an approximate number of characters highly educated native Chinese speakers usually know. The purpose of another, last group was to provide distractors, therefore it was made out of extremely rare characters, that even native speakers are unlikely to know[1]. The participants of the test have been presented 90 characters in random order, and asked to indicate as fast as possible whether they know a particular character. Each group was represented by a random sample of 13 or 12 characters.

The results of the test allow us to estimate the number of characters one can recognise. The percentage of recognised characters in each sample group indicates what percentage of characters in each group one can recognise. For example, if someone recognised 6 characters from the 1-190 group, out of 12, it would indicate that he probably knows about 50% characters from that group, that is, about 95 characters. Since the test subjects were extremely unlikely to really know the distractors, the number of distractors marked as recognised give an indication how likely someone was to mistakenly recognise an unknown character. This information was then used to appropriately decrease the estimate of known characters. For example, if 2 out of 12 distractors were marked as recognised, we would conclude that about 2/12 (16%) of the non-distractor characters were also marked incorrectly, and decrease the final estimate by 16%.

12 of the people who filled in the questionnaire and took the test indicated that they had been learning Chinese for at least 2 years and that they have contact with written Chinese at least once a week. We will focus on them, as only very few people can achieve fluency in any language with less than 2 years of practice or without regular contact with the language. Table 4.1 presents the data from the questionnaire along with an estimate of the number of known characters, based on the recognition test.

As discussed above, about 2000 characters is most likely not enough for comprehending texts aimed at native speakers, since it corresponds to the CEFR A2 level, or at most lower B1. The reading comprehension of the informants, apart from the last, seems therefore to be below the B2 level. Since the choice between intermediate and upper-intermediate levels in the self-assessment is not correlated with the character knowledge, it is likely to be a subjective difference in opinion what level can be called intermediate and what can be called upper-intermediate.

[1]The characters that have been chosen do not belong to the most common 20992 ones that make up a block of the so-called *CJK Unified Ideographs* (http://www.unicode.org/charts/PDF/U4E00.pdf), established by the Unicode Consortium.

Period of learning Chinese (years)	Self-estimate of written Chinese proficiency (CEFR scale)	Number of days in a week having contact with written Chinese	Estimated number of recognised characters
2.5	A1 (Newbie)	6	347
2	A2 (Elementary)	1	774
4	B1 (Intermediate)	2	832
3	B1 (Intermediate)	1	1063
2	B2 (Upper-intermediate)	4	1178
6	B1 (Intermediate)	1	1558
4	B2 (Upper-intermediate)	7	1779
2	B1 (Intermediate)	3	1846
4	B2 (Upper-intermediate)	3	1903
3.5	B1 (Intermediate)	2	2192
3	B1 (Intermediate)	1	2308
3	C1 (Advanced)	7	3106

Table 4.1

We can therefore conclude that most learners in this experiment experience a plateau at the intermediate stage. Even after 3 years of learning, the majority of people who attained intermediate level do not progress further, despite regular contact with the written language, and their reading comprehension does not seem to be enough to read texts aimed at native speakers with sufficient comprehension. Possible reasons and solutions to this problem will be discussed below. We can hypothesise that finding an effective method to learn to read enough Chinese characters in reasonable time would be a large step towards achieving adequate comprehension.

4.2 Character learning approaches

There are many ways character teaching approaches can be classified, Figure 4.1 presents one possibility of a partial and approximate systematisation. The word-centred approaches concentrate on teaching characters in context on words and the character-centred ones involve specific attention to each individual character, regardless of whether it represents a word, a bound morpheme or just a syllable. Traditional character-centred approaches involved learning to write by hand, but recently we can see the growing number of learners who use modern technology to type characters and therefore have much less experience with handwriting than learners in the past. This led to the development of approaches that teach recognising characters, but do not focus on writing them. One of such approaches involves learning systematic correspondences of some character components to meaning and pronunciation. The fact that these components are relevant for teaching suggests that they may play an important role in the

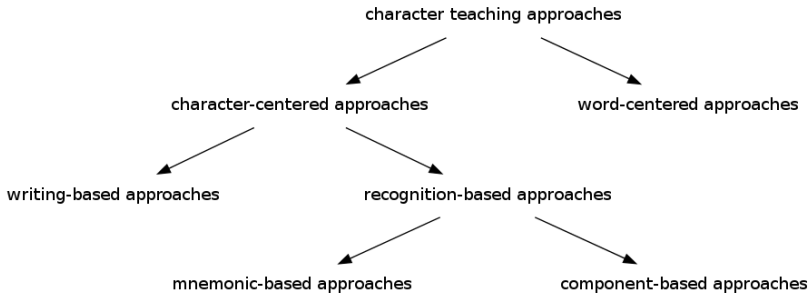

Figure 4.1: A possible systematisation of some of the approaches to teaching Chinese characters.

representation of characters in the learner's mind. Therefore, in the next sections we will consider a particular problem that learners are often faced with, namely character confusion, and pose research questions that may lead us to better understanding the role of character components in the representation of characters.

4.2.1 Difficulty with building the graphemic conceptualisation

The big problem with learning Chinese characters is building graphemic conceptualisation. Sound and meaning representation can be transferred from one's native language. They may be slightly incorrect, but it is a starting point, that can be refined as one gets more input. However, we have seen in chapter 2 that the Chinese graphemes are very different from components of other writing systems, and the lexemic transfer is therefore extremely difficult, unless the learner knows at least pronunciation of a character in question. It means that the learners, even if they use context to pick up a meaning of an unknown character, are much less likely to commit it into the memory, if they do not already recognise the components of that character. To solve this problem, one may rely less on a meaning-centred approach to learning to read, and switch a more character-centred approach, as suggested by Lam (2011). In this approach, the focus is on learning individual characters, out of context, which allows the student to pay greater attention to new characters and notice their graphical components, and link the graphemic representation to the meaning and pronunciation.

4.2.2 Relation between reading and writing characters

Traditionally, the character-centred approach was closely related to learning to write characters by hand. It certainly requires paying attention to the elements of the character and improves noticing of how they are built. However, learning to write takes very much time, both in the case of first- and second-language learners.

Until recently, everyone who wanted to attain a high level of proficiency

in the language had to learn to write by hand anyway. However, following the development of computer technology, handwriting skills are less and less useful in practice. Computer-based input methods require typing pronunciation of a character or a word, and choosing the right one from a list, in the case of homophones. Therefore, typing Chinese can be done by anyone who can speak and can read. We can also look at this issue from the perspective of second-language proficiency levels, such as CEFR. Should they, in general, be based on what native speakers know, or on what they actually do? Chinese natives themselves forget how to write some characters increasingly often and this does not impair their ability to read. Moreover, Chinese usually can write only one of the character variants (traditional or simplified), but can read both. This shows that learning to read without learning to write by hand is possible, and in many situations desirable.

However, it does not necessarily mean that one should not learn to write at least some of the characters. Handwriting may be in some cases the best method to strengthen the knowledge of some characters and to commit them into memory. Still, such an approach clearly differs from the traditional one: learning to write is not a goal in itself, but just a mean to attain the goal of learning to read. The ubiquity of computer technology is very recent, therefore there has not been enough research of this approach, and it is not known what the best strategy is, for which characters one should learn reading and writing, and for which ones learning to read is sufficient.

The issue of usefulness and scope of learning to write by hand is controversial. The debate started with Xu & Jen's (2005) article and continues until today (Peng 2016). Bi, Han & Zhang (2009) postulate that even in the case of Chinese reading does not depend on writing, and Allen (2008) makes the point much stronger: "learning to write Chinese is a waste of time". On the other hand, Zhang & Reilly (2015) found that writing Chinese characters facilitates their subsequent recognition.

There may be no final answer to this debate, as it may depend to a large degree on the learner's goals. What is important to note, however, is that many self-learners never get to learning to write, because the communicative tasks they are exposed to never require them to do so. Therefore, regardless of the usefulness of learning to write, the question we should ask is: What problems are specific to recognition-based approaches to learning characters?

4.2.3 Semantic and phonetic character components

Among teaching approaches that are focused on learning to recognise characters, and not necessarily write them, we can distinguish two groups (which may overlap): the first one is based on mnemonics, and the second one is based on learning systematic correspondences between components and pronunciation and/or meaning. As argued in subsection 2.4.5, the mnemonics may vary a lot from approach to approach and from learner to learner. Therefore, any representation of the characters that involves mnemonics is very learner-specific and is hard to generalise. On the other hand, the lists of semantic and phonetic components are relatively constant across approaches (despite some differences), and therefore, the

representations based on these methods should not vary so much and are much more likely to be generalisable.

Ann (1982) presented a noteworthy component-centred approach by systematising 5888 Chinese characters. However, Ann's decomposition of characters is based on traditional Chinese views on what constitutes a semantic and a phonetic component of a character. Because of diacronic changes of sound and meaning, in many cases phonetic components of the characters have lost their relation to the pronunciation, and semantic components are not related to the meaning of the character.

The above-mentioned problem has been partially solved by Guder-Manitius (1998), who analysed the semantic and phonetic components of 3867 characters. He has investigated which of the components really have some semantic or phonetic value in modern Chinese. He analyses objective features of the characters – for example whether a particular component has some common meaning or common pronunciation across different characters. This can be a starting point for the study of representation of characters in the learner's mind.

Guder's conclusion is that the components he found are relevant for teaching characters. Looking at it from the psycholinguistic perspective, we can ask the question: if these components are indeed pedagogically useful, does it mean that they are used to represent the characters in the learner's mind?

We can look at lexical models as a starting point to handle this question. Pavlenko's (2009) Modified Hierarchical Model, presented in section 3.9, postulates that phonetic and graphemic information is at the lexemic level. This model was primarily created with phonetic scripts in mind. Here we are concerned with learners of Chinese who have a language with a phonetic script as their mother tongue, and therefore we can expect that such organisation of the mental lexicon is transferred to some degree to their L2, Chinese. If a learner can guess the pronunciation of an unknown character, this phonetic information may additionally strengthen the new character in memory, therefore components that may serve as reliable phonological cues may be good candidates for early learning.

The scope of the problem of character representation is very large. In order to make it more focused, we may concentrate on one particular problem with learning, gather relevant data from the learners, and see whether the above-mentioned character features can be used to describe the data. The problem that we will now concentrate on is character confusion, which occurs when one character is mistaken for another by the learner.

4.3 Phonetic, semantic and graphemic character confusion

The problem with learning characters is not restricted to graphical complexity, that was discussed before, but also to problems with acquiring appropriate contrasts. As Saussure already pointed out, acquiring appropriate phonological and semantic contrasts is an important part of

language learning. In the case of the Chinese characters we need to additionally take graphemic contrasts into account.

Acquiring the semantic distinction between two characters may be hard, especially when the pronunciation is the same or similar, yet this is something very frequent in Chinese. For example, both 作 zuò and 做 zuò have the same pronunciation (*zuò*), with the former meaning 'to write, to compose' and the latter meaning 'to do'. Somewhat surprisingly, the word 工作 gōng zuò 'work, labour' uses the former, not the latter, which shows that the distinction between the two is often not obvious for a learner. But even if the semantic distinction is clear, there are many cases where the characters are easy to confuse because of same or similar pronunciation and belonging to the same semantic field, as in 燕 yàn and 雁 yàn, both pronounced *yàn*, which mean 'a swallow' and 'a goose', respectively.

Even if both the meanings and the pronunciations are distinct, two characters may be confused because they look similar. The cases of mistaking one character for another do not seem to be directly related to the complexity of the characters, but rather to their structure and the similarity of their most salient components. An example of the graphical structure of the characters being a likely reason for confusion can be seen in the pair 豫 yù 'pleased, delighted' and 橡 xiàng 'oak'. They are quite easy to mistake for one another, even though their meanings and pronunciations are different. A possible factor is that, despite the different pronunciations, the two characters actually share the component 象 xiàng, which does indicate pronunciation in many characters: 象 xiàng 'elephant', 橡 xiàng 'oak' and 像 xiàng 'be like'. Another factor is that the components 子 and 木, while different, are easy to confuse, especially when they are squeezed on the left-hand side of the character.

The problem with character confusion is especially likely to occur when the learner uses a recognition-based approach. Learning to write characters makes the learner more aware of individual strokes, and makes it easier to pay attention to relatively small differences. In recognition-based approaches, the learner probably pays less attention to the stroke level and begins with a more top-down approach, trying to recognise characters and their components as a whole.

4.4 Research questions

This chapter began with a pilot study of character recognition that showed that long-time learners of Chinese may often have problems recognising Chinese characters. Then it was argued that recognition-based character teaching approaches are becoming more popular, and therefore more research on these styles of learning is needed. The class of approaches that focus on shared phonetic and semantic components was considered as a source of data that may show phenomena that are present among many learners. A particular problem was described: character confusion, which is likely to occur when learning characters using recognition- and component-

based methods. Therefore, the goal of my study is to answer the following questions:

1. What characters are likely to be confused in the process of learning Chinese script?
2. If two or more characters are confused with one another, we may assume that they share some features or a combination of features. Which features cause character confusion?
3. What role do character components, meaning and pronunciation play in the representation of Chinese characters?

Chapter 5

Methods and data

5.1 Definition of *character confusion*

Let us say that we represent each character with a triple: <graphical form, pronunciation, definition>, where each element of the triple has a form that the learner is likely to encounter when learning the character. For example, a reasonable choice would be to have the graphical form represented as the picture of the character in the regular script, pronunciation represented as Pinyin romanisation, and meaning represented as a list of most common English definitions of the character taken from the learner's dictionary. For example, the character 花 could be represented as `<花, huā, "a flower/ to spend (money, time)">`.

Let us say that we present the user with the graphical form of a character, which we will call A. Character A is *confused* with another character, B, if:

1. the learner CAN provide a definition that he/she believes to be associated with the character A,
2. the definition provided by the learner does NOT match the definition of the character A,
3. the definition provided by the learner DOES match the definition (or one of the definitions) of the character B.

Confusion with more than one character is possible, too. Character A is *confused* with a set of characters B, C, ..., if:

1. the learner CAN provide several definitions, and he/she believes that one of them is associated with the character A, but he/she is not sure which one,
2. NONE of the definitions provided by the learner match the definition of the character A,
3. the definitions provided by the learner DO match the definitions of respective members of the set B, C, ...

Sometimes, even if the learner identifies the character correctly, a kind of confusion may take place. Let us say that character A is *almost confused* with character B, if:

1. the learner CAN provide several definitions, and he/she believes that one of them is associated with the character A,

2. after possible hesitation, the learner chooses one of the definitions as most likely, and it DOES match the definition of character A

For the confusion to take place, the learner needs to somehow recognise the character. There is no confusion when the definition provided by the user is correct, but there is no confusion either when the user cannot come up with any definition at all.

Note that *definition* is a string of English words, and it is something that only indirectly represents *meaning*. In the minds of learners who already have some experience with the language, the meaning of a character will likely be represented by a prototypical image of what it represents and the contexts the character has been seen in, and it may, but certainly does not have to be linked to an overt English definition. However, our primary concern here is whether the learner associates the graphical form of one character with the meaning of another character. The accuracy of the learner's meaning representation is not relevant here. For these purposes, the representation of the meaning with English definitions is sufficient. For example, one learner may associate the above-mentioned character huā 花 with a prototypical image of a flower, another learner may associate it with the phrase huā qián 花錢 'spend money', and yet another learner may associate it with the English verb *spend*. Let us suppose that these learners are asked to provide the pronunciation and an English definition of the character. Despite the different meaning representations, they will likely come up with something similar to either 'flower' or 'spend time or money'. Despite that they can provide the English definitions, it may happen that they cannot to use this character in the right way in the right context. But as far as we are concerned, the definitions are good enough: no other frequent character has meaning that can be expressed in English as 'flower' or 'spend time or money', so we can conclude that the character has not been confused with any other. Conversely, if the definition provided by the learner was just 'spend', we could not rule out confusion with the character guò 過 'to cross / to go over / to spend (time) / to undergo / to exceed'. In this case we need to look at the pronunciation provided by the learner. If it resembles *hua*, there was no confusion (and the meaning that the learner associated with the character may or may not be accurate, but this is irrelevant). If the pronunciation resembles *du*, the confusion was likely. Finally, if the pronunciation resembles neither *hua* nor *du*, it means that the character huā 花 is not learnt properly, but, since there is no frequent character that can be translated as 'spend' and is not pronounced *du* or *hua*), there is no evidence that the learner confused huā 花 with another character. For similar reasons, we are not concerned with accuracy, so if a character has several meanings and the learner comes up with a definition that matches only one of them (e.g. only 'flower' or only 'spend time or money' in the above-mentioned case), we can still consider the character to be correctly recognised: it means that

the initial association between the shape and the sound and meaning has been created, even if it is not fully accurate.

As we can see, the definitions of confusion are primarily concerned with the semantic information, and pronunciation is a secondary factor. It stems from the fact that a single meaning is usually associated with one character form and one pronunciation. On the other hand, a pronunciation of a syllable is usually associated with many meanings and character forms, so even if we interpreted wrong pronunciation as a confusion with another character, it would be hard to find out which character that would be.

5.2 Data gathering

5.2.1 Diary study and self-observation

With the definition of confusion in place, we can discuss the way of gathering the data. The most straightforward way to find candidates for confusion is to list characters, and ask learners to provide their definitions. With such a method of data collection, it is hard to get more than a few dozen answers from a given learner, and only a part of them is likely to exhibit character confusion. Moreover, confusion is a dynamic process that does not occur every time, so an individual questionnaire is not very reliable if it is filled in by a learner only once. Since there are about 3000 characters we are interested in, this way of data collection requires a large number of informants, informants that are willing to fill in very long questionnaires, or that are willing to repeatedly fill in questionnaires over a relatively long period of time. This is hard to organise in practice, and another solution was therefore adopted: an introspection of my own learning, supplemented by a diary that I used to record what characters I had confused in the process of learning.

There is no consensus about definition of a diary study, in particular, whether it includes studies that involve self-observation (Matsumoto 1987). In any case, regardless of how they are called, self-observation has been employed by second language acquisition researchers to put forward important hypotheses, e.g. the noticing hypothesis (Schmidt & Frota 1986).

Diary studies have some disadvantages: they are time-consuming and may not be generalisable to other learners. Moreover, a large degree of subjectivity is involved, and this is even larger in the case of a self-observation study, where the researcher and the test subject is the same person. Despite of these limitations, self-observation may be a useful tool. It makes it possible to access data that otherwise are hard to obtain. These data are not sufficient to draw conclusions, however, they may indicate some tendencies and be useful for formulating hypotheses, which may be later tested by other means. Moreover, while the subjectivity is impossible to avoid, a well-defined format of the diary may make different entries possible to compare. Moreover, in order to be as objective as possible, repeatable parts of data analysis can be done by computer.

5.2.2 The learner's profile

Since the data gathering is concerned with my own learning, some background information needs to be provided. I learned to write about 1000 most frequent characters during the Bachelor-level study of Chinese between 2010 and 2012. In the later period I did not practice writing, but continued to learn characters using a recognition-based approach. I learned characters in batches, grouping them by recurring phonetic components, paying attention to differences made by different semantic components. This is a typical example of a character-centred, recognition-based and component-based learning method that was described in the previous chapter. Using this method I ultimately learned 3437 characters, which covered all HSK levels (see section 2.5) in both simplified and traditional variants.

I continued to review the characters using flashcards in the mobile application Pleco[1]. The flashcards used spaced repetition (Woźniak & Gorzelańczyk 1994), which automatically schedules repetition of items, based on their difficulty. That is, characters that were recalled without difficulties were repeated in increasingly large intervals, while the characters that were not recalled correctly, were repeated as often as needed. An important consequence of this kind of repetition is that it significantly diminishes the frequency effects: the characters that are seen most often are not the ones that are most frequent in natural language texts, but the ones that are the hardest to remember by the learner.

5.2.3 Format of the diary

The diary has been kept for about two years and contains a log of the cases where one character was confused for another. The format of the diary is as follows:

$$AB_1B_2 \ldots B_n$$

where A is the target character that was to be recognised, and $B_1 \ldots B_n$ are the characters that were confused with A. The line may contain an optional annotation *(almost)* which indicates that $B_1 \ldots B_n$ were *almost confused* with A (see definitions in section 5.1).

Moreover, the entries contain the Pinyin transcription and information about my intuition of the most basic meanings of the characters, e.g.

獲猛護 **huo4** *trad* 'capture, catch' vs **meng3** 'fierce' vs **hu4** *trad* 'protect'

The diary contains about 2500 character pairs, which correspond to over 1500 distinct cases of confusing one character for another. If we base the research only on these data, it will have to be treated as a case study, as it is hard to predict to what degree the findings are generalisable. On the other hand, these data can also be a basis for questionnaires that can verify

[1] http://www.pleco.com

whether other learners of Chinese characters tend to have a similar pattern of character confusion. Moreover, these data can be used to formulate hypotheses about how the characters are represented in the learner's mind. In the next sections we will look at the confusion patterns in the data, their possible interpretation and the ways we can use them to build a connectionist model of character acquisition.

5.3 Confusion patterns in the gathered data

A subset of the data has been analysed and categorised according to my own assessment of the reasons for confusion; Table 5.1 presents some examples. As mentioned in the previous section, such an assessment is inevitably highly subjective, but it is a way to formulate hypotheses about how characters get confused. The analysis showed three main categories of the reasons for confusion: component similarity, phonetic similarity and semantic similarity. However, a more detailed inspection revealed that some of them have subcategories, and some cases of mixed confusion cannot be classified under any of these main categories, as shown below.

- **Component similarity/graphical similarity** may have a lot of reasons, and therefore this category has a few subcategories:
  - the same semantic component, e.g. 纟 in 绩 (jī) 'achievement' and 缀 (zhui) 'sew'
  - graphically similar semantic component, e.g. 火 and 饣 in 炊 (chuī) 'cook' and 饮 (yin) 'drink'
  - semantically similar semantic component, e.g. 水 'water' and 酉 'wine vessel' in 浆 (jiāng) 'thick liquid' and 酱 (jiàng) 'soy sauce'
  - the same phonetic component, e.g. 伺 (cì) 'spy on; wait on; wait upon, serve' and 饲 (sì) 'rear; feed'
  - graphically similar phonetic component, e.g. 陡 (dǒu) 'steep' and 徙 (xǐ) 'move'
  - general graphical similarity, when two characters taken as whole seem graphically similar, even though it is hard to explain it by similarity of individual components, e.g. 黨 (dǎng) 'political party' and 墨 (mò) 'ink'
- **Phonetic similarity** takes place when the actual pronunciation of the two characters is the same or similar, regardless of whether the it is indicated by the phonetic components, for example 秤 (chèng) 'scales' and 称 (chēng) 'weigh'.
- **Semantic similarity** takes place when the meanings of the two characters are similar, e.g. the above-mentioned 炊 (chuī) 'cook' and 饮 (yǐn) 'drink'

Target character	Confused with	Likely reason
jī 绩 'achievement'	zhui 缀 'sew'	same semantic component
chèng 秤 'scales'	chēng 称 'weigh'	similar pronunciation same semantic component; similar meaning;
chóng 崇 'sublime'	zōng 宗 'ancestor'	same phonetic component; similar pronunciation
chéng 懲 'penalise'	fá 罚 'penalise'	both characters co-occur in the word 懲罰 'penalise'
chǒu 丑 'ugly'	chòu 臭 'smelly'	similar pronunciation; similar meaning
chǒu 丑 'ugly'	qiū 丘 'mound'	similar graphical form
chuī 炊 'cook'	yǐn 饮 'drink'	semantic component; same phonetic component; graphically similar similar meaning
shī 施 'carry out'	tuō 拖 'pull'	same phonetic component
dí 敌 'enemy'	gù 故 'incident; former'	same semantic component; similar phonetic component
dì 递 'hand over'	dǐ 抵 'arrive at'	similar pronunciation; the semantic component 辶 'walk' is more related to the concept 'arrive' than to the concept 'hand over'
cán 残 'damage'	jiān 歼 'annihilate'	same semantic component; similar meaning; the phonetic component suggests the pronunciation *jian*

Table 5.1: Example confusion data with hypothesised reason for confusion

– collocational co-occurrence is a special case of semantic similarity; characters in some bisyllabic Chinese words have the same meaning, or do not have any particular meaning on their own, and occur only in a specific combination, e.g. 惩罚 (chéng fá) 'punish', 蝴蝶 (hú dié) 'butterfly' and 珊瑚 (shān hú) 'coral'; we may classify the cases of confusion 惩 (chéng) and 罚 (fá) as a confusion due to semantic and collocational similarity

As mentioned above, there are cases that do not fit into any of the categories above. They can be categorised as likely occurrences of a mixed confusion:

- **Semantic component of the target character related to the meaning of the confused character**. For example, the character 递 (dì) 'hand over' contains the component 辶 'walk'. This character was confused with the character 抵 (dǐ) , which has several meanings that include, among others, 'arrive'. A likely reason for confusion is that the semantic component 辶 'walk' seems more associated with the meaning 'arrive' than with the meaning 'hand over'. Interestingly, the confusion is also likely to occur in the opposite direction: the character 抵 (dǐ) contains the semantic component 扌 'hand', which is likely to be more associated with the meaning 'hand over' than with the meaning 'arrive'.

- **Phonetic component of the target character related to the actual pronunciation of the confused character**. The confusion between 残 (cán) 'damage' and 歼 (jiān) 'annihilate' may be an example: the component 戋 (jiān) suggests a pronunciation similar to *jian*, but 残 (cán) is pronounced *cán*. Therefore, the correspondence between the sound value of the phonetic component of 残 (cán) with the actual pronunciation of 歼 (jiān) makes the two characters easy to confuse.

5.4 Connectionist model of character learning

The learner data described in the previous sections come from one particular learner. The likely reasons for confusion discussed above suggest that the reasons for confusion may not be learner-specific, as they are based on quite general features of characters, such as their meaning, pronunciation, and components that can be found in several characters. The next step toward generalising the findings to other learners is to create a model that can be used to predict what characters are likely to be confused. Later, predictions of such a model can be tested against data from other learners.

Chapter 3 described advantages of using connectionist models to simulate the acquisition of Chinese characters. A study that is closest to the research goal of this thesis has been done by Xing, Shu & Li (2002). They modelled acquisition of Chinese characters with self-organising feature

maps, particularly with the DISLEX model, described in section 3.7. The present study is going to be different in a few ways. While Xing et al. use only parameters that are related to the pronunciation and the shape of a character, this study needs to include the semantic information. It is likely that similarity of meanings of two characters contributes to the likelihood of confusing them, so the information about such semantic similarity must be provided *a priori*. The goal of the present study is different, too: Xing et al. are interested in the accuracy of mapping from the characters' shape to their pronunciation, while this study is concerned with mapping from the shape to the meaning, and finding out how much confusion it causes (that is, how often a shape gets associated with a meaning of another character). Finally, here we are concerned with modelling non-native speakers that can recognise, but not necessarily write Chinese characters.

Section 3.7 stressed that a proper representation of input data is crucial for the usefulness of connectionist models. In other words, we need to represent the features of Chinese characters that are important during their recognition, decoding their sound and meaning; these features should also be relevant to the process of confusion, which we are interested in. The following description discusses the initial representation. The next chapter shows that it turned out not to be sufficient and describes how the representation was improved.

Subsection 2.4.5 concluded that two types of character components are particularly important: semantic and phonetic, and referred to Guder's (1998) list of 122 semantic components and 683 phonetic components that exhibit clear patterns in modern Chinese characters. Learning to read involves creating a mapping from the graphical forms to the meaning, with various grades of phonological mediation. In the model, the orthographic map needs to contain information about the graphical structure of characters (which includes semantic and phonetic components), and the semantic map, must have information about the meaning. Our definition of character confusion involves only meaning and not pronunciation. Therefore, representing the phonetic information in a separate map is not crucial. The DISLEX software can only use two maps in a simulation, so the phonological information can be either joined with the orthographic representation, or with the semantic representation. The orthographic map can represent characters in the form of semantic and phonetic components. If a character does not contain elements from Guder's list, we treat it as indecomposable.

Pronunciation is represented as a sequence of a consonant, a glide, a vowel and a coda, and a tone, and different possible values are mapped into different discrete numbers in the range from 0 to 1. This can be done using the Phonological Representation Database for Chinese Characters by Zhao & Li (2009). Since we assume that the learner becomes acquainted with meanings of new characters through a Chinese-English dictionary, we can conclude, in accordance with Pavlenko's MHM model, that conceptual access is mediated by English lemmas, presumably taken from the character's English definition. Therefore, we can approximate the meaning representation with the representation of English words. We can use all words from the English definitions of characters in question as

the features; if the word X occurs in the English definition of the Chinese character Y, it mean that the feature X of the character Y is equal to 1, and otherwise it is equal to 0.

Our main goal is to model character confusion. The DISLEX model provides a psychologically plausible account of some aspects of word learning and "captures some of the physical structures underlying the lexical system in the brain" (Miikkulainen 1997, p. 356). Therefore, we can expect that a reasonable model will represent easily confusable characters as points that are close to each other on at least one of the maps. This proximity may lead to higher activation of such items when the target character is supposed to be activated, and the cases of confusion would occur when the activation becomes higher than the activation of the target character.

Chapter 6

Experiments

The present chapter describes the connectionist simulations performed with the DISLEX package, which has been adapted to the problem of training the recognition of 3437 Chinese characters that cover all the HSK levels in both simplified and traditional variants.

6.1 Lists of semantic and phonetic components

The discussion in the previous chapter made it clear that semantic and phonetic components are likely to play important roles in the representation of the characters. Deciding which components are truly semantic is not easy; for the purpose of the experiment Guder's (1998) list of 122 components was used, as they had been shown to have a semantic value.

The phonetic components are easier to discover automatically, and a program was written specifically for this purpose. For each component it checked the pronunciation of the characters that contain it (limited to the 3437 characters that we are interested in). Some components are associated with too many different pronunciations to be useful. Others, however, indicate pronunciation quite reliably. For example, all the 4 characters that contain the component 曼 (慢$^{\text{màn}}$, 漫$^{\text{màn}}$, 馒$^{\text{mán}}$ and 蔓$^{\text{màn}}$) are pronounced *man* (with different tones). Guder (1998) noted that the number of pedagogically interesting phonetic components can be increased if we consider not only components associated with one specific pronunciation, but also consider

Group	Pinyin	IPA
bilabial/labiodental stop/fricative	b, p, f	p, p^h, f
alveolar stop/fricative	d, t	t, t^h
velar stop/fricative	g, k, h	k, k^h, x
alveolo-palatal affricate/fricative	j, q, x	$t\c{c}$, $t\c{c}^h$, $\c{c}$
retroflex affricate/fricative	zh, ch, sh, r	$t\c{s}$, $t\c{s}^h$, $\c{s}$, $z\c{}$
alveolar affricate/fricative	z, c, s	ts, ts^h, s

Table 6.1: Grouping of the initial consonants

pronunciation variants. Therefore, following Guder, the syllables with similar initial consonants were grouped according to the place of articulation and disregarding the manner of articulation. For example, *ta* was grouped with *da* (which differ in the aspiration of the initial stop consonant), and *ca* with *za* and *sa* (which all begin by alveolar consonants: an aspirated affricate, an unaspirated affricate and a fricative, respectively). All the groups are presented in Table 6.1.

A component was classified as phonetic when it fulfilled the following conditions:

1. over 40% of the characters that contain this component have the same pronunciation (disregarding the tone) or a similar pronunciation, belonging to one of the above-mentioned groups (for example, treating *ca* and *sa* as equal)

2. there are at least two characters associated with the component that have the same or similar pronunciation

In this way, a list of 787 phonetic components was created.

6.2 The initial setup

The setup for the experiment consisted of two 60x60 maps, that is, with 3600 cells on each map. This size is just right for an unambiguous representation of the 3437 characters in question. The first map contained the orthographic representation of the characters (the orthographic map), and the second one contained both the phonetic and the semantic representation (the semantophonetic map). This division was made under the assumption that the graphical form must be processed first, as this is the only information available to the reader (especially when the character is presented out of context and no guessing is possible), and the phonetic and semantic representation must be activated on the basis of the graphical form. This assumption was made to keep the initial model relatively simple; as we discussed in chapter 3, it is very likely that retrieval of the semantic form is a result of an interplay of a direct connection between the graphical form and semantics, and an indirect connection mediated by phonology.

The orthographic representation consisted of three parts. Two of the parts represented components: one for the semantic component and one for the phonetic component. The phonetic component was represented by the most common pronunciation associated with it. That is, phonetic components with a different graphical form, but associated with the same pronunciation got the same representation.

However, this component representation was not enough to differentiate the possible characters. Therefore, a representation of graphical complexity was added. It consisted of three numbers. The first number was the number of strokes that the character has. It is a simple measure of character complexity. A character such as 燃 (which has 16 strokes) is likely to be graphically confused with a character with similar number of strokes, but

a graphical confusion with a character with few strokes, such as 古, is not probable.

In the previous chapter we discussed cases of confusion caused of graphical similarity of one component. Consider the case of confusion of 伏 fú 'lie prostrate' with 扶 fú 'support with the hand'. They have no common components, yet the graphical similarity is a likely cause of confusion. The former character has 6 strokes, and the latter – 7. However, if we only consider the right-hand component, both have 4 strokes. We can see that considering the number of strokes of individual components may provide additional measures of similarity. Therefore, two values were added to the representation: the first one was the number of strokes in the semantic component (which in this case is 2 and 3, respectively) and the number of strokes in the remaining part of the character.

The semantophonetic map contained representation of meanings of the characters, combined with their pronunciation. The meanings were represented with the words that appeared in English definitions of the characters taken from *A Chinese-English Dictionary* (1995) – the assumption was that definitions of semantically related characters tend to use similar sets of words. The words had been stemmed in order to combine related word forms together, e.g. the words *assume*, *assumed* and *assuming* were all represented by the feature `assum`. Even after stemming, using every single word that appeared in the definitions would lead to an impractically high number of features. Therefore, the stop words (very frequent words with little semantic content, such as *a*, *the*) and very infrequent words were deleted from the list.

The specification of the DISLEX software package requires the features to be encoded as real numbers between 0 and 1. The representation of pronunciation was made using the Phonological Representation Database for Chinese Characters (Zhao & Li 2009), which maps each Chinese syllable into 15 real numbers, plus one more number for the tone. The same representation was used for phonetic components as they were also represented by their associated pronunciation. The semantic components were simply represented by a list of 122 numbers. The components were put in an arbitrary order and assigned numbers from 1 to 122. The first number on the list represented the presence of the first component (its value was 1 if it was present, and 0 if it was not). Most characters have only one semantic component, therefore the lists consisted mostly of zeroes, and usually did not have more than one 1. The advantage of such a sparse representation is that the arbitrary ordering of the semantic components does not have any influence on the computation of similarity, which would not be the case if we wanted to represent it, for example, with a single variable with 122 different numeric values.

6.3 Evaluation of the initial results

With the above setup, the DISLEX model was trained for 150 epochs, with each epoch consisting of self-organising of each of the two maps, and Hebbian

learning, which simulated the exposure to the characters (the concepts of self-organisation and Hebbian learning were introduced in section 3.7). Analysis of the resulting map showed that they have not changed in the last few epochs, which means that further training was unlikely to improve the representation.

The goal of the evaluation was to check whether it can predict which character pairs are likely to be confused. DISLEX is meant to provide a neurologically plausible model of reading, and therefore, if we find that confusable characters are represented close to each other on either of the two maps, the act of confusing one character for another may be explained by activation due to physical proximity.

The way the model was built was inspired by patterns found in patterns of character confusion in one learner. Since the features that have been used are quite general, it is reasonable to expect it to be applicable to more learners. Therefore, it should be ideally tested against pairs of confused characters gathered from other learners. However, such data are currently not available, and therefore the model was tested against the pairs of characters from the diary presented in the previous chapter. Even though this kind of evaluation is not ideal, it needs to be noted that the model was not directly trained using the pairs of confused characters, and they only served as a general inspiration to choose the features. Therefore, this evaluation may give an indication about the value of the model.

During the evaluation, the 1568 pairs of confused character from the diary were looked up on both maps, and the minimal distance between them was calculated. Then, 1568 character pairs were sampled from all possible combinations of characters. Histograms of the two resulting distributions are presented in Figure 6.1. As expected, we can see that among the confused pairs there are many more pairs that are close to each other than in the random sample, with the most of them within the distance of 2. However, apart from these closest pairs the two distributions look quite similar.

The evaluation of statistical significance of the difference between these distributions cannot be done with parametric statistical tests. We can see that in the case of the confused pairs the distribution is not normal. Moreover, in the evaluation we are comparing distances in a two-dimensional space, which makes the distances correlated. Therefore, the statistical significance was measured by a non-parametric two-tailed Monte Carlo permutation test, which does not require any particular statistical distribution of the data.

In order to perform the test, we assumed that the characters on the map could be permuted. That is, the location of each character representations would remain unchanged, but it would be assigned another, randomly chosen character. As a result of such a permutation, the map would contain the same 3437, but each of them could be in a position that was occupied by another character before. The null hypothesis is that our initial representation is a result of such a random permutation. If the null hypothesis was true, the mean distance between confused characters in a randomly permuted representation would sometimes have a value similar to the mean distances between confused characters in the initial representation.

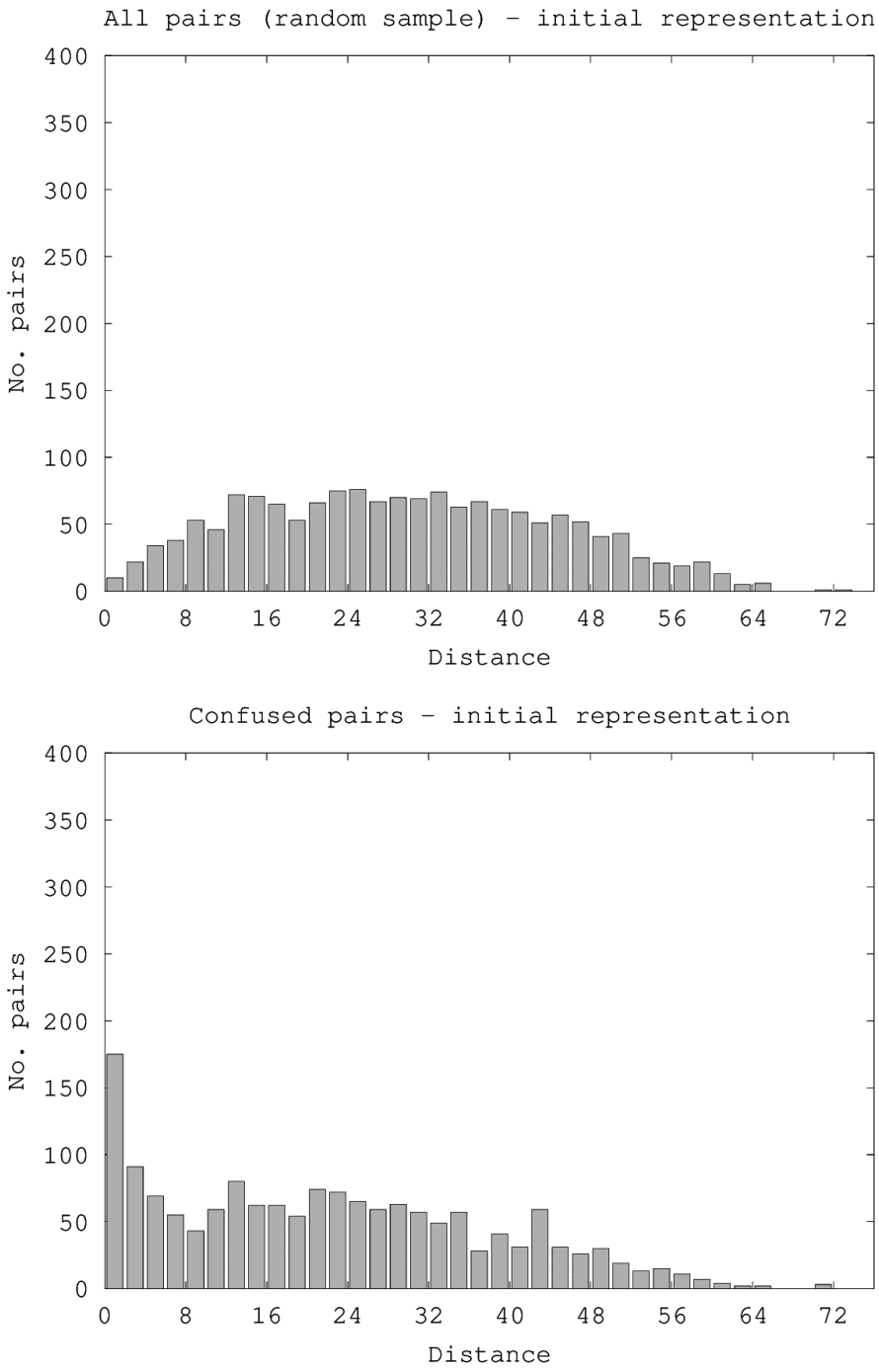

Figure 6.1: Results of the initial experiment

By repeating the random permutation and calculation of the mean distances we can increase the precision of the estimate of the p-value.

The random permutation was performed 10,000 times. The mean distance between the confused characters in the initial representation was 15.53, while the mean distances of confused characters in the random permutations were distributed in the range between 20.19 and 22.69 (with the average mean 21.50 and the standard deviation 0.31), with no value equal or lower than 15.53. This lets us reject the null hypothesis with $p < 0.0002$, and conclude that the average distance between the confused characters is lower than in the case of a representation made by a random permutation.

However, even though we can be reasonably sure that the positions of the characters in the maps were not random, it does not necessarily mean that the characters that are likely to be confused are clearly separated from the ones that are not likely to be confused. In an ideal representation two characters should be close to each other (within some arbitrarily chosen distance) if, and only if, the learner confuses these two characters for one another at some point. However, the data we have do not allow us to check this quantitatively. The character pairs that were used for evaluation are just a subset of potentially confusable pairs. If a pair of confused characters is represented as two distant points, we can regard is as an inaccuracy, because we know for a fact that the two characters have been confused for one another. This would be a clear case of a false negative (a pair of characters that are regarded as unconfusable by the model, which in fact have been confused). However, if two characters are represented by points that are close to each other, and this character pair is not on the list of confused pairs, we cannot know the reason. It may be a false positive (a pair of characters that are regarded as confusable by the model, which in fact are not confusable), but it may as well be a pair of confusable characters that did not happen to occur in the data. The list of confused characters can in no way be regarded as exhaustive.

Given all these issues, a manual inspection of the orthographic map was performed. It was possible to find pairs of characters that occupied the same place, but did not seem to be likely to be graphically confused, e.g. 门 (mén) 'gate, door' and 夕 (xī) 'evening'. The way the inspection was done is shown in Figure 6.2, which presents a 25x25 fragment of the map. The darker the cell is, the more characters is contains. The list on the left-hand side of the picture was obtained by clicking on the cell marked as 合. We can see a rather large number of characters, and the only thing they have is common is the semantic component 口 'mouth'. Several of these characters are associated with actions related to the mouth, e.g. 吞 (tūn) 'swallow, gulp down', 吻 (wěn) 'lip, kiss', but some of them are not, e.g. 右 (yòu) 'the right-hand side', 串 (chuàn) 'string together'. The 口 component has very different locations in different characters, so it is doubtful that its presence alone may cause two characters to be mistaken for one another. We can also see that due to the way the characters were represented, semantic components such as 扌 'hand' and 亻 'man' had a very large influence on how characters were grouped. Semantic components, such

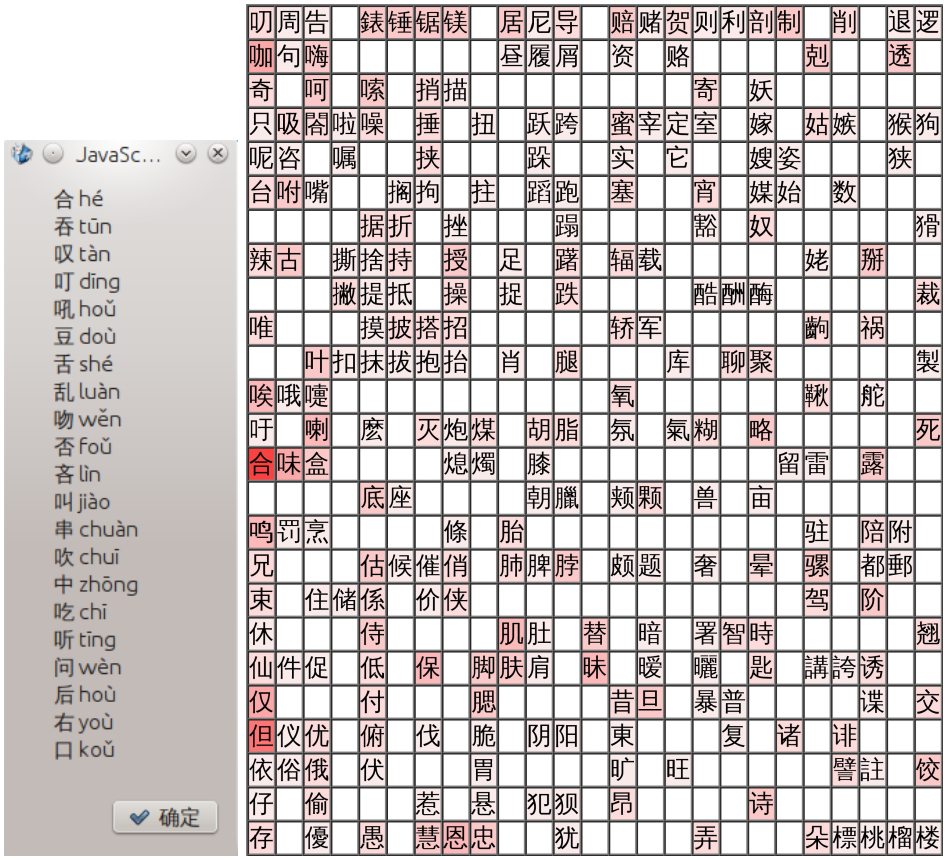

Figure 6.2: A 25x25 fragment of the orthographic map after 150 epochs in the initial experiment. All the characters located in the cell marked as 合 are shown on the left-hand side.

as 口, 扌 and 亻 are usually less visually salient than other components, so we may hypothesise that the model attached too large importance to them.

An inspection of the input vectors showed that they were identical for many pairs of very distinct characters, e.g. for 吞 (tūn) and 串 (chuàn), as well as for 门 (mén) and 夕 (xī). That is, the representation was too simple and DISLEX had no way to tell them apart. Similarly, in the case of semantophonetic map, many characters with no clear semantic or phonetic relation were represented in the same place, e.g. 汽 (qì) 'vapour, steam' and 狐 (hú) 'fox'. That was, again, caused by insufficient information that would tell them apart, as the words in their English definition were infrequent enough as not to be a part of the representation.

6.4 Improvement of the representation

The 1568 pairs of confused characters have been analysed as to whether I considered them to be semantically related. The resulting lists contained 312 pairs, assessed to have similar meanings. Such a binary classification of relatedness cannot be precise. The idea, however, was to have a subset of pairs of confused characters that were more semantically related than the rest, so the presence of some possibly unclear cases should not have a big effect on that average.

The list of the related pairs was then used to test the semantic representation. Since all the representations were in the end converted to numeric values, it was possible to compute the distances between meanings of characters. The average distance between meanings of semantically related characters was not significantly different from the average distance between meanings of characters that were assessed to be semantically unrelated. It was a clear sign that the semantic representation needed an improvement.

The first step to prepare an improved experimental setup was to revise the semantic representation. It turned out that the model based on co-occurrence of words did not work well, because the definitions were in many cases very short, and definitions of words that were assessed to be related did not necessarily have similar words. A common case was that two words had clearly different meaning, but their referents belonged to the same category, e.g. 鹊 (què) 'magpie' and 鹅 (é) 'goose' both indicate birds, while 脖 (bó) 'neck' and 膊 (bo) 'arm' both indicate parts of the body. Therefore, the WordNet database (Miller 1995) was used to represent words in the definitions with all their hypernyms, up to very general ones. In WordNet, *arm* has the following hypernym list:

arm => limb => extremity, appendage, member => external body part =>

body part => part, piece => thing => physical entity => entity

The hypernyms of *neck* are as follows:

neck, cervix => external body part => body part => part, piece =>

thing => physical entity => entity

In the old representation the meanings of the characters 脖[bó] 'neck' and 膊[bo] 'arm' had nothing in common, because their definitions had no shared words. In the new representation, however, the meanings of the two characters share several features: external body part, body part, part, piece, thing, physical entity and entity.

Also the graphical representation was revised. It turned out that representation of characters with the phonetic and semantic components, and the number of strokes did not take many important differences into account. There are often many hundreds of characters with the same number of strokes, and many of them do not have any specific phonetic or semantic component. Therefore, the old representation was not able to distinguish between such characters. In the new representation all the components that occur in more than one character were explicitly represented, and not just the ones that have a phonetic or a semantic value.

Finally, the representation of the pronunciation of the characters was moved from the semantic map to the orthographic map. The most principled approach would be to have a separate phonetic map, but the current version of DISLEX can only simultaneously train two maps, and the decision to represent phonology on one or on the other is to a large degree arbitrary. Merging the orthography and pronunciation on one map has the advantage that the representation of the pronunciation of the phonetic component of the character and the actual pronunciation of the character can be combined. In order to achieve that, the new model does not use the Phonological Representation Database for Chinese Characters, but rather makes a separate and independent feature of each initial and each final of the syllable (merging the similar initials according to Table 6.1). For example, in the old model the character 残[cán] had separate phonological representations of its actual pronunciation *cán* and of the pronunciation *jian* indicated by its phonetic component 戋[jiān] . In the new model, there is one phonological representation, and initials and finals of both pronunciations can be represented independently. In this way the new model can account for the cases mentioned in the previous chapter, where the phonetic component of one character is related to the pronunciation of the other character.

Introduction of each of the above features was preceded by testing whether it is likely to improve the model, that is, whether it makes the pairs of confused characters closer to each other, compared to the average distance between characters.

6.5 Results of the final experiment

In the final experiment, the DISLEX model with the new features was trained for 300 epochs, which consisted of self-organisation and Hebbian learning. The results were evaluated in different ways ways. Firstly, the random permutation significance test was performed. Secondly, the distance between 1568 pairs of confused characters on the resulting maps

were compared against the average distance between any two characters. It was done to check whether characters that are likely to be confused are represented close to each other on at least one of the two maps. Finally, for each of the 1568 pairs, the representation of the first character in a pair was activated in the graphical-phonetic map and the semantic map was checked to see what item is activated as a response. If the second character of the pair was activated, it means that the model reproduced that particular pattern of confusion.

The statistical significance of the result was checked with the Monte Carlo permutation test, using the same procedure as in the case of the initial representation. The mean distance between confused characters was 12.86, which is lower than in the initial model. The mean distances between confused characters in 10,000 randomly permuted representations were all between 21.81 and 24.30. The mean of these means was 23.10, which is more than in the initial model; the standard deviation was 0.34. Again, we can reject the null hypothesis with $p < 0.0002$.

A test was also performed to check whether the new model is significantly better than the previous one. Consider moving representation of each character in the initial model by a random number of steps, vertically and horizontally. The test checked how likely it is that the new model was such a random modification of the initial model. The checks were performed with the number of allowed steps in each directions ranging from 1 to 60. Each check consisted of generating and evaluating 1,000 such random models. In all these models, the difference between the mean distance between randomly chosen characters and the mean distance between confused characters was lower than 6.36, while in the new model this distance is greater than 8.95 ($21.81 - 12.86$). This lets us conclude that the new model is significantly better than the previous one with $p < 0.002$. The comparison of the distribution of distances between confused characters and distances between all character pairs is presented in Figure 6.3 and Table 6.2. Since the total number of character pairs is very large ($3437^2 = 11,812,969$), the percentages in the last column were based on a random sample of 1568 character pairs. We can see that about 37% of the character pairs that were confused for one another are within the distance of 4, and over 56% of them are within the distance of 12. For all the characters the respective percentages are 2.3% and 20%, respectively. We can conclude that the characters that are confused are likely to be represented close to each other on at least one of the two maps.

During the evaluation of the initial model we argued that there are aspects of the representation that are hard to check quantitatively. Therefore, a manual check of the new maps was performed. We cannot predict which characters are going to be confused, but unlike in the initial representation, it was hard to find character pairs that occupied the same place, but had nothing in common and seemed unlikely to be confused. This, together with the above quantitative results, gives an indication that the model has been improved in comparison to the initial one.

Figure 6.4 shows a fragment of the map and a list of characters occupying

	Confused pairs				All pairs (random sample)			
	Non-cumulative		Cumulative		Non-cumulative		Cumulative	
Distance range	Count	Percentage	Count	Percentage	Count	Percentage	Count	Percentage
0–4	578	36.86%	578	36.86%	36	2.30%	36	2.30%
4–8	170	10.84%	748	47.70%	114	7.27%	150	9.57%
8–12	135	8.61%	883	56.31%	161	10.27%	311	19.83%
12–16	124	7.91%	1007	64.22%	159	10.14%	470	29.97%
16–20	130	8.29%	1137	72.51%	195	12.44%	665	42.41%
20–24	100	6.38%	1237	78.89%	190	12.12%	855	54.53%
24–28	93	5.93%	1330	84.82%	177	11.29%	1032	65.82%
28–32	81	5.17%	1411	89.99%	135	8.61%	1167	74.43%
32–36	55	3.51%	1466	93.49%	136	8.67%	1303	83.10%
36–40	42	2.68%	1508	96.17%	88	5.61%	1391	88.71%
40–44	32	2.04%	1540	98.21%	81	5.17%	1472	93.88%
44–48	13	0.83%	1553	99.04%	48	3.06%	1520	96.94%
48–52	10	0.64%	1563	99.68%	21	1.34%	1541	98.28%
52–56	5	0.32%	1568	100.00%	19	1.21%	1560	99.49%
56–60					7	0.45%	1567	99.94%
60–64					1	0.06%	1568	100.00%

Table 6.2: Distances between pairs of confused characters compared to randomly sampled character pairs

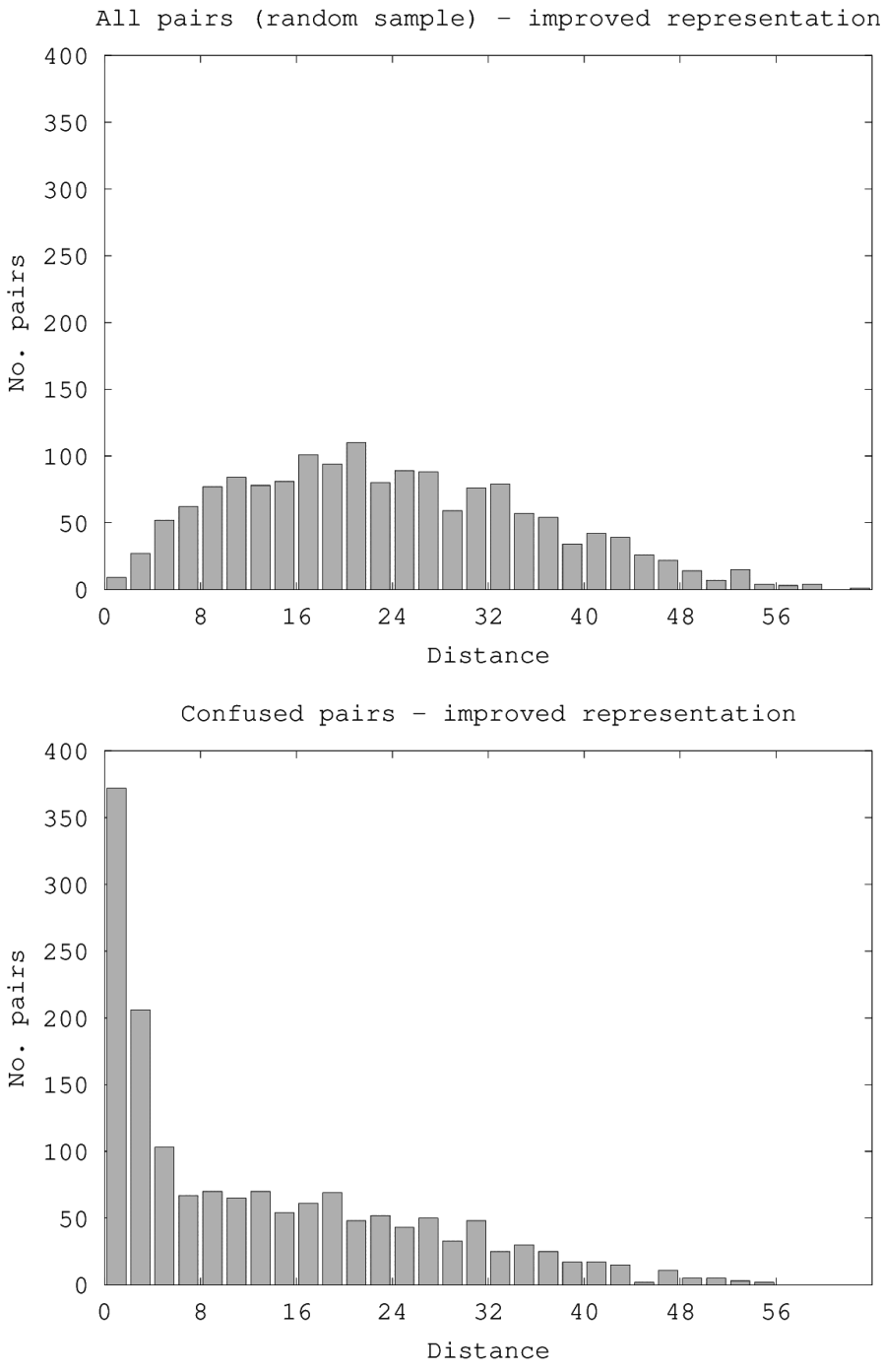

Figure 6.3: Results of the experiment with the improved representation

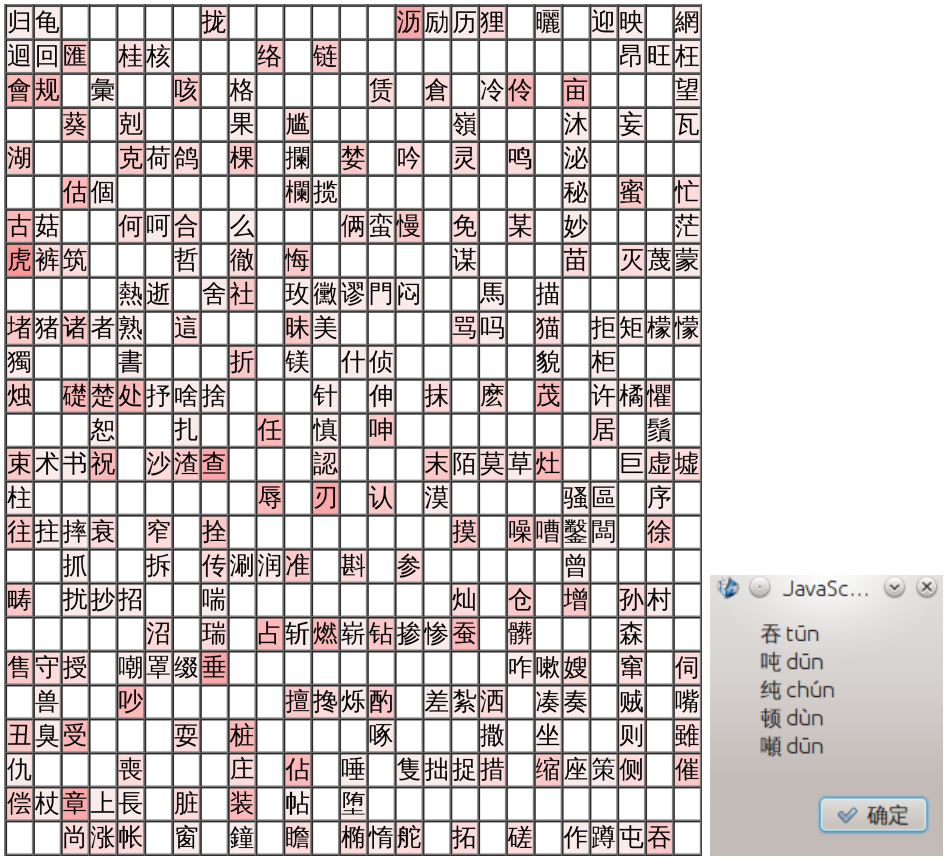

Figure 6.4: A 25x25 fragment of the orthographic-phonetic map after 300 epochs in the second experiment. All the characters located in the cell marked as 吞 are shown on the right-hand side.

the cell marked as 吞. We can see that the model predicts 吞 (tūn) to be confused with 吨 (dūn), 纯 (chún), 顿 (dùn) and 嘟 (dūn). The characters 吞 (tūn) and 吨 (dūn) not only have the common component 口, but perhaps more importantly, they have similar pronunciations. Therefore, we may expect them to be more likely to be confused than e.g. 吞 (tūn) and 串 (chuàn), which was predicted by the previous model. However, a definitive conclusion about the scope of applicability of the present model can be made only after testing it against character pairs confused by more learners.

The last type of evaluation was done after 20, 100, 200 and 300 epochs of training, and did not show the patterns of confusion higher than 5%. This is a very low result. Given that we can see that the confused characters are represented close to each other, it is likely that such a low result is related to the way the model was trained. The training of the model has several parameters that can be chosen relatively freely, such as the number of epochs, the relation between the epochs and the neighbourhood size and the order of self-organisation and Hebbian learning. Only a few combinations of such parameters have been tested. Moreover, the learning was evaluated by activating only one cell in one of the maps, and looking at the response on the other map. Perhaps a more detailed testing procedure that checks the results of activating more cells at once can show more patterns of confusion. We can conclude that the maps can be regarded as a reasonable approximation of the representation of the graphical, phonetic and semantic information related to the characters, but the process of learning and establishing connections between the two maps requires more research and testing.

Chapter 7

Conclusion

7.1 Summary of the thesis

This thesis investigated the problems related to second-language learning of Chinese characters. Chapter 1 introduced the problem by discussing why Chinese is generally regarded as a difficult language for Western learners, and the role Chinese characters play in this difficulty. Chapter 2 introduced important information about the Chinese writing system, the evolution and the structure of the Chinese characters, and investigated the number of characters a learner should be able to recognise to facilitate adequate reading comprehension. Chapter 3 introduced several models of reading, in particular different kinds of computational models, and discussed their advantages and disadvantages. The subsequent, practical part of the thesis built on the knowledge introduced in the first part, most importantly the structure of Chinese characters, the way they relate to each other, the number of characters advanced learners need to know and the ways to model the mental lexicon and the acquisition of reading.

Chapter 4 discussed results of a pilot study of character recognition, which motivated further research into character learning approaches. It focused on recognition-based and component-based approaches and posed the following research questions:

1. What characters are likely to be confused in the process of learning Chinese script?

2. If two or more characters are confused with one another, we may assume that they share some features or a combination of features. Which features cause character confusion?

3. What role do character components, meaning and pronunciation play in the representation of Chinese characters?

The first question was addressed in chapters 5 and 6. In chapter 5 it was argued that the data may be gathered in a self-observation diary study. The results of data gathering are presented in Appendix A. Chapter 6 presented the most important contribution of this thesis: a model of character acquisition based on general features of the characters, which was

able to account for much of the confusions found in the data, and gave examples of character pairs that, according to the model, are likely to be confused for one another.

The second question was addressed in chapter 5. The analysis of the data showed a variety of patterns that may cause confusion: graphical similarity, component similarity, phonetic similarity, semantic similarity. In some cases a similarity was likely to work across different aspects of the characters: a semantic component of one character might be related to the meaning of another character, and a phonetic component of one character might be related to the pronunciation of another character.

Chapter 6 addressed the third question. We could see that the initial representation of semantic components and the pronunciation of the phonetic components provided some significant results, which means that these features play a role in character confusion. However, there were several deficiencies in the representations, which were subsequently improved in the second version of the model. The most important improvements were:

- the introduction of hypernyms into the representation of meaning;
- the introduction of recurring components that did not have any clear semantic or phonetic value;
- the merging of the representation of pronunciation of the character with the pronunciation of the phonetic component;
- the merging of the initials in the phonetic representation by disregarding the manner of the articulation.

From the above we can conclude that meaning, pronunciation and graphical components indicate closeness of the representation of Chinese characters. The character meanings belonging to the same category, as indicated by shared hypernyms, can be used as an indicator of semantic relatedness. Apart from that, the representation of Chinese characters needs to take all recurring components into account, not just the ones that have a semantic or phonetic value. And finally, the actual pronunciation of the character and the pronunciation indicated by its phonetic component both seem to play a role in the assessment of a similarity, and they are likely to be represented together.

7.2 Future work

The work done in this thesis can be extended in several ways:

- **Evaluation on data from more learners**. This would allow to assess the scope of applicability of the model. In particular, a future study could test the predictions of the model on groups of learners using different learning methods: component- and recognition-based, as well as more traditional methods which focus on handwriting.

- **Tests of relative importance of the model's features**. The second version of the model performed better than the first one. However, more simulations need to be done to assess how each separate modification contributed to the overall improvement.
- **More detailed evaluation of Hebbian learning**. In the present experiments, learning was evaluated by activating only one cell in one of the maps, and looking at the response on the other map. More patterns of confusion may possibly be found if the testing procedure involves activating more cells at once.
- **Simulations with three maps**. Since each character has its semantic, phonetic and graphemic side, a model with three separate maps seems to be a good fit. It remains to be seen how interactions between three maps should exactly be modelled.

Appendix A

Confusable characters

The appendix presents characters that have been confused for one another, gathered in the self-observation diary study. The first two columns present the target character and its pronunciation. The next two columns present the character it was confused with and its pronunciation. The number of alternative pronunciations is limited to 2. The codes used in the subsequent columns present an approximate classification of the possible reason for the confusion. Apart from the semantic relatedness, the classification was performed automatically. The codes are as follows:

- `G`: graphical similarity, some shared components
- `SC`: the same semantic components
- `PC`: phonetic components with the same pronunciation
- `PCPR`: phonetic component of the first character indicates the pronunciation of the second character
- `PRON`: similar pronunciation
- `SEM`: semantic similarity (annotated manually)

哎	āi	艾	ài/yì	G		PC	PCPR	PRON	
唉	āi/ài	唆	suō	G	SC				
騃	ái	挨	āi	G		PC	PCPR	PRON	
矮	ǎi	短	duǎn	G					SEM
藹	ǎi	蕴	yùn	G	SC				
暧	ài/nuǎn	暖	nuǎn	G	SC			PRON	
爱	ài	受	shòu	G					
碍	ài	确	què	G	SC				
嚶	ài	凝	níng	G					
隘	ài	碍	ài					PRON	
昂	áng	印	yìn	G					
熬	áo	傲	ào	G		PC	PCPR	PRON	
袄	ǎo	沃	wò	G		PC			
袄	ǎo	妖	yāo	G		PC			
傲	ào	傍	bàng	G	SC				
傲	ào	骄	jiāo						SEM
叭	ba	琶	pá					PRON	SEM
捌	bā	拔	bá/ba	G	SC			PRON	
捌	bā	彆	biè						
壩	bà	霸	bà	G		PC	PCPR	PRON	

壩	bà	壇	tán	G	SC				
罷	ba/bà	態	tài	G					
霸	bà	壩	bà	G		PC	PCPR	PRON	
霸	bà	霜	shuāng	G	SC				
掰	bāi	扮	ban/bàn	G					
掰	bāi	斑	bān	G					
掰	bāi	颁	bān	G					
掰	bāi	瓣	bàn	G					
掰	bāi	辩	biàn	G					
般	bān	搬	bān	G		PC	PCPR	PRON	
般	bān	船	chuán	G	SC				
颁	bān	烦	fán/fan	G	SC			PRON	
瓣	bàn	贬	biǎn				PCPR		
瓣	bàn	辨	biàn	G		PC	PCPR		
瓣	bàn	辩	biàn	G		PC	PCPR		
瓣	bàn	辨	biàn	G		PC	PCPR		
瓣	bàn	孤	gū	G		PC	PCPR		
瓣	bàn	辜	gū	G		PC	PCPR		
幇	bāng	封	fēng	G					
榜	bǎng	棒	bàng	G	SC		PCPR	PRON	
绑	bǎng	榜	bǎng			PC	PCPR	PRON	
膀	bǎng	肪	fáng	G	SC	PC	PCPR	PRON	
傍	bàng	榜	bǎng	G		PC	PCPR	PRON	
镑	bàng	磅	bàng	G		PC	PCPR	PRON	SEM
胞	bāo	胎	tāi	G	SC				SEM
爆	bào	炮	pào	G	SC	PC	PCPR	PRON	SEM
卑	bēi	脾	pí	G		PC			
備	bèi	倍	bèi	G	SC			PRON	
備	bèi	佣	yōng/yòng	G	SC				
備	bèi	傭	yōng	G	SC				
备	bèi	各	gè	G	SC				
惫	bèi	备	bèi	G	SC			PRON	
惫	bèi	狈	bèi					PRON	SEM
惫	bèi	辈	bèi					PRON	
辈	bèi	靠	kào	G		PC			
蹦	bèng	崩	bēng	G		PC	PCPR	PRON	
迸	bèng	崩	bēng					PRON	
迸	bèng	蹦	bèng					PRON	
彼	bǐ	此	cǐ						SEM
壁	bì	垫	diàn	G	SC				
壁	bì	辟	pì/bì	G		PC	PCPR	PRON	
幣	bì	弊	bì	G		PC		PRON	
幣	bì	斃	bì	G		PC		PRON	
庇	bì	蔽	bì				PCPR	PRON	SEM
弊	bì	斃	bì	G		PC		PRON	
弊	bì	毙	bì					PRON	
畢	bì	華	huá	G					
畢	bì	畫	huà	G	SC				
痹	bì	鼻	bí	G		PC	PCPR	PRON	
痹	bì	疫	yì	G	SC				SEM
碧	bì	鄙	bǐ	G				PRON	
碧	bì	簸	bǒ						
碧	bì	翠	cuì						SEM
臂	bì/bei	肩	jiān	G	SC				SEM
臂	bì/bei	譬	pì	G		PC	PCPR	PRON	
蔽	bì	憋	biē	G		PC	PCPR		
蔽	bì	弊	bì	G	SC	PC		PRON	
蔽	bì	繭	jiǎn	G	SC				
闭	bì	逼	bī					PRON	
闭	bì	必	bì					PRON	
扁	biǎn	偏	piān	G		PC	PCPR	PRON	
辩	biàn	辨	biàn	G		PC	PCPR	PRON	
标	biāo	表	biǎo	G				PRON	
飙	biāo	飘	piāo	G	SC	PC		PRON	
憋	biē	弊	biè	G		PC	PCPR	PRON	
憋	biē	逼	bī						SEM
弊	biè	憋	biē	G		PC	PCPR	PRON	

濒	bīn	滨	bīn	G	SC	PC	PCPR	PRON	SEM
賓	bīn	賣	mài/mai	G					
賓	bīn	實	shí/shi	G	SC				
秉	bǐng	兼	jiān	G					
剥	bō	剂	jì	G	SC				
剥	bō	碌	lù	G		PC	PCPR		
剥	bō	刨	páo	G	SC				
剥	bō	削	xuē/xiāo	G	SC				SEM
拨	bō	拔	bá/ba	G	SC				SEM
搏	bó	膊	bo	G		PC	PCPR	PRON	
脖	bó	膊	bo	G	SC	PC	PCPR	PRON	SEM
蔔	bo	補	bǔ			PC	PCPR		
蔔	bo	菊	jú	G	SC				
蔔	bo	葡	pú	G	SC	PC	PCPR		
蔔	bo	萄	táo	G	SC				
哺	bǔ	補	bǔ	G		PC	PCPR	PRON	SEM
捕	bǔ	补	bǔ	G			PCPR	PRON	
捕	bǔ	補	bǔ	G		PC	PCPR	PRON	
捕	bǔ	撲	pū	G	SC	PC	PCPR	PRON	SEM
埠	bù	壇	tán	G	SC				
纔	cái	纏	chán	G	SC	PC	PCPR		
裁	cái	錶	biǎo	G					
睬	cǎi	眯	mī	G	SC				
惭	cán	慨	kǎi	G	SC				
残	cán	惭	cán					PRON	SEM
残	cán	惨	cǎn					PRON	SEM
残	cán	歹	dǎi	G	SC				
残	cán	歼	jiān	G	SC		PCPR		
蚕	cán	吞	tūn	G					
惨	cǎn	惭	cán	G	SC		PCPR	PRON	SEM
惨	cǎn	残	cán				PCPR	PRON	SEM
惨	cǎn	深	shēn	G					
惨	cǎn	慎	shèn	G	SC				
掺	càn/chān	残	cán	G				PRON	
掺	càn/chān	惨	cǎn	G				PRON	
倉	cāng	艙	cāng	G		PC		PRON	SEM
沧	cāng	汪	wāng	G	SC				SEM
苍	cāng	仓	cāng	G		PC	PCPR	PRON	
嘈	cáo	噪	zào	G	SC	PC	PCPR	PRON	SEM
槽	cáo	糟	zāo	G		PC	PCPR	PRON	
层	céng	曾	céng					PRON	
層	céng	曾	céng	G		PC	PCPR	PRON	
層	céng	屑	xiè	G	SC				
察	chá	餐	cān	G					
察	chá	查	chá					PRON	SEM
察	chá	奈	nài	G					
诧	chà	差	chà/chā					PRON	
诧	chà	剎	shā/chà					PRON	
诧	chà	驼	tuo/tuó	G					
诧	chà	托	tuō	G					
诧	chà	异	yì						
诧	chà	宅	zhái	G					
拆	chāi	驳	bó						
拆	chāi	斥	chì	G					
拆	chāi	诉	su/sù	G					
柴	chái	紫	zǐ	G		PC	PCPR		
搀	chān	掺	càn/chān	G	SC	PC	PCPR	PRON	SEM
缠	chán	厘	lí	G					
讒	chán	護	hù	G	SC				
谗	chán	馋	chán	G		PC	PCPR	PRON	
饞	chán	纔	cái	G		PC			
饞	chán	餐	cān						
饞	chán	残	cán						
颤	chàn	擅	shàn	G		PC	PCPR	PRON	
昌	chāng	晶	jīng	G	SC				
猖	chāng	昌	chāng	G		PC	PCPR	PRON	
偿	cháng	尝	cháng	G		PC	PCPR	PRON	
償	cháng	嘗	cháng	G		PC	PCPR	PRON	

偿	cháng	價	jià	G	SC				
廠	chǎng	场	chǎng/cháng				PCPR	PRON	
廠	chǎng	産	chǎn	G					SEM
廠	chǎng	廣	guǎng	G	SC				
敞	chang	般	bān	G					
敞	chang	畅	chàng				PCPR	PRON	SEM
倡	chàng	昌	chāng	G		PC	PCPR	PRON	
倡	chàng	猖	chāng	G		PC	PCPR	PRON	
畅	chàng	敞	chang					PRON	
潮	cháo	朝	cháo/zhāo	G		PC	PCPR	PRON	
潮	cháo	湖	hú	G	SC				SEM
吵	chǎo	悄	qiāo						SEM
炒	chǎo	炊	chuī	G	SC				SEM
炒	chǎo	烧	shāo	G	SC			PRON	SEM
扯	chě/che	彻	chè					PRON	
扯	chě/che	耻	chǐ	G					
彻	chè	扯	chě/che					PRON	
彻	chè	撤	chè					PRON	
彻	chè	切	qiè/qiē	G		PC	PCPR		SEM
彻	chè	锲	qiè			PC	PCPR		SEM
徹	chè	辙	zhé	G		PC	PCPR	PRON	
撤	chè	扯	chě/che	G	SC		PCPR	PRON	
撤	chè	彻	chè				PCPR	PRON	
撤	chè	撒	sā/sǎ	G	SC				
撤	chè	辙	zhé	G		PC	PCPR	PRON	SEM
撤	chè	掷	zhì	G	SC				
澈	chè	渐	jiàn	G	SC				
澈	chè	缴	jiǎo	G					
塵	chén	慶	qìng	G					
尘	chén	尖	jiān	G	SC				
晨	chen/chén	辰	chén	G		PC	PCPR	PRON	
衬	chèn	趁	chèn					PRON	
衬	chèn	裹	guǒ						
衬	chèn	讨	tǎo	G					
称	chēng/chèn	秤	chèng	G	SC			PRON	
撑	chēng/cheng	掌	zhǎng/zhang	G		PC			SEM
乘	chéng	乖	guāi	G		PC			
承	chéng	乘	chéng					PRON	
程	chéng	乘	chéng	G	SC	PC	PCPR	PRON	SEM
遲	chí	违	wéi	G	SC				
充	chōng	允	yǔn	G	SC				
崇	chóng	耸	sǒng				PCPR		SEM
稠	chóu	绸	chóu	G		PC	PCPR	PRON	
稠	chóu	调	diào/tiáo	G					
稠	chóu	稚	zhì	G	SC				
筹	chóu	畴	chóu	G		PC	PCPR	PRON	
绸	chóu	稠	chóu	G		PC	PCPR	PRON	
绸	chóu	绵	mián	G	SC				SEM
踌	chóu	躇	chú	G	SC				
醜	chǒu	魂	hún	G	SC				
醜	chǒu	酿	niàng	G	SC				
醜	chǒu	酝	yùn	G	SC				
躇	chú	踌	chóu	G	SC				SEM
畜	chù/xù	蓄	xù	G		PC	PCPR	PRON	
幢	chuáng/zhuàng	幅	fú	G	SC				
炊	chuī	欢	huan/huān	G					
炊	chuī	欣	xīn	G					
脣	chún	辱	rǔ	G		PC			
蠢	chǔn	蠶	cán	G	SC				
慈	cí	滋	zī	G		PC	PCPR	PRON	
瓷	cí	磁	cí			PC	PCPR	PRON	SEM
辞	cí	括	kuò	G					
辞	cí	锡	xī						
雌	cí	慈	cí			PC	PCPR	PRON	SEM
凑	còu	凄	qī	G	SC				
促	cù	粗	cū				PCPR	PRON	
窜	cuàn	患	huàn	G					SEM
竄	cuàn	窟	kū	G	SC				

摧	cuī	催	cuī	G		PC	PCPR	PRON	
粹	cuì	精	jīng	G	SC				SEM
翠	cuì	羿	yì	G	SC				
翠	cuì	翼	yì	G	SC				
脆	cuì	危	wēi	G		PC	PCPR		
搓	cuō	拆	chāi	G	SC				
磋	cuō	搓	cuō	G		PC	PCPR	PRON	
挫	cuò	措	cuò	G	SC		PCPR	PRON	
逮	dǎi/dài	代	dài					PRON	
耽	dān	审	shěn						
担	dān/dàn	胆	dǎn	G				PRON	
黨	dǎng	熏	xūn	G					
悼	dào	担	dān/dàn						
悼	dào	惮	dàn	G	SC				
悼	dào	惦	diàn	G	SC				SEM
瞪	dèng	盯	dīng	G	SC				SEM
瞪	dèng	瞻	zhān	G	SC				SEM
涤	dí	滴	dī	G	SC			PRON	SEM
涤	dí	潦	lǎo/liǎo	G	SC				
弟	dì/di	第	dì	G		PC	PCPR	PRON	
缔	dì	蒂	dì	G		PC	PCPR	PRON	
缔	dì	绪	xù	G	SC				
垫	diàn	坚	jiān	G	SC				
抖	dǒu	斗	dòu	G				PRON	
妒	dù	嫉	jí	G	SC				SEM
杜	dù	灶	zào	G	SC				
断	duàn	继	jì	G					
盾	dùn	直	zhí	G	SC				
堕	duò/huī	降	jiàng/xiáng	G					SEM
堕	duò/huī	墅	shù	G	SC				
掠	è	虐	nvè						SEM
贰	èr	腻	nì	G					
伐	fá	仪	yí	G	SC				
阀	fá	阁	gé	G					
犯	fàn	范	fàn	G				PRON	
妨	fáng/fāng	防	fáng	G		PC	PCPR	PRON	SEM
份	fèn/fen	扮	ban/bàn	G					
锋	fēng	降	jiàng/xiáng	G					
缝	fèng/féng	链	liàn	G					
鳳	fèng	凰	huáng	G					SEM
伏	fú	仆	pū/pú	G	SC		PCPR	PRON	SEM
俘	fú	仔	zǐ	G	SC				
扶	fú	扑	pū	G	SC			PRON	
服	fú/fu	肤	fū	G	SC			PRON	
辐	fú	幅	fú	G		PC	PCPR	PRON	
俯	fǔ	府	fǔ	G		PC	PCPR	PRON	
抚	fǔ	辅	fǔ					PRON	
辅	fǔ	辐	fú	G	SC	PC	PCPR	PRON	
辅	fǔ	抚	fǔ				PCPR	PRON	
辅	fǔ	辖	xiá	G	SC				
辅	fǔ	辙	zhé	G	SC				
妇	fù/fu	嫂	sǎo/sao	G	SC				SEM
竿	gān	秆	gǎn	G		PC	PCPR	PRON	SEM
纲	gāng	刚	gāng	G		PC	PCPR	PRON	
稿	gǎo	搞	gǎo	G		PC	PCPR	PRON	
搁	gē/ge	割	gē	G			PCPR	PRON	
格	gé/ge	阁	gé	G		PC	PCPR	PRON	
阁	gé	格	gé/ge	G		PC	PCPR	PRON	
阁	gé	阂	hé	G		PC	PCPR	PRON	
隔	gé	隘	ài	G	SC				
隔	gé	逼	bī	G					
隔	gé	割	gē	G				PRON	SEM
革	gé	韧	rèn	G					
攻	gōng	公	gōng					PRON	
躬	gōng	射	shè	G	SC				SEM
股	gǔ/gu	穀	gǔ	G				PRON	
谷	gǔ	容	róng	G					

僱	gù	顧	gù	G		PC	PCPR	PRON	
雇	gù	肩	jiān	G	SC				
雇	gù	截	jié	G	SC				
顧	gù	僱	gù	G		PC	PCPR	PRON	
顾	gù	蹲	dūn						
惯	guàn	怪	guài	G	SC				
龟	guī	乌	wū						SEM
還	hái/huán	遠	yuǎn	G	SC				
函	hán	涵	hán	G		PC	PCPR	PRON	
焊	hàn	旱	hàn	G		PC	PCPR	PRON	
浩	hào	耗	hào				PCPR	PRON	
核	hé	硋	ài	G		PC			
狠	hěn	恨	hèn	G	SC	PC	PCPR	PRON	SEM
衡	héng	橫	héng					PRON	
哄	hōng/hǒng	轰	hōng				PCPR	PRON	SEM
轟	hōng	蟲	chóng	G					
厚	hòu	厦	shà	G					
瑚	hú	蝴	hú	G		PC	PCPR	PRON	
葫	hú	菇	gu	G	SC	PC	PCPR	PRON	SEM
葫	hú	萌	méng	G	SC				SEM
畫	huà	書	shū	G					
畫	huà	晝	zhòu	G					
缓	huǎn	援	yuán	G					
缓	huǎn	缘	yuán	G	SC				
痪	huàn	患	huàn				PCPR	PRON	
痪	huàn	疫	yì	G	SC				SEM
慌	huāng/huang	荒	huāng	G		PC	PCPR	PRON	
恍	huǎng	缓	huǎn						
晃	huang/huàng	恍	huǎng	G	SC	PC	PCPR	PRON	
谎	huǎng	荒	huāng	G		PC	PCPR	PRON	
恢	huī	挥	huī			PC	PCPR	PRON	
悔	huǐ	侮	wǔ	G		PC			
悔	huǐ	悟	wù	G	SC				
毁	huǐ	悔	huǐ					PRON	
惠	huì	慧	huì	G	SC			PRON	
慧	huì	惠	huì	G	SC			PRON	
慧	huì	讳	huì					PRON	
秽	huì	穗	suì	G	SC				
贿	huì	讳	huì					PRON	
贿	huì	赂	lù	G	SC				SEM
贿	huì	豫	yù						
贿	huì	郁	yù	G					
昏	hūn/hún	晨	chen/chén	G	SC				SEM
昏	hūn/hún	混	hùn/hún	G		PC	PCPR	PRON	
昏	hūn/hún	婚	hūn	G		PC	PCPR	PRON	
浑	hún	挥	huī	G		PC	PCPR		
豁	huō	畅	chàng						
惑	huò/huo	感	gǎn	G	SC				
惑	huò/huo	或	huò	G		PC	PCPR	PRON	
惑	huò/huo	霍	huò				PCPR	PRON	
獲	huò	独	dú	G	SC				
獲	huò	獨	dú	G	SC				
獲	huò	護	hù	G		PC			
獲	huò	狂	kuáng	G	SC				
獲	huò	猛	měng	G	SC				
祸	huò	锅	guō	G		PC	PCPR	PRON	
祸	huò	货	huò	G			PCPR	PRON	
祸	huò	霍	huò				PCPR	PRON	
穫	huò	積	jī	G	SC				
穫	huò	穗	suì	G	SC				
穫	huò	稀	xī	G	SC				
霍	huò	祸	huò					PRON	
霍	huò	震	zhèn	G	SC				
圾	jī	积	jī	G			PCPR	PRON	
圾	jī	极	jí	G		PC	PCPR	PRON	
積	jī	蹟	jī	G		PC	PCPR	PRON	
绩	jī	择	zé						

嫉	jí	嫌	xián	G	SC				SEM
棘	jí	刺	cì	G					
级	jí	纪	jì	G	SC	PC	PCPR	PRON	SEM
辑	ji/jí	辐	fú	G	SC				
寂	jì	椒	jiāo	G		PC			
寂	jì	宿	sù/xiǔ	G	SC				
技	jì	支	zhī	G		PC	PCPR		
技	jì	肢	zhī	G		PC	PCPR		
技	jì	执	zhí	G	SC	PC	PCPR		
纪	jì	级	jí	G	SC	PC	PCPR	PRON	SEM
嘉	jiā	辜	gū	G	SC				
夹	jiā/jiá	甩	shuǎi						
夹	jiā/jiá	爽	shuǎng	G	SC				
歼	jiān	奸	jiān	G				PRON	
殲	jiān	歹	dǎi	G	SC				SEM
溅	jiàn	渐	jiàn	G	SC		PCPR	PRON	
溅	jiàn	浅	qiǎn	G	SC	PC	PCPR	PRON	
煎	jiān	焦	jiāo	G	SC				SEM
煎	jiān	烹	pēng	G	SC				SEM
煎	jiān	煮	zhǔ	G	SC				SEM
肩	jiān	启	qǐ	G	SC				
肩	jiān	胃	wèi	G	SC				SEM
柬	jiǎn	帖	tiē/tiě						SEM
检	jiǎn	俭	jiǎn	G		PC	PCPR	PRON	
繭	jiǎn	蠶	cán						SEM
渐	jiàn	溅	jiàn	G	SC			PRON	
鉴	jiàn	剑	jiàn	G				PRON	
键	jiàn	鉴	jiàn		SC		PCPR	PRON	
键	jiàn	炼	liàn						
键	jiàn	链	liàn	G	SC				
键	jiàn	铸	zhù	G	SC				
僵	jiāng	疆	jiāng	G		PC	PCPR	PRON	
僵	jiāng	薑	jiāng	G		PC	PCPR	PRON	
将	jiāng/jiàng	浮	fú	G					
浆	jiāng	酱	jiàng	G		PC	PCPR	PRON	SEM
疆	jiāng	僵	jiāng	G		PC	PCPR	PRON	
薑	jiāng	董	dǒng	G	SC				
薑	jiāng	监	jiān	G					
薑	jiāng	監	jiān	G					
薑	jiāng	检	jiǎn						
奖	jiǎng	將	jiāng/jiàng	G		PC	PCPR	PRON	
桨	jiǎng	奖	jiǎng	G		PC	PCPR	PRON	
娇	jiāo	育	yù						
教	jiào/jiāo	校	xiào	G		PC	PCPR	PRON	
浇	jiāo	饶	ráo	G		PC			
浇	jiāo	洗	xǐ	G	SC				SEM
膠	jiāo	谬	miù	G		PC			
蕉	jiāo	薦	jiàn	G	SC				
骄	jiāo	轿	jiào	G		PC	PCPR	PRON	
骄	jiāo	桥	qiáo	G		PC	PCPR	PRON	
绞	jiǎo	缴	jiǎo	G	SC		PCPR	PRON	
缴	jiǎo	绑	bǎng	G	SC				
缴	jiǎo	绩	jī	G	SC				
缴	jiǎo	绕	rào/rǎo	G	SC				
缴	jiǎo	邀	yāo	G					
轿	jiào	较	jiào	G	SC	PC	PCPR	PRON	
揭	jiē	遏	è	G		PC			
節	jié	结	jié/jiē					PRON	SEM
節	jié	筋	jīn	G	SC				SEM
階	jiē	皆	jiē	G		PC	PCPR	PRON	
傑	jié	價	jià	G	SC				
傑	jié	節	jié					PRON	
截	jié	裁	cái	G		PC	PCPR		SEM
截	jié	戴	dài	G		PC			
截	jié	载	zài/zǎi	G		PC	PCPR		
捷	jié	逮	dǎi/dài	G					
捷	jié	堤	dī	G					
捷	jié	截	jié	G				PRON	

捷	jié	提	tí	G	SC				
洁	jié	吉	jí	G		PC			
潔	jié	傑	jié				PCPR	PRON	
介	jiè	价	jià	G		PC			
戒	jiè	截	jié	G			PCPR	PRON	SEM
戒	jiè	诫	jiè	G		PC	PCPR	PRON	SEM
斤	jīn	切	qiè/qiē	G					
儘	jǐn	僅	jǐn	G	SC	PC	PCPR	PRON	
盡	jǐn/jìn	僅	jǐn	G		PC	PCPR	PRON	SEM
紧	jin	僅	jǐn					PRON	
紧	jǐn	繫	xì	G	SC				
谨	jǐn	僅	jǐn	G		PC	PCPR	PRON	
锦	jǐn	吊	diào	G					
锦	jǐn	掉	diào	G					
晋	jìn	普	pǔ	G	SC				
近	jìn	进	jìn	G	SC			PRON	
颈	jǐng	领	lǐng	G	SC				SEM
兢	jīng	競	jìng	G				PRON	
驚	jīng	警	jǐng	G		PC	PCPR	PRON	
警	jǐng	譬	pì	G	SC				
警	jǐng	誓	shì	G	SC				
警	jǐng	誉	yù	G	SC				
敬	jìng	静	jìng/jing			PC	PCPR	PRON	
敬	jìng	净	jìng				PCPR	PRON	
敬	jìng	竟	jìng			PC	PCPR	PRON	
竟	jìng	竞	jìng	G	SC	PC	PCPR	PRON	
競	jìng	兢	jīng	G			PCPR	PRON	
静	jìng/jing	净	jìng	G		PC	PCPR	PRON	
揪	jiū	救	jiù	G		PC	PCPR	PRON	
揪	jiū	鍬	qiāo	G		PC			
灸	jiǔ	炎	yán	G	SC				SEM
玖	jiǔ	灸	jiǔ	G		PC	PCPR	PRON	
桔	jú	枯	kū	G	SC				
桔	jú	苦	kǔ	G					
桔	jú	酷	kù	G					
俱	jù	具	jù/ju	G				PRON	
拘	jū	苟	gǒu	G			PCPR		
拘	jū	揪	jiū	G	SC				
拘	jū	鞠	jū	G				PRON	
鞠	jū	菊	jú	G		PC	PCPR	PRON	
咀	jǔ	嚼	jué	G	SC				SEM
剧	jù	據	jù	G		PC	PCPR	PRON	
懼	jù	櫻	yīng	G					
拒	jù	鉅	jù	G		PC	PCPR	PRON	
距	jù	轨	guǐ						
鉅	jù	距	jù	G		PC	PCPR	PRON	
捐	juān	肩	jiān	G					
捐	juān	损	sǔn	G	SC				SEM
卷	juǎn/juàn	捐	juān			PC	PCPR	PRON	
倦	juàn	卷	juǎn/juàn	G		PC	PCPR	PRON	
倦	juàn	捐	juān			PC	PCPR	PRON	
倦	juàn	捲	juǎn	G		PC	PCPR	PRON	
倦	juàn	俊	jùn	G	SC				
嚼	jué	囊	nāng/náng						
掘	jué	绝	jué					PRON	
掘	jué	抓	zhuā	G	SC				
绝	jué	截	jié						SEM
绝	jué	决	jué					PRON	
绝	jué	拒	jù						SEM
俊	jùn	峻	jùn	G		PC	PCPR	PRON	
俊	jùn	傻	shǎ	G	SC	PC			
俊	jùn	唆	suō	G		PC			
凯	kǎi	慨	kǎi					PRON	
慨	kǎi	概	gài	G		PC	PCPR	PRON	
勘	kān	劫	jié	G	SC				
勘	kān	截	jié						
勘	kān	刊	kān			PC	PCPR	PRON	SEM
勘	kān	堪	kān	G		PC	PCPR	PRON	

勘	kān	勸	quàn	G	SC				
堪	kān	甚	shén/shèn	G		PC			
康	kāng	厘	lí	G					
康	kāng	隶	lì	G		PC			
慷	kāng	惦	diàn	G	SC				
抗	kàng	航	háng	G		PC	PCPR	PRON	
抗	kàng	坑	kēng	G		PC			
磕	kē	罢	ba/bà	G					
磕	kē	摆	bǎi	G					
科	kē	料	liào	G					
颗	kē	棵	kē	G		PC	PCPR	PRON	
殼	ké/qiào	穀	gǔ	G					
克	kè	剋	kè/kēi	G		PC	PCPR	PRON	
克	kè	刻	kè			PC	PCPR	PRON	
肯	kěn	恳	kěn			PC	PCPR	PRON	
坑	kēng	炕	kàng	G		PC	PCPR		
抠	kōu	寇	kòu					PRON	
寇	kòu	盗	dào						SEM
寇	kòu	冠	guān/guàn	G		PC			
扣	kòu	寇	kòu					PRON	
窟	kū	窜	cuàn	G	SC				
裤	kù	库	kù	G		PC	PCPR	PRON	
垮	kuǎ	挎	kuà	G		PC	PCPR	PRON	
垮	kuǎ	墟	xū	G	SC				SEM
挎	kuà	垮	kuǎ	G		PC	PCPR	PRON	
款	kuǎn	砍	kǎn	G					
款	kuǎn	亲	qīn/qin						
款	kuǎn	欺	qī	G	SC				
筐	kuāng	框	kuāng/kuàng	G		PC	PCPR	PRON	
愧	kuì	惭	cán	G	SC				SEM
廓	kuò	廠	chǎng	G	SC				
廓	kuò	廊	láng	G	SC				
廓	kuò	轮	lún						SEM
廓	kuò	庭	tíng	G	SC				
扩	kuò	括	kuò	G	SC			PRON	SEM
扩	kuò	阔	kuò					PRON	
括	kuò	拉	lā/la	G	SC				
括	kuò	舌	shé	G					
括	kuò	折	zhé/zhē	G	SC				
腊	là/xī	膝	xī	G	SC			PRON	
臘	là	蠟	là	G		PC	PCPR	PRON	
臘	là	腻	nì	G	SC				
蜡	là	烛	zhú	G	SC				SEM
蠟	là	青	qīng						
蠟	là	鼠	shǔ	G					
啦	la/lā	喇	lǎ	G	SC		PCPR	PRON	
赖	lài	敞	chang	G	SC				
赖	lài	豁	huō						
婪	lán	禁	jìn/jīn	G	SC				
斓	lán	欄	lán	G		PC	PCPR	PRON	
篮	lán	蓝	lán	G		PC	PCPR	PRON	
揽	lǎn	援	yuán	G	SC				SEM
滥	làn	涝	lào	G	SC				SEM
廊	láng	堂	táng/tang						SEM
廊	láng	唐	táng	G	SC				
廊	láng	亭	tíng						SEM
廊	láng	庭	tíng	G	SC				SEM
狼	láng	狠	hěn	G	SC				
朗	lǎng	郎	láng/làng	G			PCPR	PRON	
浪	làng	郎	láng/làng	G				PRON	
浪	làng	朗	lǎng	G				PRON	
捞	lāo	涝	lào	G		PC	PCPR	PRON	SEM
唠	láo/lào	啰	luō	G	SC				SEM
涝	lào	捞	lāo	G		PC	PCPR	PRON	SEM
牢	láo	牵	qiān	G	SC				
潦	lǎo/liǎo	涝	lào	G	SC			PRON	
樂	lè/yuè	葉	yè	G					
勒	lēi	革	gé	G	SC				

勒	lēi	韧	rèn	G					
累	léi/lěi	壘	lěi	G		PC	PCPR	PRON	SEM
壘	lěi	累	léi/lěi	G		PC	PCPR	PRON	SEM
淚	lèi	妒	dù	G					
累	lèi/lěi	类	lèi	G				PRON	
類	lèi	数	shù/shǔ	G					
棱	lēng/léng	陆	lù	G		PC			
棱	lēng/léng	睦	mù	G		PC			
棱	lēng/léng	穆	mù	G					
梨	lí	栗	lì	G	SC		PCPR	PRON	SEM
李	li	季	jì	G	SC				
李	li	梨	lí	G	SC			PRON	SEM
理	lǐ/li	埋	mái/mán	G					
禮	lǐ	體	tǐ	G					
裏	lǐ/li	裹	guǒ	G	SC				SEM
俐	lì	伶	ling	G	SC				SEM
厉	lì	害	hài/hai						SEM
厉	lì	励	lì	G		PC	PCPR	PRON	
厉	lì	迈	mài	G					
歷	lì	裏	lǐ/li				PCPR	PRON	
荔	lì	蕾	lěi	G	SC				
隶	lì	录	lù	G	SC				
隸	lì	录	lù	G	SC				
隸	lì	錄	lù	G					
帘	lián	罕	hǎn	G					
怜	lián	恋	liàn					PRON	SEM
聯	lián	聊	liáo	G	SC				
练	liàn	联	lián				PCPR	PRON	
练	liàn	连	lián	G		PC	PCPR	PRON	
练	liàn	炼	liàn	G		PC	PCPR	PRON	SEM
凉	liáng	冷	lěng	G	SC				
梁	liáng/liang	渠	qú	G	SC				
梁	liáng/liang	染	rǎn	G	SC				
粱	liáng	粮	liáng	G	SC		PCPR	PRON	SEM
量	liàng/liang	童	tóng	G					
量	liàng/liang	晝	zhòu	G	SC				
遼	liáo	遥	yáo	G	SC				SEM
列	liè	裂	liè	G		PC	PCPR	PRON	
烈	liè	裂	liè	G		PC	PCPR	PRON	
裂	liè	列	liè	G		PC	PCPR	PRON	
临	lín	邻	lín	G				PRON	SEM
磷	lín	硫	liú	G	SC				SEM
臨	lín	鄰	lín					PRON	SEM
赁	lìn	吝	lìn					PRON	
賃	lìn	任	rèn	G		PC			
凌	líng	冷	lěng	G	SC	PC			
凌	líng	灵	líng/ling				PCPR	PRON	
凌	líng	凄	qī	G	SC				
岭	lǐng	耸	sǒng						SEM
岭	lǐng	崭	zhǎn	G	SC				SEM
灵	líng/ling	零	líng					PRON	
陵	líng	隘	ài	G	SC				
陵	líng	棱	lēng/léng	G		PC	PCPR	PRON	
陵	líng	楞	léng						
靈	líng/ling	零	líng	G	SC			PRON	
靈	líng/ling	营	yíng						
靈	líng/ling	赢	yíng						
龄	líng	岭	lǐng	G		PC	PCPR	PRON	
龄	líng	铃	líng	G		PC	PCPR	PRON	
龄	líng	零	líng	G		PC	PCPR	PRON	
龄	líng	临	lín						
龄	líng	邻	lín	G		PC			
嶺	lǐng	耸	sǒng						
榴	liú	柳	liǔ	G	SC	PC	PCPR	PRON	SEM
浏	liú	览	lǎn						SEM
浏	liú	溜	liū	G	SC			PRON	SEM
浏	liú	流	liú	G	SC			PRON	SEM
瘤	liú	癌	ái	G	SC				SEM

窿	long	隆	lóng/lōng	G		PC	PCPR	PRON	
聋	lóng	耳	ěr	G	SC				SEM
隆	lóng/lōng	降	jiàng/xiáng	G	SC				
隆	lóng/lōng	陵	líng	G	SC				
隆	lóng/lōng	窿	long	G		PC	PCPR	PRON	
龍	lóng	韵	yùn	G					
龍	lóng	韻	yùn	G					
垄	lǒng	笼	lóng/long	G		PC	PCPR	PRON	
垄	lǒng	袭	xí	G		PC			
拢	lǒng	笼	lóng/long	G		PC	PCPR	PRON	
拢	lǒng	垄	lǒng	G		PC	PCPR	PRON	
拢	lǒng	抚	rǎo	G	SC				
漏	lòu	泪	lèi	G	SC				
漏	lòu	沃	wò	G	SC				
碌	lù	录	lù	G		PC	PCPR	PRON	
碌	lù	錄	lù	G			PCPR	PRON	
露	lù	漏	lòu	G					
蘆	lú/lu	蔔	bo	G	SC				
蘆	lú/lu	薦	jiàn	G	SC				
鲁	lǔ	兽	shòu						
鲁	lǔ	昼	zhòu	G	SC				
鲁	lǔ	晝	zhòu	G	SC				
錄	lù	碌	lù	G				PRON	
驴	lǘ	骡	luó	G	SC				SEM
驴	lǘ	骆	luò	G	SC				SEM
侣	lǚ	偶	ǒu	G	SC				SEM
屡	lǚ	履	lǚ	G	SC			PRON	
屡	lǚ	屎	shǐ	G	SC				
履	lǚ	屡	lǚ	G	SC			PRON	
旅	lǚ	放	fàng	G		PC	PCPR		
旅	lǚ	旋	xuán	G					
旅	lǚ	族	zú	G		PC			
滤	lǜ	溃	kuì	G	SC				
率	lǜ/shuài	律	lǜ					PRON	
绿	lǜ	级	jí	G	SC				
绿	lǜ	纪	jì	G	SC				
绿	lǜ	录	lù	G		PC	PCPR		
绿	lǜ	錄	lù	G			PCPR		
虑	lǜ	思	sī/si	G	SC				SEM
抡	lūn	轮	lún	G		PC	PCPR	PRON	
羅	luó	節	jié						
萝	luó	夢	mèng	G	SC				
萝	luó	梦	mèng	G					
骡	luó	骆	luò	G	SC		PCPR	PRON	SEM
络	luò	罗	luō/luó	G				PRON	SEM
骆	luò	赂	lù	G		PC			
骆	luò	路	lù	G					
麻	má/ma	嘛	ma/má	G		PC		PRON	
脉	mài	胀	zhàng	G	SC				
脉	mài	肥	féi	G	SC				
瞒	mán	眠	mián	G	SC				
蔓	mán/màn	漫	màn	G		PC	PCPR	PRON	
猫	māo	貌	mào	G	SC			PRON	
貓	māo	貌	mào	G	SC	PC	PCPR	PRON	
茂	mào	蔑	miè	G	SC				
枚	méi	梅	méi	G	SC			PRON	
梅	méi	梧	wú	G	SC				
煤	méi	媒	méi	G		PC		PRON	
玫	méi	纹	wén	G					
眉	méi	旬	xún	G					
酶	méi	霉	méi	G		PC	PCPR	PRON	
霉	méi	酶	méi	G		PC	PCPR	PRON	
黴	méi	徽	huī	G	SC				
黴	méi	酶	méi					PRON	
镁	měi	汞	gǒng						SEM
昧	mèi	暧	ài/nuǎn	G	SC				SEM
昧	mèi	暗	àn	G	SC				

弥	mí	迷	mí					PRON	
弥	mí	覓	mì					PRON	
密	mì	秘	mì	G		PC	PCPR	PRON	SEM
密	mì	蜜	mì	G	SC	PC	PCPR	PRON	
蜜	mì	密	mì	G	SC	PC	PCPR	PRON	
蜜	mì	秘	mì	G		PC	PCPR	PRON	
棉	mián	歸	guī						
棉	mián	绵	mián	G	SC	PC	PCPR	PRON	SEM
棉	mián	柿	shì	G	SC				
眠	mián	盲	máng	G	SC				
绵	mián	缠	chán	G	SC				SEM
绵	mián	锦	jǐn	G	SC	PC			
绵	mián	棉	mián	G	SC	PC	PCPR	PRON	SEM
绵	mián	绣	xiù	G	SC				
渺	miǎo	晰	xī	G					
秒	miǎo	妙	miào	G	SC			PRON	
秒	miǎo	稍	shāo	G	SC				
廟	miào	朝	cháo/zhāo	G		PC	PCPR		
蔑	miè	茂	mào	G	SC				
蠛	miè	蜥	mài	G	SC				
蠛	miè	茂	mào	G					
蠛	miè	襪	wà	G		PC			
敏	mǐn	酶	méi	G		PC	PCPR		
敏	mǐn	霉	méi	G		PC	PCPR		
鸣	míng	呜	wū	G	SC				SEM
摸	mō/mo	抹	mǒ/mā	G	SC	PC	PCPR	PRON	SEM
摸	mō/mo	模	mó/mú	G		PC	PCPR	PRON	
摸	mō/mo	摩	mó	G		PC	PCPR	PRON	SEM
磨	mó/mo	摩	mó	G		PC	PCPR	PRON	SEM
膜	mó	抹	mǒ/mā	G		PC	PCPR	PRON	
抹	mǒ/mā	摸	mō/mo	G	SC	PC	PCPR	PRON	SEM
寞	mò	末	mò			PC	PCPR	PRON	
寞	mò	宴	yàn	G	SC				
沫	mò/mo	漠	mò	G	SC	PC	PCPR	PRON	
沫	mò/mo	沐	mù	G	SC				
漠	mò	莫	mò/mo	G		PC	PCPR	PRON	SEM
默	mò	陌	mò					PRON	
谋	móu/mou	媒	méi	G		PC			
谋	móu/mou	某	mǒu	G		PC	PCPR	PRON	
谋	móu/mou	诈	zhà	G	SC				SEM
畝	mǔ	玖	jiǔ	G		PC	PCPR		
畝	mǔ	庙	miào						
墓	mù	基	jī	G	SC				
暮	mù	幕	mù	G	SC	PC		PRON	
暮	mù	慕	mù	G	SC	PC		PRON	
睦	mù	穆	mù	G				PRON	
穆	mù	稀	xī	G	SC				
乃	nǎi	且	qiě	G					
奈	nài	耐	nài					PRON	
挠	náo	交	jiāo			PC	PCPR		
挠	náo	饶	ráo	G		PC			
嫩	nèn	椒	jiāo	G					
尼	ní	匙	shi	G					
捻	niǎn	拧	níng	G	SC				SEM
鳥	niǎo	烏	wū	G					SEM
纽	niǔ	绸	chóu	G	SC				
纽	niǔ	扭	niǔ/niu	G	SC	PC	PCPR	PRON	
农	nóng	衣	yī	G					
農	nóng	麴	qū	G					
怒	nù	奴	nú	G		PC	PCPR	PRON	
怒	nù	恕	shù	G	SC				
怒	nù	絮	xù	G					
怒	nù	怨	yuàn/yuan	G	SC				
暖	nuǎn	暧	ài/nuǎn	G	SC			PRON	
暖	nuǎn	暄	xuān	G	SC				SEM
虐	nvè	掠	è						SEM
虐	nvè	略	è						SEM
欧	ōu	殴	ōu	G	SC	PC	PCPR	PRON	

徘	pái	匪	fěi	G		PC	PCPR		
牌	pái	板	bǎn	G					SEM
牌	pái	版	bǎn	G		PC			SEM
盤	pán/pan	盆	pén	G					SEM
叛	pàn	判	pàn	G		PC	PCPR	PRON	
畔	pàn	鉡	xìn	G		PC			
庞	páng	宠	chǒng	G		PC			
抛	pāo	攀	pān						
袍	páo	包	bāo	G		PC	PCPR	PRON	
袍	páo	刨	páo	G		PC	PCPR	PRON	
陪	péi	部	bù	G	SC	PC			
陪	péi	培	péi	G		PC	PCPR	PRON	
佩	pèi	配	pèi					PRON	
沛	pèi	沸	fèi	G	SC	PC	PCPR	PRON	
配	pèi	培	péi					PRON	
配	pèi	赔	péi					PRON	
配	pèi	陪	péi					PRON	
配	pèi	佩	pèi					PRON	
喷	pēn	损	sǔn	G					
盆	pén	岔	chà	G	SC	PC			
盆	pén	盘	pán/pan	G					SEM
盆	pén	盼	pàn	G	SC	PC			
棚	péng/peng	绷	běng/bēng	G		PC	PCPR	PRON	
棚	péng/peng	篷	peng			PC	PCPR	PRON	SEM
篷	peng	蓬	péng	G		PC	PCPR	PRON	
蓬	péng	勃	bó						SEM
蓬	péng	莲	lián	G	SC				SEM
蓬	péng	篷	peng	G		PC	PCPR	PRON	
捧	pěng	棒	bàng	G		PC			
闢	pī/pì	避	bì	G		PC	PCPR	PRON	
劈	pī	剪	jiǎn	G	SC				SEM
披	pī	搏	bó	G	SC	PC	PCPR		
疲	pí	脾	pí					PRON	SEM
屁	pì	庇	bì	G		PC	PCPR	PRON	
屁	pì	毙	bì	G		PC	PCPR	PRON	
屁	pì	尸	shī	G	SC				
譬	pì	誓	shì	G	SC				
譬	pì	誉	yù	G	SC				
譬	pì	譽	yù	G	SC				
飘	piāo	飙	biāo	G	SC	PC	PCPR	PRON	SEM
撇	piē/piě	撒	sā/sǎ	G	SC				SEM
瞥	piē	督	dū	G	SC				SEM
瞥	piē	监	jiān						SEM
拼	pīn	秉	bǐng				PCPR		
贫	pín	费	fèi	G	SC				
贫	pín	穷	qióng	G					SEM
贫	pín	贪	tān	G	SC				
贫	pín	污	wū						
蘋	píng	藥	yào	G	SC				
凭	píng	评	píng					PRON	
评	píng	平	píng	G		PC	PCPR	PRON	
坡	pō	堆	duī	G	SC				
坡	pō	披	pī	G		PC			
坡	pō	破	pò	G		PC	PCPR	PRON	
颇	pō	披	pī	G		PC			
颇	pō	坡	pō	G		PC	PCPR	PRON	SEM
破	pò	迫	pò				PCPR	PRON	SEM
魄	pò	魂	hún	G	SC				SEM
扑	pū	挂	guà	G	SC				
扑	pū	仆	pū/pú	G				PRON	SEM
扑	pū	铺	pù/pū	G				PRON	
扑	pū	朴	pǔ	G				PRON	
撲	pū	僕	pú	G		PC	PCPR	PRON	SEM
铺	pù/pū	補	bǔ	G		PC	PCPR	PRON	
葡	pú	蔔	bo	G	SC	PC			
瀑	pù	暴	bào	G		PC	PCPR		
瀑	pù	爆	bào	G		PC	PCPR		
瀑	pù	泡	pào	G	SC	PC	PCPR		

岂	qǐ	壹	yī	G					
凄	qī	沏	qī	G			PCPR	PRON	
凄	qī	漆	qī	G		PC	PCPR	PRON	
柒	qī	渠	qú	G	SC				
柒	qī	染	rǎn	G	SC				
欺	qī	期	qī	G		PC	PCPR	PRON	
沏	qī	澈	chè	G	SC				SEM
沏	qī	漆	qī	G	SC			PRON	
齊	qí	弃	qì	G				PRON	
齊	qí	棄	qì					PRON	
弃	qì	奇	qí	G				PRON	
弃	qì	齐	qí	G				PRON	
弃	qì	异	yì	G	SC				
棄	qì	案	àn	G	SC				
棄	qì	葉	yè	G					
棄	qì	業	yè						
汽	qì	气	qì/qi	G	SC	PC	PCPR	PRON	SEM
迄	qì	疙	gē	G					
迄	qì	乞	qǐ	G				PRON	
掐	qiā	陷	xiàn	G		PC	PCPR		
掐	qiā	馅	xiàn	G		PC	PCPR		
恰	qià	洽	qià	G				PRON	
牵	qiān	沿	yán/yàn						SEM
籤	qiān	截	jié	G					
縴	qiàn	從	cóng/cōng						
縴	qiàn	截	jié						
谦	qiān	怜	lián	G					
谦	qiān	签	qiān			PC	PCPR	PRON	
谦	qiān	谴	qiǎn	G	SC	PC	PCPR	PRON	
遷	qiān	選	xuǎn	G	SC				
韆	qiān	遷	qiān	G		PC	PCPR	PRON	
潜	qián	涕	tì	G	SC				
钳	qián	嵌	qiàn	G				PRON	
谴	qiǎn	谦	qiān	G	SC	PC	PCPR	PRON	
谴	qiǎn	潜	qián				PCPR	PRON	
谴	qiǎn	遣	qiǎn	G		PC	PCPR	PRON	
遣	qiǎn	遷	qiān	G	SC	PC	PCPR	PRON	SEM
遣	qiǎn	谴	qiǎn	G		PC	PCPR	PRON	
歉	qiàn	欠	qiàn/qian	G	SC		PCPR	PRON	
歉	qiàn	赚	zhuàn	G		PC			
翘	qiáo	膏	gāo	G	SC				
悄	qiāo	俏	qiào	G		PC	PCPR	PRON	
侨	qiáo	僚	liáo	G	SC				SEM
瞧	qiáo	翘	qiào			PC	PCPR	PRON	SEM
俏	qiào	悄	qiāo	G		PC	PCPR	PRON	
窃	qiè	锲	qiè			PC	PCPR	PRON	
窃	qiè	穷	qióng	G	SC				
窃	qiè	弃	qì						
锲	qiè	鐵	tiě		SC				
钦	qīn	嵌	qiàn	G					
钦	qīn	铅	qiān	G	SC				
勤	qín	竭	jié						SEM
勤	qín	谨	jǐn	G		PC	PCPR	PRON	
勤	qín	勸	quàn	G	SC				
清	qīng	青	qīng	G	SC	PC	PCPR	PRON	
蜻	qīng	青	qīng	G	SC	PC	PCPR	PRON	SEM
蜻	qīng	蜓	tíng	G	SC				SEM
蜻	qīng	蝇	ying	G	SC				SEM
擎	qíng	诚	chéng						
擎	qíng	挚	zhì	G	SC				
庆	qìng	厌	yàn	G					
庆	qìng	忧	yōu	G					
庆	qìng	憂	yōu						
慶	qìng	愛	ài	G	SC				
慶	qìng	请	qǐng					PRON	
穷	qióng	究	jiū/jiu	G	SC				
穷	qióng	帘	lián	G	SC				
穷	qióng	劣	liè	G	SC				SEM

穷	qióng	强	qiáng/qiǎng	G					
区	qū	凶	xiōng	G					
屈	qū/qu	掘	jué	G					
趋	qū	超	chāo	G	SC				
趋	qū	驱	qū					PRON	
趋	qū	追	zhuī	G					
驱	qū	趋	qū					PRON	
權	quán	椎	chuí/zhuī	G	SC				
勸	quàn	動	dòng	G	SC				
勸	quàn	觀	guān	G		PC	PCPR		
勸	quàn	歡	huan/huān	G		PC	PCPR		
勸	quàn	勤	qín	G	SC				
勸	quàn	權	quán	G		PC		PRON	
瘸	qué	腐	fǔ/fu	G					
瘸	qué	瘫	tān	G	SC				SEM
榷	què	椎	chuí/zhuī	G	SC				
榷	què	椿	zhuāng	G	SC				
榷	què	准	zhǔn	G					
确	què	触	chù	G					
确	què	角	jiǎo/jué	G				PRON	
確	què	榷	què	G		PC	PCPR	PRON	
確	què	准	zhǔn	G					SEM
確	què	準	zhǔn	G					SEM
雀	què	霍	huò	G	SC				
雀	què	鹊	que					PRON	SEM
鹊	que	鹅	é	G	SC				SEM
鹊	que	鸽	gē	G	SC				SEM
群	qún	裙	qún	G		PC	PCPR	PRON	
染	rǎn	究	jiū/jiu	G					
染	rǎn	梁	liáng/liang	G	SC				
染	rǎn	渠	qú	G	SC				
嚷	rǎng	囊	nāng/náng	G					
嚷	rǎng	嗓	sǎng	G	SC				SEM
嚷	rǎng	響	xiǎng						SEM
壤	rǎng	壞	huài	G	SC				
讓	ràng	護	hù	G	SC				
讓	ràng	嚷	rǎng	G		PC	PCPR	PRON	
讓	ràng	聽	tīng/ting	G					
饶	ráo	饺	jiǎo	G	SC	PC	PCPR		
扰	rǎo	绕	rào/rǎo	G				PRON	
擾	rǎo	搗	dǎo	G	SC				SEM
擾	rǎo	摄	shè	G	SC				
擾	rǎo	投	tóu	G	SC				
擾	rǎo	優	yōu	G		PC	PCPR		
擾	rǎo	憂	yōu	G		PC	PCPR		
绕	rào/rǎo	饶	ráo	G		PC		PRON	
惹	rě	忍	rěn	G	SC				
热	rè	熟	shú	G	SC	PC			
饪	rèn	蚀	shí	G	SC				
溶	róng	熔	róng	G		PC	PCPR	PRON	SEM
溶	róng	浴	yù	G	SC				
荣	róng	宋	sòng	G					
柔	róu	揉	róu	G		PC	PCPR	PRON	
乳	rǔ	俘	fú	G		PC	PCPR		
乳	rǔ	浮	fú	G		PC	PCPR		
乳	rǔ	辱	rǔ					PRON	
辱	rǔ	夺	duó	G					
辱	rǔ	乳	rǔ					PRON	
瑞	ruì	端	duān	G					
瑞	ruì	锐	ruì					PRON	
撒	sā/sǎ	撤	chè	G	SC				SEM
洒	sǎ	酒	jiǔ	G	SC				
灑	sǎ	灌	guàn	G	SC				
灑	sǎ	撒	sā/sǎ					PRON	SEM
灑	sǎ	曬	shài	G		PC			
塞	sāi/sè	寨	zhài	G	SC	PC			
赛	sài	债	zhài	G					
叁	sān	参	cān/shēn	G		PC	PCPR	PRON	

啬	sè	塞	sāi/sè	G				PRON	
啬	sè	吝	lìn	G					SEM
啬	sè	塞	sāi/sè	G				PRON	
厦	shà	厚	hòu	G					
曬	shài	寒	hán						
曬	shài	麗	lì	G		PC	PCPR		
珊	shān	删	shān	G		PC	PCPR	PRON	
擅	shàn	颤	chàn	G		PC	PCPR	PRON	
擅	shàn	壇	tán	G		PC			
傷	shāng	陽	yáng	G					
赏	shǎng	奖	jiǎng	G					SEM
捎	shāo	捐	juān	G	SC				
捎	shāo	敲	qiāo				PCPR		
稍	shāo	秒	miǎo	G	SC				
稍	shāo	悄	qiāo	G		PC	PCPR		
稍	shāo	敲	qiāo				PCPR		
稍	shāo	销	xiāo	G		PC	PCPR		
勺	sháo	勾	gōu	G					
哨	shào	啸	xiào	G	SC		PCPR		SEM
绍	shào	招	zhāo	G		PC	PCPR	PRON	
奢	shē	奔	bēn/bèn	G	SC				
摄	shè	缀	zhui						
绅	shēn	绳	shéng/sheng	G	SC				
婶	shěn/shen	嫂	sǎo/sao	G	SC				SEM
婶	shěn/shen	绅	shēn	G		PC	PCPR	PRON	
审	shěn	伸	shēn	G		PC	PCPR	PRON	
审	shěn	申	shēn	G		PC	PCPR	PRON	
審	shěn	番	fān	G		PC	PCPR		
渗	shèn	深	shēn	G	SC			PRON	SEM
肾	shèn	肩	jiān	G	SC		PCPR		SEM
盛	shèng/chéng	剩	shèng			PC	PCPR	PRON	SEM
牲	shēng	牺	xī	G	SC				SEM
胜	shèng	膀	bǎng	G	SC				
胜	shèng	生	shēng/sheng					PRON	
胜	shèng	腾	téng/teng	G	SC	PC			
圣	shèng	怪	guài	G					
聖	shèng	勝	shèng					PRON	
聖	shèng	神	shén/shen						SEM
胜	shèng	牲	shēng	G		PC		PRON	
適	shì	遮	dì	G	SC				
適	shì	遭	zāo	G	SC				
適	shì	遵	zūn	G	SC				
实	shí/shi	买	mǎi	G		PC	PCPR		
實	shí/shi	賣	mài/mai	G					
拾	shi/shí	恰	qià	G					
拾	shi/shí	洽	qià	G					
势	shì/shi	努	nǔ	G	SC				
誓	shì	譬	pì	G	SC				
誓	shì	誉	yù	G	SC				
誓	shì	譽	yù	G	SC				
逝	shì	折	zhé/zhē	G		PC	PCPR		
逝	shì	遮	zhē	G	SC	PC	PCPR		
释	shì	译	yì	G		PC	PCPR		SEM
壽	shòu	辜	gū	G	SC				
叔	shū/shu	权	quán	G					
叔	shū/shu	述	shù					PRON	
抒	shū	舒	shū	G				PRON	
書	shū	晝	zhòu	G	SC				
梳	shū	榴	liú	G	SC				
梳	shū	柳	liǔ	G	SC				
梳	shū	疏	shū	G		PC	PCPR	PRON	
殊	shū	珠	zhū	G		PC	PCPR	PRON	
殊	shū	蛛	zhū	G		PC	PCPR	PRON	
舒	shū	抒	shū	G				PRON	
舒	shū	豫	yù	G					
蔬	shū	疏	shū	G		PC	PCPR	PRON	
输	shū	與	yú	G	SC		PCPR		SEM
输	shū	御	yù			PC	PCPR		SEM

熟	shú	熱	rè	G	SC	PC			
属	shǔ	嘱	zhǔ	G		PC	PCPR	PRON	
屬	shǔ	遲	chí	G					
暑	shǔ	署	shǔ	G	SC	PC	PCPR	PRON	
薯	shǔ	署	shǔ	G		PC	PCPR	PRON	
薯	shǔ	著	zhe/zháo	G	SC	PC	PCPR	PRON	
恕	shù	怒	nù	G	SC				
恕	shù	饶	ráo						SEM
恕	shù	絮	xù	G		PC			
摔	shuāi	挥	huī	G	SC				
摔	shuāi	率	lǜ/shuài	G		PC	PCPR	PRON	
摔	shuāi	撵	niǎn	G	SC				
衰	shuāi	哀	āi	G	SC				
衰	shuāi	挨	āi						
衰	shuāi	摔	shuāi	G				PRON	
衰	shuāi	衷	zhōng	G	SC				
甩	shuǎi	爽	shuǎng						
帅	shuài	師	shī	G	SC	PC			
爽	shuǎng	夹	jiā/jiá	G	SC				
睡	shuì	眠	mián	G	SC				SEM
税	shuì	兑	duì	G					
税	shuì	悦	yuè	G					
税	shuì	稚	zhì	G	SC				
烁	shuò	硕	shuò					PRON	
诵	sòng	背	bèi/bēi						SEM
诵	sòng	讼	sòng	G	SC			PRON	
颂	sòng	讼	sòng	G		PC	PCPR	PRON	
搜	sōu	骚	sāo						
搜	sōu	瘦	shòu	G		PC			
嗽	sou	喇	lǎ	G	SC				
塑	sù	素	sù					PRON	
塑	sù	望	wàng/wang	G					
素	sù	肃	sù	G				PRON	
肃	sù	素	sù	G				PRON	
肃	sù	啸	xiào	G					
肃	sù	素	sù					PRON	
肃	sù	繡	xiù	G					
酸	suān	醋	cù	G	SC				SEM
酸	suān	峻	jùn	G		PC	PCPR		
酸	suān	酗	xù	G	SC				
蒜	suàn	棘	jí	G					SEM
歲	suì	滅	miè	G					
碎	suì	破	pò	G	SC				SEM
穗	suì	惠	huì	G					
穗	suì	慧	huì	G					
穗	suì	秽	huì	G	SC				
穗	suì	讳	huì						
隧	suì	随	suí	G	SC			PRON	
笋	sǔn	损	sǔn					PRON	
塌	tā	蹋	tà	G		PC	PCPR	PRON	
塔	tǎ	搭	dā	G		PC	PCPR	PRON	SEM
蹋	tà	塌	tā	G		PC	PCPR	PRON	
抬	tái	始	shǐ	G		PC			
抬	tái	执	zhí	G	SC				
抬	tái	制	zhì	G					
瘫	tān	瘸	qué	G	SC				SEM
贪	tān	贫	pín	G	SC				
潭	tán	滩	tān	G	SC			PRON	
痰	tán	瘫	tān	G	SC	PC	PCPR	PRON	SEM
叹	tàn	汗	hàn						
嘆	tàn	喊	hǎn	G	SC				
探	tàn/tan	控	kòng	G	SC				
炭	tàn	岸	àn	G	SC				
炭	tàn	岩	yán	G	SC				
碳	tàn	碍	ài	G	SC				
碳	tàn	灰	huī	G					
碳	tàn	炭	tàn	G		PC	PCPR	PRON	SEM
趟	tàng	超	chāo	G	SC				

堂	táng/tang	常	cháng	G					
塘	táng	堤	dī	G	SC				SEM
塘	táng	坛	tán	G	SC				
塘	táng	壇	tán	G	SC				
倘	tǎng	尚	shàng/shang	G		PC	PCPR		
躺	tǎng	背	bèi/bēi						SEM
掏	tāo	抅	jū	G	SC				
掏	tāo	捞	lāo	G	SC				SEM
涛	tāo	滔	tāo	G	SC			PRON	SEM
滔	tāo	稻	dào	G		PC	PCPR	PRON	
滔	tāo	涛	tāo	G	SC		PCPR	PRON	SEM
滔	tāo	淘	táo	G	SC	PC	PCPR	PRON	
桃	táo	梨	lí	G	SC				SEM
桃	táo	兆	zhào	G		PC			
淘	táo	捞	lāo						
萄	táo	掏	tāo	G		PC	PCPR	PRON	
陶	táo	淘	táo	G		PC	PCPR	PRON	
陶	táo	窑	yáo	G					SEM
疼	téng	痛	tòng	G	SC	PC	PCPR		SEM
剔	tī	剥	bō	G	SC				
剔	tī	剃	tì	G	SC	PC	PCPR	PRON	
剔	tī	惕	tì	G		PC	PCPR	PRON	
剔	tī	削	xuē/xiāo	G	SC				
蹄	tí	蒂	dì	G		PC	PCPR	PRON	
體	tǐ	禮	lǐ	G					
嚏	tì	嚏	dāng	G	SC				
惕	tì	惮	dàn	G	SC				
惕	tì	警	jǐng						
惕	tì	锡	xī	G		PC			
涕	tì	滴	dī	G	SC		PCPR	PRON	SEM
涕	tì	涤	dí	G	SC		PCPR	PRON	
涕	tì	泣	qì	G	SC				SEM
涕	tì	嚏	tì				PCPR	PRON	SEM
舔	tiǎn	舌	shé	G	SC				SEM
舔	tiǎn	添	tiān	G		PC	PCPR	PRON	
舔	tiǎn	填	tián				PCPR	PRON	
舔	tiǎn	甜	tián	G			PCPR	PRON	
挑	tiāo/tiǎo	逃	táo	G		PC			
帖	tiē/tiě	贴	tiē	G	SC	PC		PRON	
帖	tiē/tiě	粘	zhān	G	SC	PC	PCPR		
帖	tiē/tiě	佔	zhàn	G	SC	PC	PCPR		
贴	tiē	粘	zhān	G	SC	PC	PCPR		SEM
廳	tīng	聽	tīng/ting	G		PC	PCPR	PRON	
童	tóng	撞	zhuàng	G		PC	PCPR		
筒	tǒng	笛	dí	G	SC				
偷	tōu	输	shū	G		PC			
徒	tú	途	tú	G				PRON	SEM
徒	tú	徙	xǐ	G	SC				
團	tuán	圖	tú	G	SC				
退	tuì	痕	hén	G	SC				
托	tuō	拖	tuō	G	SC			PRON	SEM
拖	tuō	施	shī	G					
驼	tuo/tuó	骆	luò	G	SC				SEM
妥	tuǒ	觅	mì	G					
椭	tuǒ	率	lǜ/shuài	G					SEM
唾	tuò	吐	tǔ/tù	G	SC				SEM
娃	wá/wa	佳	jiā	G					
娃	wá/wa	姓	xìng	G	SC				
弯	wān	变	biàn	G					
弯	wān	恋	liàn	G					
顽	wán	玩	wán	G		PC		PRON	
顽	wán	碗	wǎn	G				PRON	
汪	wāng	旺	wàng	G	SC			PRON	
枉	wang	拄	zhǔ	G					
枉	wang	柱	zhù	G	SC				
忘	wàng	念	niàn	G	SC				SEM
唯	wéi	准	zhǔn	G	SC				
维	wéi	围	wéi	G				PRON	

伟	wěi	唯	wéi	G			PCPR	PRON	
纬	wěi	维	wéi	G	SC		PCPR	PRON	
纬	wěi	伟	wěi	G		PC	PCPR	PRON	
胃	wèi	畏	wèi	G	SC	PC	PCPR	PRON	
衛	wèi	街	jiē	G	SC				
衛	wèi	围	wéi					PRON	
瘟	wēn	瘫	tān	G	SC				SEM
稳	wěn	隐	yǐn	G		PC	PCPR		
紊	wěn	吝	lìn	G					
握	wò	据	jù	G	SC				
乌	wū	鸟	niǎo	G		PC			SEM
乌	wū	呜	wū	G		PC	PCPR	PRON	
呜	wū	鸣	míng	G	SC				SEM
污	wū	夸	kuā	G		PC	PCPR		
污	wū	亏	kuī	G		PC			
烏	wū	鳥	niǎo	G					SEM
诬	wū	误	wù/wu	G	SC		PCPR	PRON	SEM
梧	wú	桐	tóng	G	SC				SEM
梧	wú	晤	wù	G		PC	PCPR	PRON	
侮	wǔ	悔	huǐ	G		PC			
悟	wù	误	wù/wu	G			PCPR	PRON	
晤	wù	悟	wù	G		PC	PCPR	PRON	
繫	xì	擊	jī	G		PC	PCPR	PRON	
息	xi/xī	悉	xī	G	SC		PCPR	PRON	
犧	xī	牲	shēng	G	SC				SEM
犧	xī	特	tè	G	SC				
膝	xī	漆	qī	G		PC	PCPR	PRON	
锡	xī	锦	jǐn	G	SC				
锡	xī	绣	xiù						
锡	xī	银	yín	G	SC				SEM
媳	xí	熄	xī	G		PC	PCPR	PRON	
席	xí	度	dù/du	G					
席	xí	廣	guǎng	G					
習	xí	摺	zhé	G					
徙	xǐ	徒	tú	G	SC				
徙	xǐ	途	tú						
徙	xǐ	喜	xǐ					PRON	
瞎	xiā	害	hài/hai	G		PC			
瞎	xiā	盼	pàn	G	SC				
辖	xiá	辐	fú	G	SC				
掀	xiān	撕	sī	G	SC				
纖	xiān	殲	jiān	G		PC	PCPR	PRON	
纖	xiān	籤	qiān	G		PC	PCPR	PRON	
咸	xián	或	huò	G	SC				
嫌	xián	奸	jiān	G	SC		PCPR	PRON	
嫌	xián	廉	lián	G					
嫌	xián	赚	zhuàn	G		PC			
衔	xián	镶	xiāng	G	SC				
衔	xián	衍	yǎn	G					
险	xiǎn	检	jiǎn	G		PC	PCPR	PRON	
险	xiǎn	怜	lián	G					
县	xiàn	悬	xuán	G	SC				
獻	xiàn	勸	quàn						
縣	xiàn	懸	xuán	G		PC			
羡	xiàn	次	cì	G	SC	PC	PCPR		
厢	xiāng	箱	xiāng	G		PC	PCPR	PRON	SEM
鄉	xiāng	绑	bǎng	G					
祥	xiáng	详	xiáng	G				PRON	
翔	xiáng	羽	yǔ	G	SC				
详	xiáng	美	xiàn	G					
響	xiǎng	鄉	xiāng	G		PC	PCPR	PRON	
響	xiǎng	音	yīn	G	SC				SEM
嚮	xiàng	鄉	xiāng	G		PC	PCPR	PRON	
消	xiāo	销	xiāo	G		PC	PCPR	PRON	SEM
销	xiāo	钞	chāo	G	SC				
销	xiāo	消	xiāo	G		PC	PCPR	PRON	SEM
销	xiāo	钥	yào	G	SC				

晓	xiǎo	晌	shǎng	G	SC				SEM
晓	xiǎo	暄	xuān	G	SC				
协	xié	胁	xié	G		PC	PCPR	PRON	
協	xié	携	xié				PCPR	PRON	
協	xié	脅	xié	G		PC	PCPR	PRON	
挟	xié	持	chí	G	SC				SEM
挟	xié	携	xié	G	SC			PRON	SEM
携	xié	挟	xié	G	SC			PRON	SEM
斜	xié	科	kē	G					
斜	xié	料	liào	G					
胁	xié	膀	bǎng	G	SC				SEM
胁	xié	旁	páng						SEM
胁	xié	挟	xié	G			PCPR	PRON	SEM
脅	xié	挟	xié	G			PCPR	PRON	SEM
卸	xiè	泄	xiè					PRON	SEM
械	xiè	诫	jiè	G		PC	PCPR	PRON	
泄	xiè	泻	xiè	G	SC			PRON	SEM
锌	xīn	锡	xī	G	SC				SEM
鬻	xìn	寡	guǎ	G					
腥	xīng	牲	shēng	G		PC			SEM
腥	xīng	醒	xǐng	G		PC	PCPR	PRON	
興	xìng/xīng	與	yǔ/yù	G	SC				
兇	xiōng	兜	dōu	G	SC				
兇	xiōng	兒	r/ér	G	SC				
羞	xiū	丑	chǒu	G	SC	PC			
羞	xiū	寿	shòu	G					
朽	xiǔ	巧	qiǎo	G		PC			
嗅	xiù	臭	ào						
秀	xiù	香	xiāng	G	SC				
秀	xiù	优	yōu	G					SEM
绣	xiù	修	xiū				PCPR	PRON	
绣	xiù	锈	xiù	G		PC	PCPR	PRON	
绣	xiù	诱	yòu	G		PC			
锈	xiù	朽	xiǔ				PCPR	PRON	SEM
锈	xiù	铀	yóu	G	SC				
墟	xū	坛	tán	G	SC				
墟	xū	壇	tán	G	SC				
徐	xú	循	xún	G	SC				
叙	xù	絮	xù					PRON	
叙	xù	绪	xù					PRON	
絮	xù	挠	náo						
絮	xù	饶	ráo						
絮	xù	恕	shù	G		PC	PCPR		
絮	xù	述	shù			PC	PCPR		
絮	xù	繫	xì	G	SC				
续	xù	序	xù					PRON	SEM
续	xù	绪	xù	G	SC			PRON	
蓄	xù	蕴	yùn	G	SC				SEM
宣	xuān	暄	xuān	G		PC	PCPR	PRON	
悬	xuán	嫌	xián						
悬	xuán	息	xi/xī	G	SC				
懸	xuán	慰	wèi	G	SC				
懸	xuán	惜	xī				PCPR		
旋	xuán	族	zú	G					
雪	xuě	雷	léi	G	SC				SEM
寻	xún/xín	询	xún	G				PRON	
循	xún	旬	xún					PRON	
循	xún	徐	xú	G	SC				
询	xún	讯	xùn	G	SC	PC	PCPR	PRON	
迅	xùn	寻	xún/xín				PCPR	PRON	
迅	xùn	讯	xùn	G		PC	PCPR	PRON	
迅	xùn	逊	xùn	G	SC		PCPR	PRON	
轧	yà	扎	zhā/zhá	G					
壓	yā	厭	yàn	G					
鸦	yā	鸭	yā	G	SC		PCPR	PRON	SEM
雅	yā/yǎ	鸦	yā	G		PC	PCPR	PRON	
雅	yā/yǎ	鸭	yā				PCPR	PRON	
讶	yà	雅	yā/yǎ	G		PC	PCPR	PRON	

烟	yān	焰	yàn	G	SC			PRON	SEM
巖	yán	歲	suì	G					
巖	yán	嚴	yán	G		PC	PCPR	PRON	
延	yán	沿	yán/yàn					PRON	SEM
言	yán	信	xìn	G	SC				
掩	yǎn	淹	yān	G	SC	PC	PCPR	PRON	
眼	yǎn	眠	mián	G	SC				
厌	yàn	庆	qìng	G					
焰	yàn	熄	xī	G	SC				
艳	yàn	截	jié						
艳	yàn	绝	jué	G		PC			
艳	yàn	色	sè	G		PC			SEM
验	yàn	脸	liǎn	G		PC			
验	yàn	骗	piàn	G	SC				
殃	yāng	央	yāng	G				PRON	
杨	yáng	扬	yáng	G				PRON	
洋	yáng	样	yàng	G				PRON	
癢	yǎng	残	cán						
妖	yāo	沃	wò	G		PC			
邀	yāo	邊	biān/bian	G	SC				
邀	yāo	变	biàn	G					
摇	yáo	掏	tāo	G	SC				
窑	yáo	坛	tán						SEM
窑	yáo	罈	tán	G					SEM
钥	yào	阴	yīn	G	SC				
钥	yào	银	yín	G	SC				
葉	yè	業	yè	G				PRON	
頁	yè	貝	bèi	G					
壹	yī	壺	hú	G	SC				
壹	yī	壺	hú	G					
壹	yī	臺	tái	G	SC				
壹	yī	喜	xǐ	G					
仪	yí	亿	yì	G	SC		PCPR	PRON	
移	yí	夥	huǒ	G					
倚	yǐ	奇	qí	G		PC	PCPR		
倚	yǐ	依	yī	G	SC			PRON	SEM
椅	yǐ	核	hé	G	SC				
億	yì	仪	yí	G	SC	PC	PCPR	PRON	
役	yì	毅	yì	G				PRON	
役	yì	绎	yì	G				PRON	
忆	yì	议	yì	G				PRON	
意	yì/yi	义	yì			PC	PCPR	PRON	SEM
意	yì/yi	義	yì				PCPR	PRON	SEM
憶	yì	意	yì/yi	G		PC	PCPR	PRON	
憶	yì	疑	yí				PCPR	PRON	
憶	yì	议	yì	G		PC	PCPR	PRON	
抑	yì	毅	yì					PRON	
易	yì	宜	yi/yí	G				PRON	
毅	yì	抑	yì					PRON	SEM
疫	yì	痹	bì	G	SC				SEM
疫	yì	疾	jí	G					SEM
益	yì	易	yì	G				PRON	
绎	yì	绪	xù	G	SC				
绎	yì	择	zé	G		PC			
羿	yì	習	xí	G	SC				
羿	yì	翼	yì	G	SC			PRON	
翼	yì	羿	yì	G	SC		PCPR	PRON	
藝	yì	术	shù						SEM
藝	yì	術	shù						SEM
裔	yì	裳	shang	G	SC				
裔	yì	商	shāng	G					
译	yì	择	zé	G		PC			
谊	yì	议	yì	G	SC	PC	PCPR	PRON	
姻	yīn	婚	hūn	G	SC				
姻	yīn	咽	yàn/yè	G					
陰	yīn	会	huì/kuài	G					
陰	yīn	會	huì/kuài	G					

阴	yīn	绘	huì	G					
吟	yín	玲	líng	G		PC	PCPR		
瘾	yǐn	隐	yǐn	G		PC	PCPR	PRON	
隐	yǐn	稳	wěn	G		PC			
隐	yǐn	穩	wěn	G					
饮	yǐn	炊	chuī	G					SEM
饮	yǐn	欲	yù	G					
應	yīng/yìng	雁	yàn	G		PC			
應	yīng/yìng	鹰	yīng	G		PC	PCPR	PRON	
樱	yīng	桑	sāng	G	SC				SEM
盈	yíng	盛	shèng/chéng	G					
盈	yíng	淫	yín						
荧	yíng	营	yíng	G	SC			PRON	
荧	yíng	灾	zāi	G					
蝇	ying	蝉	chán	G	SC				SEM
蝇	ying	绳	shéng/sheng	G					
迎	yíng	毅	yì						
颖	yǐng	秉	bǐng	G					
映	yìng	央	yāng	G					
映	yìng	阳	yáng	G	SC				SEM
庸	yōng	佣	yōng/yòng	G			PCPR	PRON	
庸	yōng	傭	yōng	G		PC	PCPR	PRON	
拥	yōng	佣	yōng/yòng	G				PRON	
踊	yǒng	桶	tǒng	G		PC	PCPR		
踴	yǒng	躍	yuè	G	SC				SEM
憂	yōu	愛	ài	G	SC				
憂	yōu	慶	qìng	G	SC				
憂	yōu	優	yōu	G		PC	PCPR	PRON	
犹	yóu	尤	yóu	G		PC	PCPR	PRON	
籲	yù	籤	qiān	G	SC				
娱	yú	误	wù/wu	G					
娱	yú	愉	yú					PRON	SEM
榆	yú	桐	tóng	G	SC				SEM
榆	yú	梧	wú	G	SC				SEM
與	yú	興	xìng/xīng	G	SC				
與	yú	御	yù					PRON	SEM
與	yú	禦	yù					PRON	SEM
喻	yù	吁	xū/yù	G	SC	PC	PCPR	PRON	
喻	yù	籲	yù				PCPR	PRON	
域	yù	城	chéng	G	SC				
寓	yù	萬	wàn	G	SC	PC			
寓	yù	宇	yǔ	G	SC	PC	PCPR	PRON	
寓	yù	遇	yù	G	SC	PC	PCPR	PRON	
御	yù	畜	chù/xù						
御	yù	卸	xiè	G		PC			
愈	yù	娱	yú				PCPR	PRON	
愈	yù	愉	yú	G		PC	PCPR	PRON	
浴	yù	溶	róng	G	SC				
浴	yù	裕	yù	G		PC	PCPR	PRON	
熨	yù/yùn	慰	wèi	G					
禦	yù	卸	xiè	G		PC			
裕	yù	祸	huò	G					
豫	yù	橡	xiàng	G		PC	PCPR		
豫	yù	象	xiàng	G		PC	PCPR		
豫	yù	预	yù	G				PRON	
郁	yù	豫	yù					PRON	
冤	yuān	憲	xiàn	G					
圆	yuán	员	yuán	G		PC	PCPR	PRON	
圓	yuán	園	yuán	G	SC	PC	PCPR	PRON	
園	yuán	圓	yuán	G	SC	PC	PCPR	PRON	
援	yuán	缓	huǎn	G					
援	yuán	授	shòu	G	SC				
缘	yuán	豫	yù	G					
怨	yuàn/yuan	愿	yuàn	G	SC			PRON	
嶽	yuè	狱	yù	G					
跃	yuè	沃	wò	G		PC			
阅	yuè	悦	yuè	G				PRON	
匀	yún	均	jūn	G	SC	PC			SEM

允	yǔn	充	chōng	G	SC				
蕴	yùn	蔼	ǎi	G	SC				
蕴	yùn	隘	ài						
蕴	yùn	蓄	xù	G	SC				SEM
運	yùn	遷	qiān	G	SC				SEM
醞	yùn	酿	niàng	G	SC				SEM
醞	yùn	酗	xù	G	SC				SEM
韻	yùn	音	yīn	G	SC				SEM
栽	zāi	裁	cái	G		PC	PCPR	PRON	
栽	zāi	戚	qi	G					
载	zài/zǎi	裁	cái	G		PC	PCPR	PRON	
载	zài/zǎi	截	jié	G		PC			
暂	zàn	哲	zhé	G					
赞	zàn	攒	zǎn	G		PC	PCPR	PRON	
脏	zàng/zāng	胀	zhàng	G	SC				
髒	zāng	體	tǐ	G					
髒	zāng	葬	zàng	G	SC			PRON	
葬	zàng	斃	bì	G		PC	PCPR		SEM
葬	zàng	毙	bì	G		PC	PCPR		SEM
葬	zàng	藏	cáng	G	SC			PRON	
葬	zàng	髒	zāng	G	SC			PRON	
凿	záo	砸	zá						
枣	zǎo	策	cè	G					
枣	zǎo	刺	cì	G					
枣	zǎo	冬	dōng	G					
枣	zǎo	棘	jí	G	SC				
澡	zǎo	燥	zào	G		PC	PCPR	PRON	
灶	zào	灯	dēng	G	SC				
燥	zào	烧	shāo	G	SC				
燥	zào	灶	zào	G	SC		PCPR	PRON	
燥	zào	躁	zào	G		PC	PCPR	PRON	
竈	zào	龜	guī						
责	zé	绩	jī	G		PC			
榨	zhà	窄	zhǎi	G					
榨	zhà	炸	zhà	G				PRON	
诈	zhà	咋	zǎ	G					
宅	zhái	宇	yǔ	G	SC				
窄	zhǎi	咋	zǎ	G					
窄	zhǎi	榨	zhà	G					
债	zhài	责	zé	G		PC			
债	zhài	窄	zhǎi					PRON	
寨	zhài	塞	sāi/sè	G	SC	PC	PCPR		
寨	zhài	栽	zāi	G		PC	PCPR		
粘	zhān	贴	tiē	G	SC	PC			SEM
瞻	zhān	膽	dǎn	G		PC	PCPR		
瞻	zhān	睁	zhēng	G	SC				SEM
崭	zhǎn	耸	sǒng						SEM
斩	zhǎn	崭	zhǎn	G		PC	PCPR	PRON	
斩	zhǎn	折	zhé/zhē	G					SEM
佔	zhàn	估	gū	G	SC				
佔	zhàn	站	zhàn	G	SC	PC	PCPR	PRON	
戰	zhàn	戲	xì	G	SC				
掌	zhǎng/zhang	撑	chēng/cheng	G		PC			SEM
仗	zhàng	杖	zhàng	G		PC	PCPR	PRON	
帐	zhàng	胀	zhàng	G	SC	PC	PCPR	PRON	
帐	zhàng	账	zhàng	G	SC	PC	PCPR	PRON	
杖	zhàng	丈	zhàng	G		PC	PCPR	PRON	
召	zhào	招	zhāo	G		PC	PCPR	PRON	
哲	zhé	暂	zàn	G					
折	zhé/zhē	拆	chāi	G	SC				
折	zhé/zhē	执	zhí	G	SC				
蔗	zhe	芦	lú/lu	G	SC				SEM
蔗	zhe	蘆	lú/lu	G	SC				SEM
侦	zhēn	调	diào/tiáo						
侦	zhēn	贞	zhēn	G		PC	PCPR	PRON	
斟	zhēn	堪	kān	G		PC	PCPR		
斟	zhēn	酌	zhuó						SEM
贞	zhēn	贴	tiē	G	SC				

贞	zhēn	粘	zhān	G					
贞	zhēn	侦	zhēn	G		PC	PCPR	PRON	
诊	zhěn	珍	zhēn	G		PC	PCPR	PRON	
震	zhèn	雷	léi	G	SC				
征	zhēng	微	wēi	G	SC				
癥	zhēng	废	fèi	G					
蒸	zhēng	烹	pēng	G					SEM
蒸	zhēng	征	zhēng	G				PRON	
證	zhèng	登	dēng	G		PC	PCPR		
鄭	zhèng	兆	zhào						
祇	zhǐ	纸	zhǐ	G		PC	PCPR	PRON	
肢	zhī	股	gǔ/gu	G	SC				SEM
肢	zhī	骨	gǔ/gú	G					SEM
肢	zhī	脂	zhī	G	SC	PC	PCPR	PRON	
脂	zhī	腊	là/xī	G	SC				
值	zhí	直	zhí	G		PC	PCPR	PRON	
执	zhí	支	zhī			PC	PCPR	PRON	
职	zhí	值	zhí			PC	PCPR	PRON	
挚	zhì	擎	qíng	G	SC				
挚	zhì	掌	zhǎng/zhang	G	SC				
挚	zhì	执	zhí	G		PC	PCPR	PRON	
滞	zhì	溉	gài	G	SC				
秩	zhì	跌	diē	G		PC			
秩	zhì	稚	zhì	G	SC		PCPR	PRON	
稚	zhì	秩	zhì	G	SC			PRON	
緻	zhì	製	zhì			PC	PCPR	PRON	
置	zhì/zhi	值	zhí	G		PC	PCPR	PRON	
質	zhì	斤	jīn	G					
忠	zhōng	衷	zhōng	G				PRON	
鐘	zhōng	鐵	tiě	G	SC				
腫	zhǒng	種	zhǒng/zhòng	G		PC	PCPR	PRON	
宙	zhòu	庙	miào	G					
宙	zhòu	昼	zhòu					PRON	SEM
晝	zhòu	書	shū	G	SC				
烛	zhú	蚀	shí	G	SC				
築	zhú	鞏	gǒng	G		PC	PCPR		
築	zhú	建	jiàn						SEM
築	zhú	住	zhù					PRON	
嘱	zhǔ	属	shǔ	G		PC	PCPR	PRON	
柱	zhù	枉	wang	G	SC				
专	zhuān	传	chuán/zhuàn	G		PC	PCPR	PRON	
赚	zhuàn	谴	qiǎn			PC	PCPR		
庄	zhuāng	压	yā	G					
庄	zhuāng	桩	zhuāng	G		PC	PCPR	PRON	
桩	zhuāng	脏	zàng/zāng	G		PC			
桩	zhuāng	庄	zhuāng	G		PC	PCPR	PRON	
樁	zhuāng	棒	bàng	G	SC				SEM
桌	zhuō	卓	zhuō	G				PRON	
浊	zhuó	蚀	shí	G	SC				
浊	zhuó	酌	zhuó					PRON	
酌	zhuó	酗	xù	G	SC				SEM
酌	zhuó	斟	zhēn						SEM
仔	zǐ	籽	zǐ	G	SC			PRON	SEM
吝	zī	姿	zī	G		PC	PCPR	PRON	
姿	zī	婆	pó/po	G	SC				
姿	zī	滋	zī			PC	PCPR	PRON	
踪	zōng	蹭	cèng	G	SC				
阻	zǔ	组	zǔ	G				PRON	
鑽	zuān	鐵	tiě	G	SC				
尊	zūn	遵	zūn	G		PC	PCPR	PRON	
遵	zūn	遭	zāo	G	SC				
琢	zuó	啄	zhuó	G					

Bibliography

A Chinese-English Dictionary. 1995. Foreign Language Teaching & Research Press.

Allen, Joseph R. 2008. Why learning to write Chinese is a waste of time: A modest proposal. *Foreign Language Annals* 41(2). 237–251.

Ann, T.K. 1982. *Cracking the Chinese puzzles*. Hong Kong: Stockflow.

Atkinson, Richard C. & Richard M. Shiffrin. 1968. Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation* 2. 89–195.

Balota, David A., Michael J. Cortese, Susan D. Sergent-Marshall, Daniel H. Spieler & Melvin J. Yap. 2004. Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General* 133(2). 283.

Bi, Yanchao, Zaizhu Han & Yumei Zhang. 2009. Reading does not depend on writing, even in Chinese. *Neuropsychologia* 47(4). 1193–1199.

Boltz, William G. 1994. *The origin and early development of the Chinese writing system*. Vol. 78. Eisenbrauns.

Bottéro, Françoise & Christoph Harbsmeier. 2008. The "Shuowen Jiezi" Dictionary and the Human Sciences in China. *Asia Major*. 249–271.

Brown, Gordon D.A. & Frances L. Watson. 1987. First in, first out: Word learning age and spoken word frequency as predictors of word familiarity and word naming latency. *Memory & Cognition* 15(3). 208–216.

Bullinaria, John A. 1996. Connectionist models of reading: Incorporating semantics. In *Proceedings of the First European Workshop on Cognitive Modelling*, 224–229. Citeseer.

Chen, Mengjia 陳夢家. 1988. 殷墟卜辭綜述 [Introduction to the Yinxu Oracular Texts]. Chinese. Zhonghua Book Company.

Chen, Zhiqun. 2009. *Compound ideograph: a contested category in studies of the Chinese writing system*. Monash University. Faculty of Arts. School of Languages, Cultures & Linguistics PhD thesis.

Coltheart, Max & Kathleen Rastle. 1994. Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance* 20(6). 1197.

Cortese, Michael J. & David A. Balota. 2012. Visual word recognition in skilled adult readers. In Michael Spivey, Ken McRae & Marc Joanisse (eds.), *The Cambridge Handbook of Psycholinguistics*, 159–185.

Cortese, Michael J. & Greg B. Simpson. 2000. Regularity effects in word naming: What are they? *Memory & Cognition* 28(8). 1269–1276.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Tech. rep. Cambridge.

Da, Jun. 2004. A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction. In *Proceedings of the fourth International Conference on New Technologies in Teaching and Learning Chinese*, 501–11. Beijing: Tsinghua University Press.

Da, Jun. 2005. Reading news for information: How much vocabulary a CFL learner should know. In *International Interdisciplinary Conference on Hanzi renzhi – How Western learners discover the world of written Chinese*.

Frost, Ram, Leonard Katz & Shlomo Bentin. 1987. Strategies for visual word recognition and orthographical depth: a multilingual comparison. *Journal of Experimental Psychology: Human Perception and Performance* 13(1). 104.

Goodman, Kenneth S. 1967. Reading: A psycholinguistic guessing game. *Literacy Research and Instruction* 6(4). 126–135.

Gough, Philip B. 1972. One second of reading. *Visible Language* 6(4). 291–320.

Grasemann, Uli, Chaleece Sandberg, Swathi Kiran & Risto Miikkulainen. 2011. Impairment and rehabilitation in bilingual aphasia: A SOM-based model. In *International Workshop on Self-Organizing Maps*, 207–217. Springer.

Guder-Manitius, Andreas. 1998. *Sinographemdidaktik*. Heidelberg: Julius Groos Verlag.

Harbaugh, Rick. 1998. *Chinese characters: A genealogy and dictionary*. Yale University Press.

Harm, Michael W. & Mark S. Seidenberg. 2004. Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review* 111(3). 662.

Hayden, Jeffrey J. 2005. Breaking the camel's back: Cognitive load and reading Chinese. In *International Interdisciplinary Conference on Hanzi renzhi – How Western learners discover the world of written Chinese*.

Hebb, Donald Olding. 1949. *The organization of behavior: A neuropsychological approach*. John Wiley & Sons.

Heisig, James W. & Timothy W. Richardson. 2015. *Remembering Simplified Hanzi 1: How Not to Forget the Meaning and Writing of Chinese Characters*.

Hu, Hsueh-Chao Marcella & Paul Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language* 13(1). 403–30.

Huang, Chu-Ren & Keh-jiann Chen. 1992. A Chinese Corpus for Linguistic Research. In *Proceedings of the 14th conference on Computational Linguistics*, vol. 4, 1214–1217. Association for Computational Linguistics.

Jared, Debra. 1997. Spelling–sound consistency affects the naming of high-frequency words. *Journal of Memory and Language* 36(4). 505–529.

Jared, Debra. 2002. Spelling-sound consistency and regularity effects in word naming. *Journal of Memory and Language* 46(4). 723–750.

Jarvis, Scott. 2009. Lexical transfer. In Aneta Pavlenko (ed.), *The bilingual mental lexicon: Interdisciplinary approaches*, 99–124. Multilingual Matters Clevedon.

Katz, Leonard & Ram Frost. 1992. The reading process is different for different orthographies: The orthographic depth hypothesis. *Advances in Psychology* 94. 67–84.

Koda, Keiko. 2005. *Insights into second language reading: A cross-linguistic approach*. Cambridge University Press.

Kohonen, Teuvo. 1989. Self-organising and associative memory. *Springer Series on Information Sciences*.

Kosek, Michał. 2014. *An intelligent tutoring system for learning Chinese with a cognitive model of the learner*. University of Oslo, Department of Informatics. Master's thesis.

LaBerge, David & S. Jay Samuels. 1974. Toward a theory of automatic information processing in reading. *Cognitive Psychology* 6(2). 293–323.

Lam, Ho Cheong 林浩昌. 2011. A critical analysis of the various ways of teaching Chinese characters. *Electronic Journal of Foreign Language Teaching* 8(1).

Longman Advanced Chinese Dictionary, 2nd edn. 2003. Pearson Education Asia Limited & People's Education Press.

Matsumoto, Kazuko. 1987. Diary studies of second language acquisition: A critical overview. *JALT Journal* 9(1). 17–34.

Matthews, Alison & Laurence Matthews. 2007. *Learning Chinese Characters*. Tuttle Publishing.

Matthews, Laurence. 2004. *Chinese Character Fast Finder*. Tuttle Publishing.

McClelland, James L. & David E. Rumelhart. 1981. An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review* 88(5). 375.

McClelland, James L., David E. Rumelhart, PDP Research Group, et al. 1986. *Parallel Distributed Processing*. Vol. 1 and 2. Cambridge, MA: MIT Press.

Miikkulainen, Risto. 1993. *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory*. MIT Press.

Miikkulainen, Risto. 1997. Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language* 59(2). 334–366.

Miikkulainen, Risto & Swathi Kiran. 2009. Modeling the bilingual lexicon of an individual subject. In *International Workshop on Self-Organizing Maps*, 191–199. Springer.

Miller, George A. 1995. WordNet: a lexical database for English. *Communications of the ACM* 38(11). 39–41.

MOEDict 萌典. 2015. Chinese dictionary. Retreived 2016-08-25. Taiwan Ministry of Education. http://www.moedict.tw.

Morton, John. 1969. Interaction of information in word recognition. *Psychological Review* 76(2). 165.

Moser, David. 1991. Why Chinese is so damn hard. *Sino-Platonic Papers* 27. 59–70.

Nation, Kate, M. Spivey, K. McRae & M. Joanisse. 2012. Decoding, orthographic learning and the development of visual word recognition. In Michael Spivey, Ken McRae & Marc Joanisse (eds.), *Cambridge Handbook of Psycholinguistics*.

Outlier Dictionary of Chinese Characters. 2016. Outlier Linguistics Solutions. http://www.outlier-linguistics.com/.

Paap, Kenneth R. & Ronald W. Noel. 1991. Dual-route models of print to sound: Still a good horse race. *Psychological Research* 53(1). 13–24.

Pavlenko, Aneta. 2009. Conceptual representation in the bilingual lexicon and second language vocabulary learning. In Aneta Pavlenko (ed.), *The bilingual mental lexicon: Interdisciplinary approaches*, 125–160. Multilingual Matters Clevedon.

Peereman, Ronald, Alain Content & Patrick Bonin. 1998. Is perception a two-way street? The case of feedback consistency in visual word recognition. *Journal of Memory and Language* 39(2). 151–174.

Peng, Ke. 2016. Chinese as a Foreign Language in K-12 Education. In *Chinese Language Education in the United States*, 123–140. Springer.

Perfetti, Charles A. & Ying Liu. 2006. *Reading Chinese characters: Orthography, phonology, meaning, and the lexical constituency model.*

Ping, Xu & Theresa Jen. 2005. Penless Chinese language learning: A computer-assisted approach. *Journal of Chinese Language Teachers Association* 40(2). 25–42.

Plaut, David C. 1997. Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes* 12(5–6). 765–806.

Plaut, David C., James L. McClelland, Mark S. Seidenberg & Karalyn Patterson. 1996. Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review* 103(1). 56.

Powell, Daisy, David Plaut & Elaine Funnell. 2006. Does the PMSP connectionist model of single word reading learn to read in the same way as a child? *Journal of Research in Reading* 29(2). 229–250.

Pulleyblank, Edwin G. 1991. *Lexicon of Reconstructed Pronunciation: in Early Middle Chinese, Late Middle Chinese, and Early Mandarin*. UBC Press.

Qiu, Xigui 裘锡圭. 2000. *Chinese writing*. The Society for the Study of Early China & The Institute of East Asian Studies.

Rumelhart, David E. 1994. *Toward an interactive model of reading.* International Reading Association.

Samuels, S. Jay. 1994. Toward a theory of automatic information processing in reading, revisited. In Robert B. Ruddell, Martha Rapp Ruddell & Harry Singer (eds.), *Theoretical models and processes of reading*, 4th edn., 816–837. Newark, DE, US: International Reading Association.

Sandak, Rebecca, Stephen J. Frost, Jay G. Rueckl, Nicole Landi, W. Einar Mencl, Leonard Katz & Kenneth R. Pugh. 2012. How does the brain read words. In Michael Spivey, Ken McRae & Marc Joanisse (eds.), *The Cambridge Handbook of Psycholinguistics*. Cambridge University Press New York, NY.

Schmidt, Richard & Sylvia Frota. 1986. Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In *Talking to learn: Conversation in second language acquisition*, 237–326.

Seidenberg, Mark S. 2012. Computational models of reading. In Michael Spivey, Ken McRae & Marc Joanisse (eds.), *The Cambridge Handbook of Psycholinguistics*, 186. Cambridge University Press.

Seidenberg, Mark S. & James L. McClelland. 1989. A distributed, developmental model of word recognition and naming. *Psychological Review* 96(4). 523.

Seidenberg, Mark S., Gloria S. Waters, Marcia A. Barnes & Michael K. Tanenhaus. 1984. When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior* 23(3). 383–404.

Selfridge, O.G. 1959. Pandemonium: A paradigm for learning. In *The mechanism of thought processes*. National Physical Laboratory, Teddington, England. London: Her Majesty's Stationery Office.

Sharoff, Serge. 2006. Creating general-purpose corpora using automated search engine queries. *WaCky! Working papers on the Web as Corpus*. 63–98.

Sproat, Richard William. 2000. *A computational theory of writing systems*. Cambridge: Cambridge University Press.

Stanovich, Keith E. 1980. Toward an interactive-compensatory model of individual differences in the development of reading fluency. *Reading Research Quarterly*. 32–71.

State Council of the People's Republic of China 中華人民共和國國務院. 2013. 通用規範漢字表 [Table of General Standard Chinese Characters]. Chinese. Tech. rep. http://www.gov.cn/gzdt/att/att/site1/20130819/tygfhzb.pdf.

Sturgeon, Donald. 2011. *Chinese Text Project*. http://ctext.org.

Tang, Lan 唐兰. 1979. 中国文字学 [The study of Chinese characters]. Chinese. Shanghai Guji Chubanshe.

Taraban, Roman & James L. McClelland. 1987. Conspiracy effects in word pronunciation. *Journal of Memory and Language* 26(6). 608–631.

Tracey, Diane H. & Lesley Mandel Morrow. 2012. *Lenses on reading: An introduction to theories and models*. Guilford Press.

Wang, Li 王力. 2010. 中国语言学史 [History of Chinese linguistics]. Chinese. Zhonghua Book Company.

Woźniak, Piotr A. & Edward J. Gorzelańczyk. 1994. Optimization of repetition spacing in the practice of learning. *Acta Neurobiologiae Experimentalis* 54. 59–62.

Xing, Hongbing, Hua Shu & Ping Li. 2002. A self-organizing connectionist model of character acquisition in Chinese. In *Proceedings of the Twenty-Fourth Annual Conference of the Cognitive Science Society*, 950–955.

Xing, Hongbing, Hua Shu & Ping Li. 2004. The acquisition of Chinese characters: Corpus analyses and connectionist simulations. *Journal of Cognitive Science* 5(1). 1–49.

Yin, John Jing-hua 印京华. 2006. *Fundamentals of Chinese characters*. Yale University Press.

Zevin, Jason D. & Mark S. Seidenberg. 2004. Age-of-acquisition effects in reading aloud: Tests of cumulative frequency and frequency trajectory. *Memory & Cognition* 32(1). 31–38.

Zhang, Qi & Ronan G. Reilly. 2015. Writing to read: the case of Chinese. In *29th Pacific Asia Conference on Language, Information and Computation*, 341–350.

Zhao, Cheng 趙誠. 2005. 甲骨文字學綱要 [A Study of the Oracle Bone Script]. Chinese. Zhonghua Book Company.

Zhao, Xiaowei & Ping Li. 2009. An online database of phonological representations for Mandarin Chinese. *Behavior Research Methods* 41(2). 575–583.

Zhou, Xiaolin & William Marslen-Wilson. 2000. The relative time course of semantic and phonological activation in reading Chinese. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 26(5). 1245.

Zorzi, Marco, George Houghton & Brian Butterworth. 1998. Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance* 24(4). 1131.