The Acquisition of Chinese Characters: Corpus Analyses and Connectionist Simulations

Hongbing Xing[1], Hua Shu[2], and Ping Li[3]

[1]*Beijing Language and Culture University, Beijing Normal University*
[2]*State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University*
[3]*University of Richmond*
pli@richmond.edu

Empirical evidence has accumulated that regularity and consistency in orthography-to-phonology correspondence both affect the processing and acquisition of Chinese characters, and that they interact with character frequency in complex ways. However, despite growing interests in the Chinese orthography, few previous studies have systematically analyzed these effects on a large scale, or have modeled the acquisition process using connectionist networks. In this study, we set out to analyze a realistic character corpus as learned by children in Chinese elementary school, in order to evaluate the degree to which corpus analysis can inform us of the roles that regularity, consistency, frequency, and their interactions play in children's acquisition of Chinese characters. We further modeled character acquisition in a self-organizing connectionist network, developing orthographic representations on the basis of our analysis of character properties in a large-scale character corpus. Our model is able to faithfully capture the orthographic similarities of Chinese characters, and moreover, display effects of regularity, consistency, character frequency, and the complex interactions among them, matching up well with available empirical evidence on children's acquisition of characters during the elementary school years.

How do children acquire Chinese characters with its logographic orthography? Chinese characters distinguish themselves from alphabetic letters in that the basic orthographic units map mostly onto meaningful morphemes

Journal of Cognitive Science 5 : 1 - 49, 2004.

rather than spoken phonemes, and therefore a comparison between Chinese and other written languages helps to identify both language-universal and language-specific aspects in the processing and acquisition of orthography. Because characters and meanings correspond in a phonologically non-transparent way in Chinese, a conventional wisdom is that Chinese characters must be learned by rote memory. Empirical psycholinguistic studies, however, have led us to reject this misconception (see Hu & Catts, 1998; Leong, in press; McBride-Chang & Zhong, in press; Shu & Anderson, 1998; Shu, Anderson, & Wu, 2000; Siok & Fletcher, 2001; Yang & Peng, 1997). These studies indicate that the awareness of the phonological structure of words and morphemes is important in the acquisition of Chinese characters.

In spite of growing interests in the Chinese orthography and its acquisition, there have been few systematic, large-scale, analyses of the properties of the characters that children acquire over time, and there has been no research in the modeling of the acquisition process, in contrast to what has been done in English (see review below). In particular, few empirical studies have systematically examined the complex ways in which regularity, consistency, and frequency interact in the acquisition of Chinese characters, effects that are shown to affect the acquisition of alphabetic writing systems. The present study attempts to fill this gap. Our particular focuses are to examine corpus characteristics in textbooks learned by school children and to incorporate such characteristics into computational models that can further inform us of the acquisition processes.

Properties of Chinese Characters and Chinese Phonograms

The first salient visual feature to a foreign eye is that Chinese characters are square shaped. In contrast to English-speaking children's letter-exercise books with lines, Chinese pupils' character-exercise books are made of pages of grids, and the model characters are displayed within the grids — each stroke of the character should be aligned exactly relative to the four sides. Much of the elementary school years of the child is spent on "crawling through the grids", i.e., practicing the writing of Chinese characters.

Despite years of practice, a Chinese-speaking child by the end of the elementary school years would still have acquired only a small portion

(approximately 3,000) of the large number of characters (an estimate of up to 54,600 according to the *Hanyu Da Zidian*, a comprehensive Chinese character dictionary, Hanyu Da Zidian Commission, 1990).[1] Chinese characters are complex, with respect to both the internal structure of each of the components that make up a character, and the structural hierarchies in which these components are assembled into a character (see later discussion for details). In addition, the relationships between sound and meaning in characters are also complex: although many characters are derived from pictographs, in modern Chinese a large portion of characters are the so-called phonograms that do relate to the character's pronunciation in some way. Phonograms are typically made of two parts (often called "radicals"), the phonetic part (simply called "phonetic" below) guiding the pronunciation of the character and the semantic part indicating the categorical meaning of the character. However, the phonetic part provides no reliable pronunciations for the character, and the semantic part provides only an approximation to the semantic category to which the character might be related.

Nevertheless, the appearance of phonograms is an important landmark in the history of Chinese characters and it has significant implications for the learning of characters (Peng & Jiang, in press). Modern Chinese has moved from an early dominance of pictographs (such as 田 "fields") to a dominance of phonograms in modern eras. There are 5,631 phonogram characters, accounting for 81% (Li & Kang, 1993) of the total 7,000 frequent characters (National Language Commission of China, 1989). According to Shu, Chen, Anderson, Wu, and Xuan's (2003) analysis of the "School Chinese Corpus" (which contains 2,570 characters used in the elementary school textbooks in Beijing), about 74% of the Chinese characters taught in elementary schools are phonograms. Given the prominence of phonograms and their representativeness in Chinese orthography, it is important for us to understand the functions of these characters and the processes by which they are acquired. In this paper, we focus on children's acquisition of phonogram characters.

Although the phonetic of a phonogram provides no reliable pronunciations

[1] However, the 3000 most frequent characters account for nearly 99% of all characters commonly used in daily life, according to one estimate (Sun, 1998).

for the whole character, it may relate to the pronunciation of the character in one of three ways: (a) Regular: the whole character is pronounced the same as the phonetic in isolation, that is, the same as the phonetic when it is being used as an independent character - note that many phonetic parts can act as independent characters by themselves; for example, "清/*qing1*/" and "青/*qing1*/". (b) Semi-regular: the whole character is pronounced partly as the phonetic, with a different tone (e.g., "请/*qing3*/" and "青/*qing1*/"), a different onset (e.g., "晴/*jing1*/" and "青/*qing1*/"), or a different final (e.g., "沙/*sha1*/ "and "少/*shao3*/"). (c) Irregular: the whole character is pronounced completely differently from the phonetic (e.g., "猜/*cai1*/" and "青/*qing1*/"). These patterns of regularities or irregularities in the pronunciations of phonograms influence the recognition and processing of Chinese characters, a phenomenon known as the *regularity effect* in the literature.[2]

A related phenomenon in the processing of Chinese characters is the so-called *consistency effect*. Consistency here refers to the degree of consistency in the pronunciation of the group of characters that share the same phonetic component. There are basically two possibilities: (a) Consistent: all characters that share the same phonetic are pronounced the same as the phonetic itself (e.g., 惶, 蝗, 煌, 隍, are all pronounced as /*huang2*/); (b) Inconsistent: characters that share the same phonetic could be pronounced differently, with some taking the pronunciation of the phonetic, while others not (e.g., 清 /*qing1*/, 猜/*cai1*/). Within this latter category, the degree of inconsistency can also vary, such that in some cases only a single character or a few characters deviate from the pronunciation of the majority but in other cases each character may have a different pronunciation (see later discussion on computing different degrees of character consistency). These patterns of consistency or inconsistency in the pronunciations of phonograms also influence the recognition and processing of Chinese characters.

Regularity and consistency of character pronunciations, along with the frequency of use of characters, are the three major variables that have been

[2] This regularity effect is different, but comparable to the regularity effect studied in alphabetic languages. Regularity in those languages typically refers to the degree of consistency in the grapheme-to-phoneme correspondence.

repeatedly examined in the adult psycholinguistics context. In a pioneering work, Seidenberg (1985) found that the phonetic in Chinese characters provides cues to help identify a character's pronunciation, as compared with non-phonetic in characters. This facilitation, however, appeared to hold only for low-frequency characters but not for high-frequency characters. Seidenberg interpreted this as suggesting that high-frequency characters are recognized on a whole-character basis, while low-frequency characters are processed through individual components in smaller units. Other studies that investigated the interaction between regularity, consistency, and frequency include Fang, Horng, and Tzeng (1986), Hue (1992), Yang and Peng (1997), Perfetti and Zhang (1991), Shu and Zhang (1987) (see Peng & Jiang, in press, for a summary). But how do these variables impact on children's acquisition of characters in Chinese? Relatively fewer studies have been designed to directly address this question. In what follows, we first provide an overview of previous research relevant to this question, and then present our corpus analysis on the effects of regularity, consistency, and frequency in character acquisition.

Effects of Regularity, Consistency, and Frequency: Corpus Analyses

Previous Empirical Research

Several empirical studies have examined regularity effects in children's acquisition of Chinese characters. Shu, Anderson, and Wu (2000) showed that children in Grades 2, 4, and 6 display regularity effects when they are required to write down the pronunciations of Chinese characters: they perform better on regular characters whose pronunciation is the same as their phonetic in isolation, than on irregular characters whose pronunciation is different from their phonetic. When children see unfamiliar characters, they often exploit the pronunciation of the phonetic as a possible reading of the whole character, and this ability increases with school grade (see also Shu and Wu, in press). Ho and Bryant (1997) reported similar findings for Grade 1 and 2 children in Hong Kong. Yang and Peng (1997) also found regularity effects in children's speed of naming characters: children in Grade 3 name regular characters more rapidly than irregular characters, but by Grade 6, they name both types of characters equally quickly. Frequency also plays an important role in children's character naming as it interacts with the regularity of characters. For example,

children display smaller regularity effects on characters of high frequency but show larger regularity effects on characters of low frequency (Shu, Anderson, & Wu, 2000).

In addition to regularity effects, researchers have also examined consistency effects in Chinese children's acquisition of characters. Tzeng, Zhang, Hung, and Lee (1995) studied children in Taiwan by asking them to read three types of pseudo-characters: those with only regular neighbors, those with only irregular neighbors, and those with both regular and irregular. Children were found to make more regular responses in the regular-only condition than in the irregular-only condition. The results suggest that children do not simply name a character based on its phonetic, but also take into consideration the pronunciations of other characters sharing the same phonetic (i.e., consistency of the neighbors). Yang and Peng (1997) found consistency effects in the naming of characters by both Grade 3 and Grade 6 students, suggesting that as vocabulary increases, more and more characters in the repertoire share the same components, which influences children's processing of character reading. This finding was further confirmed in a study by Shu, Zhou and Wu (2000) who showed that young children develop phonological awareness of the structures of characters and the functions of the phonetic and the semantic radical. Some 4th graders already start to acquire the awareness of the consistency of the phonetic, and by Grade 6 this awareness becomes stronger. The sensitivity to character consistency continues to increase until the college level, according to Shu et al. (2000).

Most of the above-mentioned studies of character acquisition are experimental in nature, looking at a very selective set of characters in a highly controlled manner. While this type of experimental research is undoubtedly important in unraveling the influences of regularity, consistency, and frequency in character acquisition, it falls short of a complete picture of the thousands of characters that are actually learned by children in the early school years. In this study, we have taken on the task of building a corpus that includes all the characters in elementary school textbooks, and have conducted our analyses on the basis of this corpus. In what follows, we discuss the methods used to build the corpus, and the detailed analyses of phonograms in the corpus with respect to regularity, consistency, and frequency.

Method

The elementary school textbooks for first through sixth grades (Beijing Academy of Educational Sciences, 1998) are the primary source of our corpus analyses. In adopting these textbooks as the basis of our analysis, the present corpus study differs from previous corpus studies in significant ways. In previous studies, only character type information has been considered (without character token frequency), and researchers usually use character frequency based on adult corpus as an approximation to character frequency used by children (a situation parallel to the use of the Kucéra-Francis word frequency to predict vocabulary acquisition in children, which does not work well; see Goodman, Dale, & Li, 2002). In this study, we constructed our corpus based on both type and token information of characters in the school textbooks used in classroom teaching. Thus, the information we get from this corpus (including character frequency and distribution of character regularity and consistency) provides a more realistic measure of children's knowledge, which is also important for simulations as seen in our modeling study reported below. Our study can also be compared with previous studies in terms of the distributions of regularity, consistency, and other features of Chinese characters by the comparison of child-based and adult-based corpora.

The total number of characters (tokens) in the Beijing textbooks is 160,342, and the number of characters (types) is 3306. In other words, an elementary school student will learn an average of 3306 characters during the six elementary school years, covering about 98.64% of the total characters commonly used (e.g., in a 100-million word adult corpus; Sun, 1998). The characters that appear in these textbooks were tagged with respect to the following dimensions (Shu, Chen, Anderson, Wu, & Xuan, 2003): character configuration, character pronunciation, character frequency, age (grade) of acquisition of character, radical (phonetic) configuration, pronunciation, and position in the character, phonetic type (whether the character has the same pronunciation as the phonetic), and character consistency (whether all characters taking the same phonetic have the same pronunciation). Such information, especially frequency, phonetic type, and consistency, is valuable to our understanding of the factors that affect children's acquisition of characters.

Results

Phonogram Profiles

The total number of phonograms (types) in the corpus is 2477, accounting for 75% of all character types used in the corpus, and the total number of phonogram tokens is 77,729, accounting for 48% of all character tokens used in the corpus. First, we counted the number of phonograms for each school grade, and computed the proportion of new phonogram characters that students would learn in each grade. Second, we counted the number of phonograms relative to the total number of characters that students used in each grade. Table 1 and Table 2 present the results of this analysis.

Table 1 shows that the total number (type) of new characters and new phonograms for each grade. It can be seen that although the absolute number of new characters and phonograms decreases after Grade 3, the proportion of phonograms relative to the number of new characters linearly increases across school grades. Table 2 also shows that although the total number of

Table 1. Number of new characters and phonograms for each school grade

	1st	2nd	3rd	4th	5th	6th
New characters	667	697	759	441	410	332
Phonograms	400	480	586	376	349	286
Proportion	.60	.69	.77	.85	.85	.86

Table 2. Types and tokens of characters and phonograms for each grade (cumulative)

	1st	2nd	3rd	4th	5th	6th
Token						
Number of characters	3910	11485	21369	26295	41214	56069
Number of phonograms	1801	5725	10350	12782	19846	27070
Proportion	.46	.50	.48	.49	48	.48
Type						
Number of characters	667	1260	1904	2101	2375	2630
Number of phonograms	400	803	1281	1457	1673	1870
Proportion	.60	.64	.67	.69	.70	.71

phonograms (tokens) remains constant relative to the total number of characters (around 48-50%), the types of phonograms relative to the types of characters steadily increase. These two tables indicate clearly that phonograms play an increasingly important role in character learning as the child progresses through elementary school.

Frequency of Phonograms

As seen in Table 1 and Table 2, both the proportion of new phonograms and the proportion of phonogram types increase with school grades. The number of non-phonograms (types), then, is relatively small. However, the number of non-phonogram tokens is more than half of the total character tokens used in each grade (around .52). Thus, the low type frequency and high token frequency of non-phonograms shows that it is important for us to have a clear picture of the frequency of characters used in the textbook corpus. We divided the frequency of characters in the Beijing textbook into five categories, according to the following criteria: high (character use ≥ 50 in the corpus, or ≥ 300 per million), medium-high (20-49, or 120-300 per million), medium (10-19, or 50-120 per million), medium-low (3-9, or 15-50 per million), and low (≤ 2, or ≤ 15 per million).

Fig. 1 shows the frequency profile of the characters learned in each grade. It

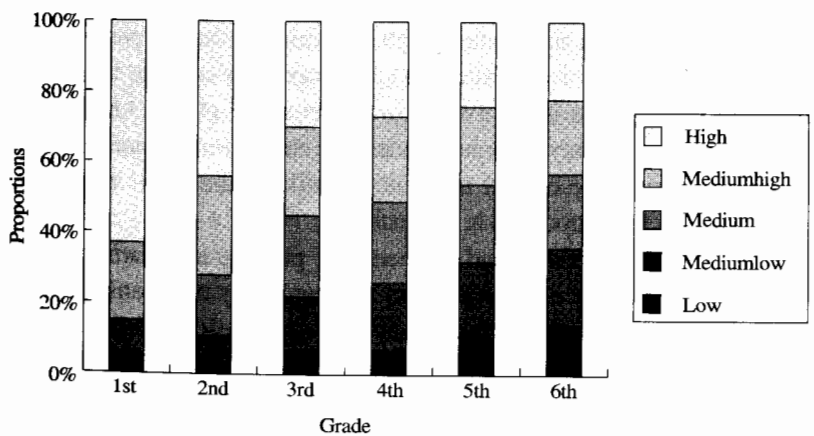

Fig 1. Frequency profile of the characters learned in each grade

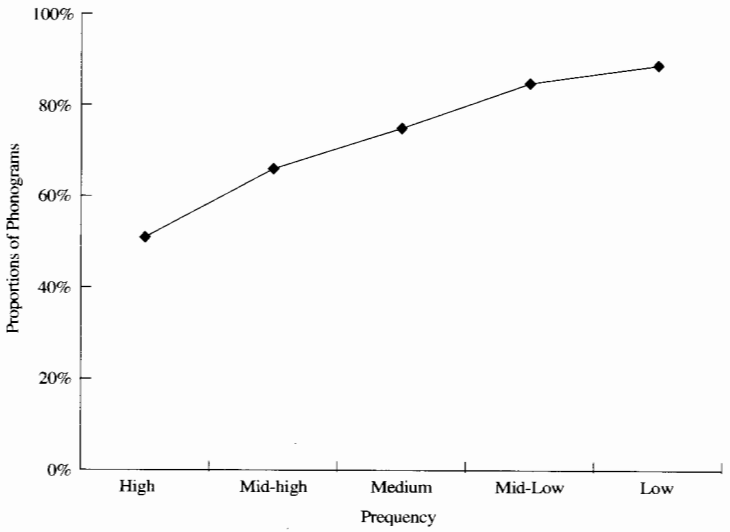

Fig. 2. Proportion of phonograms in different frequency range of characters

indicates that the characters children learn in lower grades are of relatively higher frequency, whereas in later grades they meet more and more lower frequency characters in the textbooks. Fig. 2 indicates that more phonograms are in the low to medium frequency categories, with relatively fewer phonograms in the high and medium-high categories. Given the predominance of phonograms (75% of all characters), it comes as a surprise that a large number of phonograms is relatively low in frequency of use. On the other hand, this means that only in higher grades will children meet more and more phonograms, given the overall frequency profile of the high- versus low-frequency characters introduced in the textbooks.

In addition to the number and frequency of phonograms learned in each grade, school children are sensitive to the regularity and consistency of characters in the acquisition of phonograms. We discuss each of these below.

Character Regularity

In the discussion of the properties of Chinese characters, we divided characters into three categories with respect to regularity: (a) regular, in which

the whole character is pronounced the same as the phonetic in isolation, (b) semi-regular, in which the whole character is pronounced partly as the phonetic, and (c) irregular, in which the whole character is pronounced completely differently from the phonetic (see examples discussed earlier). This last category also includes phonograms that have multiple pronunciations, and phonograms that have lost their phonetic cues when simplified.

Table 3 and Fig. 3 present the relative distribution of these three categories, and their distribution in each grade according to the use of these characters in the corpus. It shows that the percentage of regular characters increases but the percentage of irregular characters decreases across school grades. The percentage of semi-regular characters remains roughly the same across grade. These contrasting patterns for regular versus irregular characters indicate that as children get older they are exposed to more regularities in the character, thus promoting their awareness of the orthography-to-phonology correspondences

Table 3. Number and proportion of regular, semi-regular and irregular characters

Regularity	Regular	Semi-regular	Irregular
Number of characters	619	1017	841
Proportion of characters	.25	.41	.34

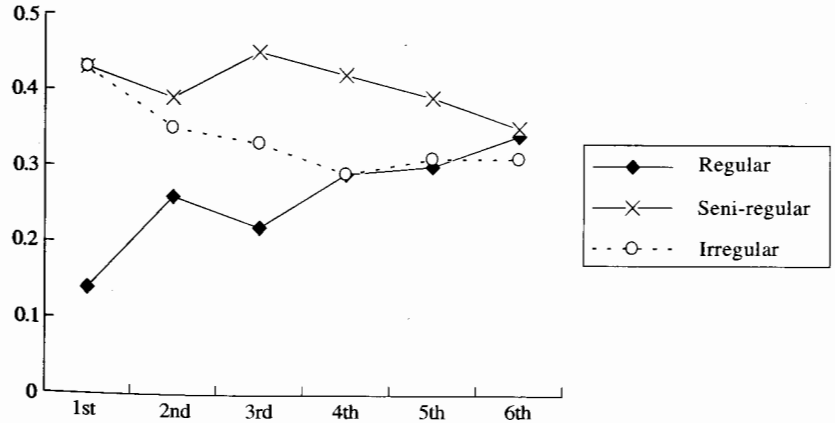

Fig. 3. Proportion of regular, semi-regular and irregular characters by grade

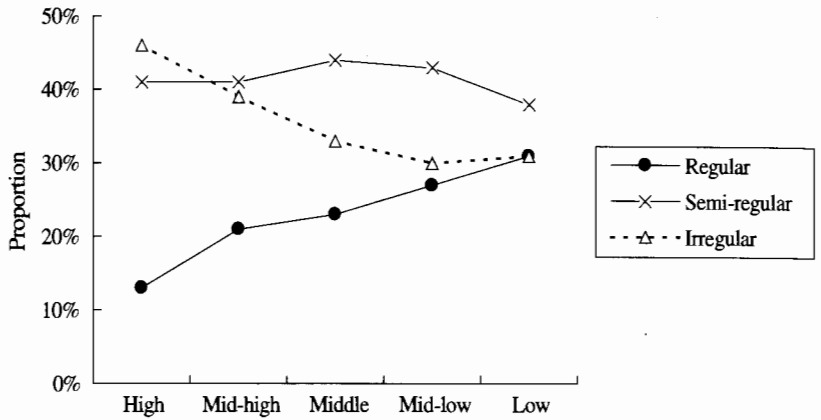

Fig. 4. Proportion of regular, semi-regular and irregular characters by frequency

(Shu & Wu, in press).

A number of studies have found that regularity also interacts with frequency in the processing and recognition of Chinese characters (Seidenberg, 1985; Perfetti & Zhang, 1991; Peng & Jiang, in press). In particular, regularity is transparent for low-frequency characters, but less so for high-frequency characters. How is this interaction reflected in the corpus? We analyzed the relative proportion of the three categories of regular and irregular characters with respect to their frequency of use for each grade, as shown in Fig. 4.

Fig. 4 confirms the interaction between regularity and frequency in our corpus. A relatively smaller proportion of regular phonograms are in the high-frequency range, but the proportion increases as frequency decreases. In contrast, irregular and semi-regular phonograms show no such tendency for high versus low frequency. In fact, there are more irregular phonograms in high frequency than in low frequency. These patterns are consistent with the results in Fig. 3, in that early on children are unlikely to be highly sensitive to the regularity of characters, owing both to the relatively small number of regular phonograms and to the relatively low frequency of these characters.

Character Consistency

As discussed earlier, Chinese characters can be classified into two categories

with respect to the consistency of pronunciation: (a) consistent, in which all characters that share the same phonetic are pronounced the same as the phonetic itself, and (b) inconsistent, in which characters that share the same phonetic are pronounced differently. Character consistency is tied to the notion of character family. We define a set of characters as in the same "family" if they share the same phonetic (including the phonetic itself as an independent character). For example, 惶*/huang2/* belong to a consistent family because all characters in this family take the same phonetic 皇(凰, 惶, 蝗, 煌, 隍, 徨, 湟, 遑, 篁, 鲷) and have the same pronunciation as the phonetic (*/huang2/*). By contrast, the character 抬*/tai2/* belongs to an inconsistent family because all characters in this family take the phonetic 饴台, but have different pronunciations (抬*/tai2/*, 胎*/tai1/*, 苔*/tai2/*, 邰*/tai2/*, 跆*/tai2/*, 鲐*/tai2/*, 驼*/tai2/*, 诒*/yi2/*, */yi2/*, 怡*/yi2/*, 贻*/yi2/*, 始*/shi3/*, 治*/zhi4/*, 冶*/ye3/*).

In elementary school, the number of character families to be learned differs from grade to grade, and therefore the same family may contain different numbers of members for children in different grades. Thus, our corpus analysis, in contrast to other empirical research, counts the consistent families or members only for children in specific grade, not for children in all grades. For example, the character "清" */qing1/* may be a consistent character for Grade 1 children, because the children so far have learned only the characters "清*/qing1/*" and its phonetic "青/qing1/". However, the same character becomes an inconsistent character for Grade 2 children, because they now have learned more characters like "猜*/cai1/*", and "请*/qing3/*".

Table 4 shows that as school grades increase, the number of members in each phonogram family (maximum or mean) also increases, as well as the absolute number of character families across grades. The "maximum number of members" refers to the characters families that have the largest members in

Table 4. Number of phonogram families in each grade

Grade	1st	2nd	3rd	4th	5th	6th
Family	123	287	478	566	632	687
Max. members	7	9	12	12	14	17
Mean members	2.49	2.97	3.23	3.49	3.76	3.94

one family, and the "mean number of members" refers to the average family sizes of all the character families.

Although the number of character families and the members in a family both increase, a detailed analysis of the consistent character families indicates that awareness of character consistency may be a slowly developing process. Table 5 shows the number of consistent and inconsistent groups with the character families for each grade. By Grade 6, the number of consistent families is 108, accounting for only 16% of the total number of character families (687), and the majority of these families have only two or three members. This result differs from that of regularity - in the regularity case, regular and semi-regular characters account for nearly 70% of the phonograms to be learned by Grade 6 (cf. Fig. 3).

It can be seen from Table 5 that both consistent and inconsistent families increase with grade. However, the absolute number of inconsistent families is much larger than that of consistent families. This is because, as mentioned earlier, as higher-grade children learn more characters, some consistent families for earlier grades become inconsistent families for later grades. Thus, elementary school children's sensitivity to consistency may only gradually develop. However, children may begin with a hypothesis that assumes that all characters are consistently pronounced as the phonetic, especially when they have learned only one or two characters in a family. Thus, earlier on they tend to overgeneralize the use of the phonetic across all members of a family, as if they had knowledge of character consistency (Shu & Wu, in press).

To summarize our corpus study, our analysis shows that there are many inherent statistical properties of the characters that children are faced with in the elementary school textbooks. These statistical properties would in many cases promote children's awareness to regularity, consistency, and frequency effects in the characters. However, such awareness does not develop

Table 5. Consistency in character families for each grade

Grade	1st	2nd	3rd	4th	5th	6th
Consistent families	14	40	56	83	94	108
Inconsistent families	109	247	422	483	538	579
Total families	123	287	478	566	632	687

automatically, or uniformly, but rather, they depend on the specific dimension of properties and the interaction of these properties with each other and with the learning system. In what follows, we discuss how a learning system, our connectionist model, may handle the task of learning the kinds of characters that Chinese-speaking children encounter in elementary school, and how the system can become sensitive to a number of factors that influence the learning process, including regularity, consistency, and frequency of the characters.

Effects of Regularity, Consistency, and Frequency: Connectionist Modeling

Connectionist Models of Reading in English and Chinese

Our corpus analysis of the profiles of Chinese phonograms learned by elementary school children shows a number of important factors that affect children's acquisition of characters. But how does the learner incorporate these factors in the process of acquisition? In other words, although the corpus analysis shows the properties of the learning environment, we need to ask about properties of the learner — in particular, what computational properties allow the learner to handle the learning environment and display specific learning patterns in the process of character acquisition? In this section, we present a connectionist model that can adequately address this issue.

Most previous connectionist models of reading have examined alphabetic languages (particularly English), using feed-forward networks with the back-propagation learning algorithm. Seidenberg and McClelland (1989) first studied word reading for English monosyllabic words with a connectionist network. Using distributed representations of word knowledge, they were able to simulate word-reading processes from orthography to phonology. The model succeeded in reading both regular and exception words in the orthography-to-phonology correspondence, relying on a single connectionist mechanism rather than two separate mechanisms using "dual routes" (phonological route for regular words vs. lexical route for exceptions; Coltheart, 1978). Their model accounted for a number of classical word reading effects, including word frequency, spelling-sound consistency, and the interaction between the two. Building on the Seidenberg and McClleland model, Plaut, McClelland, Seidenberg, and Patterson (1996) examined word

reading in normal and impaired situations. They pointed out that the original model performed poorly on the reading of pronounceable nonwords (e.g., mave), and to account for nonword as well as word reading. Plaut et al. (1996) developed orthographic and phonological representations that better capture the structures of written and spoken forms of words. The new model was able to read regular and exception words as the original Seidenberg and McClleland model, but it also performed well on nonwords. Following these two models, several other connectionist models of orthographic processing further expanded the investigation into reading impairment (phonological and surface dyslexia; e.g., Harm & Seidenberg, 1999) and the relationships between orthography, phonology, and semantics in word reading (Harm & Seidenberg, in press).

A common feature to the previous models is that they use empirical data from normal adults as the basis (and target) for their modeling (e.g., word frequency estimates from adult norms). Such an approach works well for modeling word reading in adults, but would be problematic for modeling orthographic acquisition in children. Children acquire words incrementally, such that their vocabulary gradually increases along with age/grade. In the case of Chinese, orthographic acquisition of characters goes hand-in-hand with an increasing vocabulary. A model of children's word reading needs to take into account this property. In this study, we build our connectionist model on the basis of our corpus analysis of the characters and their properties in the textbooks that children learn from. Our approach thus should provide a better approximation to the learning processes than a method based on adult corpus analyses.

In contrast to previous models, our study also uses the "unsupervised" learning algorithm. It is based on the architecture of self-organizing neural networks, in particular, self-organizing maps (SOM; Kohonen, 1989, 1995). A number of recent models have explored SOMs as viable models of language representation and language acquisition, in both monolingual and bilingual situations (Hernandez, Li, & MacWhinney, 2005; Miikkulainen, 1993, 1997; Li, 2003; Li & Farkas, 2002; Li, Farkas, & MacWhinney, 2004). Although significant progress has been made with previous models based on back-propagation (as shown by the word reading models reviewed above), there have been known limitations associated with these models as realistic models

of human cognition, in particular, as models of language acquisition (see Li, 2003; Li, Farkas, & MacWhinney, 2004; MacWhinney, 2001 for discussion). SOM relies on unsupervised learning algorithms by which learning proceeds without explicit error-correcting signals as in back-propagation. Learning in SOM is achieved by the system's self-organization in response to the input. During learning, the self-organizing process extracts an efficient and compressed internal representation from the high-dimensional input space and projects this new representation onto a 2-D topological structure (Kohonen, 1989, 1995). Several important properties of SOMs and related features make such networks particularly well suited for our investigation.

Self-organization in these networks typically occurs in a 2-dimensional topographical map, where each processing unit in the network is a location on the map that can uniquely represent one or several input patterns. At the beginning of learning, an input pattern randomly activates a set of units on the map that surrounds the best matching unit (the winner), according to how similar by chance the input pattern is to the weight vectors of the units. Once these units become active in response to a given input, the weights of the winner and those of its neighboring units are adjusted such that they become more similar to the input and these units will therefore respond to the same or similar inputs more strongly the next time. Initially activation occurs in large areas of the map (i.e., many units are active), but gradually learning becomes focused so that only the maximally responding unit or units are active. This process continues until all the inputs can elicit specific response patterns in the network. As a result of this self-organizing process, the network gradually develops concentrated areas of units on the map (the so-called "activity bubbles") that capture input similarities, and the statistical structures implicit in the high-dimensional space of the input are projected to and preserved on a 2-D space in the map.

In an attempt to model the mental lexicon, Miikkulainen (1993, 1997) connected several SOMs through associative links trained by Hebbian learning, where each SOM is dedicated to a specific type of linguistic information (orthography, phonology, or semantics). Hebbian learning is a biologically plausible learning principle, according to which the connection strength between two units is increased if the units are both active at the same time (Hebb, 1949). In Miikkulainen's (1997) model (the DISLEX model), all

units on one map are initially connected to all units on the other map. As self-organization takes place, the associations become more focused, such that in the end only the maximally active units on the corresponding maps are associated. Hebbian learning combined with SOM has strong implications for orthographic acquisition: it can account for the process of how the learner establishes relationships between phonological forms and orthographic forms of words, on the basis of how often the two co-occur and how strongly they are co-activated in the representation.

Although word reading and orthographic processing have been carefully examined in English and other languages, due to the difficulty in representing the complex structure of the Chinese orthography, very little modeling research has been done for Chinese. There have been only two preliminary attempts so far. First, Chen and Peng (1994) proposed a connectionist model of recognition and naming in Chinese, using a standard feed-forward network. Their model consisted of orthographic representations at the input layer, mapped to phonological representations at the output layer. Orthographic representations of Chinese characters focused on radical components and their structural relationships in the character. A major success of the model was its ability to show character frequency effects and to distinguish regular from irregular characters in naming. A second, more recent model, the interactive constituency model, was proposed by Perfetti and Liu (in press). Perfetti and Liu were interested in building a more general model of reading in Chinese rather than modeling specific effects in naming. Their model included four different levels of interactive constituency: radical, orthography, phonology, and semantics. The input units were 144 radicals that begin activation in orthography, which then activate phonology or semantics of the characters. One interesting pattern from their model was the oscillation effect: the onset of inhibition of orthographically similar primes coincides with the onset of facilitation of phonological priming. This effect matches with empirical observations on the time course of orthographic and phonological priming (Perfetti & Tan, 1998).

Although both models attest to the utility of connectionist networks to model existing results, their methods in character representation and in network architecture were rather crude. The first model was limited to specific naming effects and it is not clear how it can be generalized to reading acquisition. The second model used a very limited vocabulary (204 characters) and it is not clear how the model can scale up to a larger lexicon because of its localist representation (i.e., one unit per character). In addition, both models were not designed to encode the phonology of radicals, hence were unable to capture the role that radicals play in character naming (and the effects of regularity by frequency in naming). Finally, both models relied on the standard feed-forward network architecture.

In this study, we present a model of the acquisition of Chinese characters, a model that differs from previous ones in the following: (a) rather than relying on adult norms, our model builds orthographic representations of Chinese characters based on corpus analyses of the phonograms that children learn in elementary school, (b) our model uses a self-organizing neural network rather than a feed-forward network, and (c) our model attempts to connect empirical patterns, corpus analyses, and computational mechanisms in Chinese, a non-alphabetic language where the relation of orthography-phonology differs from that in English. In Xing, Shu, and Li (2002), we presented preliminary results that demonstrate the utility and promises of aspects of the model; in this study, we extend the preliminary results to a larger-scale model based on corpus derived from characters in elementary school textbooks.

Method

Network Architecture

The network architecture we used in this study is a self-organizing feature-map model developed by Miikkulainen (1997), the DISLEX model. In DISLEX, different feature maps are dedicated to different types of linguistic information (orthography, phonology, or semantics), and are connected through associative links trained by Hebbian learning. To model orthographic processing, an input pattern activates a group of units on the orthographic input map, and the resulting activation propagates through the associative links and causes an activity to form in the other map (semantic or phonological). The unit that has the highest activation in response to the input is the "winner". This process leads to the adaptive formation of associative connections between the maps. Fig. 5 presents a diagrammatic sketch of the model's reading process, from seeing the orthographic representation of dog to the comprehension of the

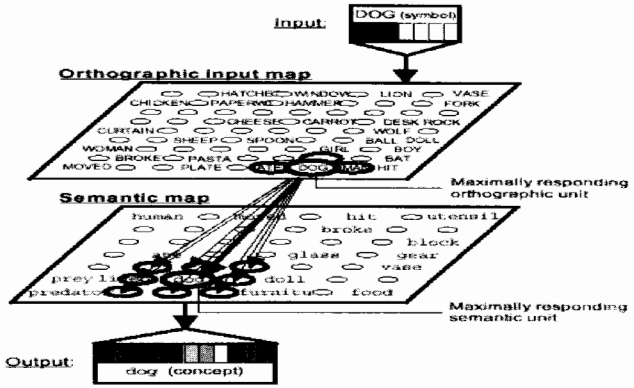

Fig. 5. Reading comprehension of dog in DISLEX (Miikkulainen, 1997; reproduced with author's permission)

word's meaning (see Miikkulainen, 1997 for technical specifications). In our simulations, we examine only the production process from orthography to phonology in order to model character naming in the acquisition of phonograms. Our network has not yet incorporated semantic information and thus no comprehension component is included in this study.

To measure the success of our network, we computed the accuracy, after each simulation, for the network's representation of orthography, phonology, and its naming ability, according to Miikkulainen's (1997) standard measures. The first two are accuracies of map representations (orthographic and phonological) that quantify the proportion of characters that are uniquely represented in the map. For example, if 惶 and 煌 are mapped onto the same node on the orthographic map, this is interpreted as evidence that the network cannot differentiate between these two characters. On the other hand, if the pronunciation of */huang2/* (tone 2) is confused with */huang1/* (*tone 1*), these two pronunciations will be mapped onto the same node in the phonological map. The third measure is on the association from orthography to phonology, measuring the accuracy of the network's naming ability. For example, if the nodes with highest activations in both maps are consistent with each other (e.g., 惶 in the orthographic map and */huang2/* in the phonological map), this is interpreted as evidence that the network correctly names the character.

Phonological Representation

A general property of Chinese characters is that one character corresponds to one monosyllable in the spoken language. This property makes it relatively simple for us to construct input representations for the phonology of characters

Table 6. Representation of Chinese phonemes by five phonological dimensions (D1-D5)

Phoneme	IPA	D1	D2	D3	D4	D5
a	A	vowel	—	—	low	central
o	o	vowel	—	round	mid	back
e	ɤ	vowel	—		mid	front
i	i	vowel	—		high	front
u	u	vowel	—		high	back
ü	y	vowel	—	round	high	front
b	p	unaspirated	bilabial	stop	—	—
p	p‘	aspirated	bilabial	stop	—	—
m	m	unaspirated	bilabial	nasal	—	—
f	f	unaspirated	labiodental	fricative	—	—
d	t	unaspirated	front	stop	—	—
t	t‘	aspirated	front	stop	—	—
n	n	unaspirated	central	nasal	—	—
l	l	unaspirated	central	lateral	—	—
g	k	unaspirated	velar	stop	—	—
k	k‘	aspirated	velar	stop	—	—
h	x	unaspirated	velar	fricative	—	—
j	tɕ	unaspirated	palatal	affricate	—	—
q	tɕ‘	aspirated	palatal	affricate	—	—
x	ɕ	unaspirated	palatal	fricative	—	—
z	ts	unaspirated	central	affricate	—	—
c	ts‘	aspirated	central	affricate	—	—
s	s	unaspirated	central	fricative	—	—
zh	tʂ	unaspirated	back	affricate	—	—
ch	ts‘	aspirated	back	affricate	—	—
sh	ʂ	unaspirated	back	fricative	—	—
r	ʐ	unaspirated	back	retroflex	—	—
ng	ŋ	unaspirated	velar	nasal	—	—
er	ɚ	vowel	—		mid	central

Note: Phonemes are given in Pinyin, along with their corresponding International Phonetic Alphabets (IPA).

in that we need not consider multiple syllables. According to tradition in Chinese linguistics, the monosyllable of each character consists of three parts: initial (onset), final (rhyme), and tone. The initial is usually a consonant. The final consists of at least the nucleus vowel, sometimes with or without a medial or an ending. The nucleus vowel may be one single phoneme or a diphthong (two phonemes). Lexical tones are supra-segmental, imposed on the initial and the final. In our representational scheme, we represent each phoneme (consonant or vowel) by 5 dimensions or features, and each feature by the phoneme's articulatory properties, on a continuous scale from 0 to 1. The overall method of representation is similar to PatPho, a phonological representation scheme for English described by Li and MacWhinney (2002).

With this method we can represent all Chinese monosyllables with four tones (a total of 1,335), with each syllable on a 30-unit feature vector. Table 6 lists the articulatory features we used to represent the Chinese phonemes.

Orthographic Representation

As mentioned earlier, Chinese characters are complex in at least two respects: first, they are visually complex in terms of the number, the shape, and the relative position of the strokes that make up the character, and second, they are structurally complex in the hierarchies in which these strokes are assembled into radicals, and radicals assembled into characters. For example, the character "鞭" can be divided into the radicals "革" and "便" in the first level. The radical "便" can be further divided into smaller components which could also be radicals by themselves. At the same time, radicals are often independent characters, in much the same way that a root can be an independent word in English; for example, the radical "便" by itself is an independent character. One implication of these complex nestings in Chinese orthography is that it is very difficult to accurately represent the orthographic similarities of characters. Crucially, this complexity poses an obstacle to modeling research because a simple decomposition of visual features is neither feasible nor sufficient when the orthography is not alphabetic. Perfetti, Liu & Tan (2002) localist representation did not solve the problem because it could not be scaled up, and Chen and Peng's distributed representations were specific to the particular characters they used in training. One contribution of the present study is our solution to character representation.

We constructed our character representations on the basis of a detailed analysis of all the characters in the *UCS Chinese Character Database* (Standards Press of China, 1994) with respect to the strokes, components, and structures of these characters. The *UCS Chinese Character Database* contains information about the structure and components for each of the 20,902 Chinese characters used in China, Japan, and Korea. This information includes the hierarchically ordered sequences of each component when characters are decomposed into smaller units of strokes. Other information includes pronunciation of the character, first-level categorization of the character, number of radicals, number of strokes, and frequency of usage. The database lists 560 basic radicals for the 20,902 Chinese characters, including each character's structural features, shape features, position of radicals, number of radical strokes, etc. Most important, in our study we included information about phonetics in phonograms as found in our school textbook corpus. This included the position of the phonetic in the character, whether the position of the radical is fixed, and the relationship between the pronunciation of the phonetic and that of the character. We also included information about the frequency of each character that appears in our school corpus.

On the basis of our analyses of this database and the textbook corpus, we represented each phonogram character with a 382-unit feature vector, along the dimensions as depicted in Fig. 6.

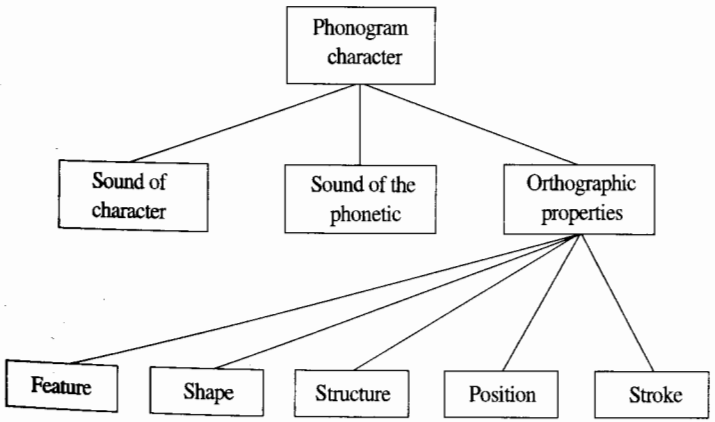

Fig. 6. Orthographic representation scheme for Chinese characters

The first 30 units represent the sound of the character, while the second 30 units represent the sound of the phonetic. The remaining 322 units represent the orthographic properties of characters, such as structure of radicals, shape of radicals, position of radicals, feature of strokes, shape of strokes, and number of strokes. The decision to include information about radicals and strokes in the representation of Chinese orthography was based on behavioral work that showed that the processing and learning of characters depends crucially on the reader's ability to decompose characters into radicals and smaller components (Taft, in press; Shu, in press). In the *Appendix* we provide details on how we coded these aspects into the orthographic representation of characters. With this representation scheme, we are able to model some of the key properties that are important in Chinese character acquisition. These include: (1) phonological similarities among phonograms, (2) orthographic similarities among phonograms, (3) positional features of the phonetic within phonograms, (4) pronunciations of the phonetic in phonograms, and (5) relationships between the pronunciations of characters and those of their phonetics.

Simulation 1: Modeling Regularity Effects

Materials

The basic training materials consisted of groups or families of Chinese phonograms — characters that have the same phonetic. It is important to note again that some phonetic components appear in many characters (large family) and some appear in relatively few (small family). Because we are modeling elementary school children's acquisition, we allowed the number character families to differ from grade to grade, and the same family may also contain different numbers of family members at each grade level (see Table 7). We randomly sampled phonogram characters from our textbook corpus for Grades 2, 4, and 6 to generate the basic training materials. Since the total number of characters in the training for each grade is limited to 300, the percentage of families sampled differs across grades: 50 percent for Grade 2, 13.5 percent for Grade 4, and 9.2 percent for Grade 6. Selection of training materials mirror introduction of characters to children. In particular, items are selected to represent distribution of phonetic at particular grade levels. Characters are selected (a) if they have been learned in or before this grade, and (b) if the family includes all phonogram characters that have been learned before. For

Table 7. Number of families sampled from Grades 2, 4 and 6 textbooks

Family members	Grade 2		Grade 4		Grade 6	
	Total	Sampled	Total	Sampled	Total	Sampled
1	503	100	560	60	565	48
2	148	27	237	24	234	21
3	83	17	131	13	156	12
4	22	4	72	7	95	7
5	19	3	53	6	72	6
6	12	2	24	3	41	3
7	4	1	21	2	38	2
8	1	0	7	1	14	1
9	1	0	10	1	11	1
10			5	0	6	1
11			2	0	7	1
12			4	0	7	1
>13					6	0
Total	793	154	1126	117	1252	104

example, in grade 2 most characters have a phonetic that appears in only one character (small family); by contrast, at grade 6 some characters have phonetics that appear in 12 characters or more. Table 7 shows the composition of our training materials in terms of character families.

Training

Each batch of characters corresponding to each grade was submitted to the network, trained for varying numbers of epochs for the self-organization of phonological representations and of orthographic representations (see details below). Upon training of the network, a phonological representation of a character was presented to the network, and simultaneously, the orthographic representation of the same character was also presented to the network. Through self-organization the network formed an activity on the phonological map in response to the phonological input, and an activity on the orthographic map in response to the orthographic input. The phonological representation of the character was also co-activated with its orthographic representation. As the

network received input and continued to self-organize on each map, it simultaneously learned associative connections between maps through Hebbian learning: initially, all units on the phonological map were fully connected to all units on the orthographic map; as learning continued, only the units that were co-activated in response to the inputs were associated. All simulations were run with the DISLEX simulator (Miikkulainen, 1999) on a Pentium 4 PC under the Linux environment.

The network was trained separately on 2, 4, 6 grade corpus samples, where average family size as well as frequency varied across samples. Frequency was modeled by the number of presentations of the input. The actual training time (number of cycles) for each character is calculated with the following formula:

Training times in the network = frequency of usage in the textbooks x 20

For example, a character will be trained for 60 times, if it appears three times in the textbooks. A pilot study shows that the network is able to learn the pronunciation of characters in the training pool with about 350 epochs of training. So, for a high frequency character which appears 18 times or more in the textbooks, it will be trained for 360 times. This coarse approximation to frequency in terms of training time guarantees that the orthographic and phonological representations and their connections for each character in the network will be different after some amount of initial training.

Testing

Once the network has self-organized on the phonological and orthographic inputs and has learned the associative connections, we tested the model's performance by presenting the network with the testing characters.

Three groups of phonograms were sampled from the training set of Grades 2, 4 and 6 characters in the model. Each included 60 characters, in which 40 were from the trained characters, and 16 unfamiliar new characters that the network was not trained on. Within the 40 trained characters, 20 were of high frequency and 20 of low frequency. High frequency and low frequency characters were defined as follows, with varying criteria for varying grades: for Grade 2, 5 times or more of occurrence were considered high frequency, while 2 times or less were considered low frequency; for Grade 4, 8 times or more of occurrence were considered high frequency, while 4 times or less were considered low frequency; for Grade 6, 18 times or more of occurrence were considered high frequency, while 9 times or less were considered low frequency. The increasing numbers used in defining frequency ranges are based on the consideration that a given character will accumulate more frequencies of usage at successive grades. For example, the frequency of a character for Grade 2 is calculated from its occurrences in first and second grade textbooks, whereas that for Grade 6 is calculated from occurrences in first through sixth grade textbooks. New characters serve the same function as non-words in English; although the model was not trained on these characters, the model probably learned aspects of the character's phonological and orthographic information, due to the resemblance of the trained characters to the new characters, because the phonetics of the new characters also appeared in the trained characters. For these 20 new characters, eight were regular and 10 were irregular or semi-regular in terms of character-to-radical pronunciation regularity.

We inputted the orthographic and phonological patterns of the testing characters to the trained network to test the model's naming ability, that is, the output pronunciations of the characters. The outputs included the pronunciations of the testing characters and the orthographies of the testing characters. No learning takes place at this stage. Each grade model was run separately for ten times, and the results were averaged across the ten runs.

Results

Overall performance of the model. The model's performance was evaluated with the standard measures of representation accuracy as used by Miikkulainen (1997). These are accuracies computed for character orthography, character phonology, and associative connections from orthography to phonology (see earlier discussion in *Method*).

The overall performance of the network after it was trained for 350 epochs on all characters corresponding to each of the three grades being considered. The network achieved an average of 76% accuracy for orthographic representations, 79% accuracy for phonological representations, and 93% accuracy for the associative connections from orthography to phonology, a highly successful naming ability.

Effects of age and frequency in the model. To see the model's ability in character naming across grade, we first tested the accuracy of its naming of

regular and irregular characters for Grades 2, 4, and 6. The mean performance of the model from 10 simulation runs indicates that, like empirical results from children, there were significant effects of age (grade) and frequency. The naming accuracy for high-frequency characters (.91) was significantly higher than that for low frequency characters (.66) and new characters (.24) ($F(2,54) = 1104.67$, $p < .01$). The naming accuracy of higher grades (.64, .67) was higher than that of lower grade (.49) ($F(2,27) = 83.21$, $p < .01$). Most important, the interaction of grade by frequency was significant, $F(4,54) = 25.42$, $p < .01$. Comparing the performances for high and low frequency characters, we found that because of changes for low-frequency characters, the model for Grade 2 showed a larger frequency effect (.43) than for Grade 4 and Grade 6 (.16, .16, respectively). This result indicates that as the network learns more characters, its sensitivity to frequency decreases. In other words, frequency plays a stronger role in earlier years of acquisition than in later years. However, for new characters, the model in Grade 4 and Grade 6 showed higher accuracy (.33, .26) than in Grade 2 (.13), $F(2,27) = 27.88$, $p < .01$). This result indicates that the higher-grade models display a larger capacity to generalize to novel characters than the lower-grade models because more characters have entered the learner's repertoire.

Regularity effects in the model. Our model displayed a significant regularity effect. The ratio of naming accuracy for regular characters (.70) was higher than that for irregular characters (.50) ($F(1,27) = 250.71$, $p < .01$). More important, the model showed significant interactions of grade by regularity ($F(2,27) = 7.8$, $p < .01$) and frequency by regularity ($F(2,54) = 6.28$, $p < .01$) , as seen in Fig. 7 and 8. In general, the effect of regularity tends to decrease with grade: the model in Grade 6 showed a smaller regularity effect (.13) than did the Grades 2 and 4 models (.23 in both cases). On the other hand, the low-frequency and new characters (.23, .21) showed a larger regularity effect than the high-frequency characters (.14). Previous computational models such as Chen and Peng (1994) and Perfetti and Liu (in press) could not capture these interaction effects. Our results match well with patterns found in empirical research with children (e.g., Ho & Bryant, 1997; Shu, Anderson, & Wu, 2000), where frequency interacts with regularity in the acquisition and the processing of Chinese characters. Consistent with interpretations from empirical studies (Seidenberg, 1985; Shu et al., 2000), these patterns indicate that when the

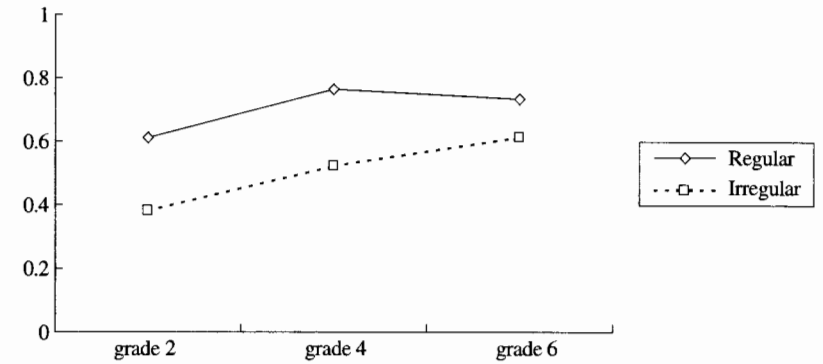

Fig. 7. Interaction of regularity by grade in Simulation 1

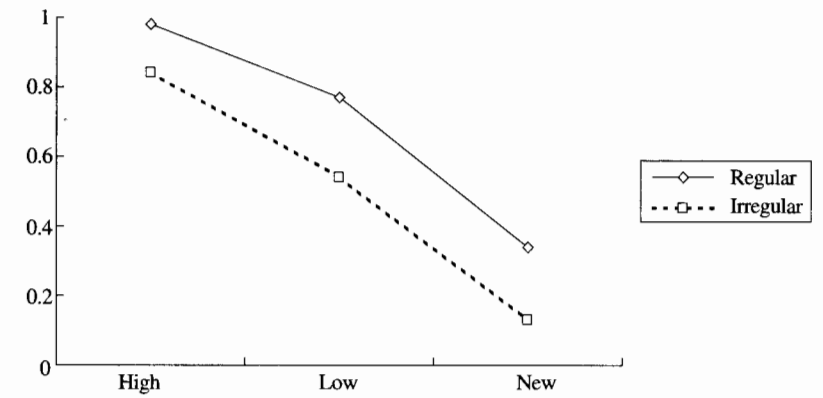

Fig. 8. Interaction of regularity by frequency in Simulation 1

learner (network or child) encounters characters that are unfamiliar or low in frequency, it makes use of the pronunciation of the existing phonetic parts, which benefits regular characters but not irregular characters.

We also analyzed the network's strategies in the naming of new characters. In naming phonograms that have not been encountered before, children as well as our network could use a variety of methods to get at the pronunciation of the character. The use of these methods would allow us to discern regularity effects in reading acquisition. There are basically three methods the learner could use:

(1) reading the character as the pronunciation of its phonetic (e.g., "橙 /*cheng2*/" as "登/*deng1*/"); (2) reading the character as another character having the same phonetic in the family (e.g., "蝙/*bian1*/" as "偏/*pian1*/"); and (3) reading the character as other unrelated characters (e.g., "枞/zong1/" as "凯 /*kai3*/"). However, each of these methods would lead to erroneous naming if the new character is irregular.

Fig. 9 shows the ratio of the network's (erroneous) methods in the naming of new irregular characters for each grade, as a function of different naming methods (M1 = Method (1); M2 = Method (2); and M3 = Method (3), as indicated above).

Several interesting patterns emerge from Fig. 9. First, for Grade 2 characters the network's strategies (errors) were mainly based on Methods 2 and 3, that is, reading the character as another character having the same phonetic in the family (.46), or reading characters as unrelated characters (.47). This indicates that although the network could capture character regularity partially, it was equally prone to random reading. However, the model for Grade 4 and Grade 6 used mainly Method 2 (.68, and .71, respectively), indicating that the network was more willing to pronounce the character as a character having similar orthographic and phonological similarities. These results, especially the dominance of Method 2 for Grades 4 and 6, indicate that our model, in the face

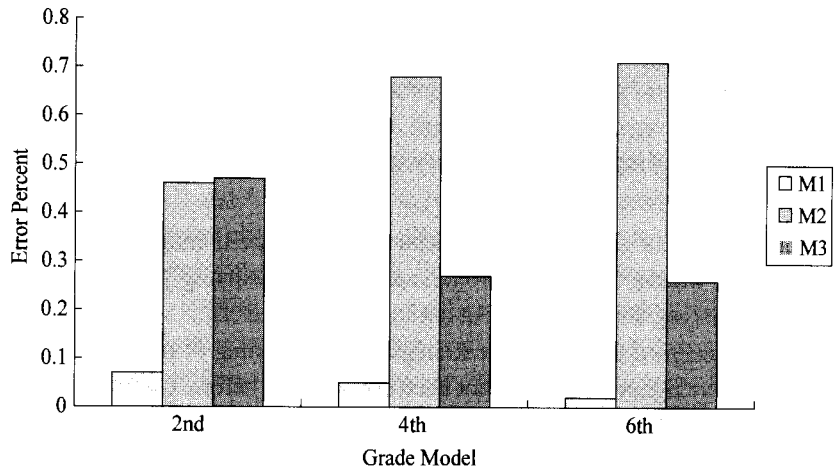

Fig. 9. Type of methods in the network's naming of new characters

of new characters, was able to exploit the sublexical orthographic information, the correlations between orthographic and phonological representations (both of which were learned from training), to generate pronunciations of new characters. Note that the new characters are similar to English nonwords, and to generalize on nonwords is a more difficult task than to produce the pronunciations of words in the training set. Our model was not exposed to the nonwords and therefore must piece together their pronunciations on the basis of exposure to other items in the training set. Thus, the model provides a computational realization of the notion of pronouncing nonwords "by analogy" to known words (Glushko, 1979).

To summarize, we found that the model displays regularity effects starting from Grade 2 learning. The regularity effect is modulated by a number of factors, including age/grade and frequency. It is more transparent for low-frequency characters than for high-frequency characters, and is more pronounced for lower grades than for higher grades. The analysis of the network's naming strategies shows that the model shifts from a heavy reliance on reading new characters as unrelated to a reliance on reading the character as its phonetic. These modeling results are consistent with empirical data available so far, in that both regularity and frequency are important factors in elementary school children's acquisition of Chinese characters. The results also map well to the predictions based on corpus analyses of school textbooks with respect to the importance of variables in character acquisition.

Simulation 2: Modeling Consistency Effects

Children and adults perform better with characters whose pronunciations are consistent for all family members, as discussed earlier. We tested our model's accuracy in naming consistent and inconsistent characters for Grades 2, 4, and 6. Here consistent characters have family members that have the same phonetic and are pronounced the same as the phonetic for characters introduced at that grade (see earlier discussion).

The method of this simulation, including network architecture, training and testing procedures, is identical to that of Simulation 1. The only difference is in the materials.

Materials

Each model (for each grade) contained roughly 260 characters. Consistent with the corpus, more characters from larger families were sampled for the higher-grade models than for the lower-grade models. We set a specific percentage of characters for given ranges of character families in the corpus; as grades increase, the number of characters in the same range of families also increases.

The testing materials included 90 characters, in a 3 (regular-consistent, regular-inconsistent, irregular-inconsistent) × 3 (high frequency, low frequency, new) × 3 (Grade 2, 4, 6) design. There were 30 high frequency, 30 low frequency, and 30 new characters for each grade model. New characters were not trained in the model. Frequency was modeled as follows (with varying criteria for varying grades, as in Simulation 1). For Grade 2, high frequency characters were those that occurred for 5-10 times in the textbook corpus, while low frequency characters1-2 times; for Grade 4, high frequency characters occurred for 8-13 times, while low frequency characters 3-5 times; for Grade 6, high frequency characters occurred for 16 times or more, while low frequency characters 5-8 times.

For each frequency range, there were 10 regular-consistent characters (whose character and phonetic are pronounced the same, including tone, for all family members that take the phonetic); 10 regular-inconsistent characters (whole character and its phonetic are pronounced the same, but other members that take the phonetic may be pronounced differently); and 10 irregular-inconsistent characters (whole character and its phonetic are pronounced differently, and other family members that take the phonetic are also pronounced differently). This configuration of characters allows us to assess consistency effect (regular-consistent vs. regular-inconsistent characters) along with regularity effect (regular-inconsistent vs. irregular-inconsistent characters). Note that whether a character was counted as consistent or not depended on the specific grade: a character could be consistent for a lower grade but become inconsistent in a higher grade as new characters that come into the family (i.e., sharing the same phonetic) but pronounced differently were introduced. We also assessed the degree of consistency for character families by using a mean frequency weighted consistency measure (Fang, Horne & Tzeng, 1986; Shu, Chen, Anderson, et al. 2003), according to which

Table 8. Degree of consistency in the testing materials for each grade

	2nd	4th	6th
Regular-consistent	1.00	1.00	1.000
Regular-inconsistent	0.716	0.321	0.123
Irregular-inconsistent	0.394	0.328	0.050

the number of characters pronounced the same in a family is divided by the total number of the characters in the family. Table 8 shows the degrees of consistency for our testing materials for each grade model. It can be seen that the degree of consistency for inconsistent characters decreases with grade, confirming the general pattern in the textbook corpus.

Results

After training for 350 epochs on all characters, the network achieved an average accuracy of 75.1% for orthographic representations, 78.3% for phonological representations, and 95.4% for the associative connections from orthography to phonology.

ANOVA on the naming accuracy for the three types of characters revealed a number of interesting effects. First, the main effect of character type was significant ($F(2,54) = 294.44$, $p < .01$). The proportion correct for regular-consistent characters was higher than that for regular-inconsistent characters ($p < .01$), which was in turn higher than that for irregular-inconsistent characters ($p < .01$). Second, the main effect of frequency was significant ($F(2,54) = 722.03$, $p < .01$). The accuracy for high frequency characters was higher than that for low frequency characters ($p < .01$), which was in turn higher than that for new characters ($p < .01$). For high frequency, except irregular-inconsistent characters, the model performed well starting from Grade 2, and for low frequency characters the model increased its performance across grade. Third, the main effect of grade was significant ($F(2,27) = 80.20$, $p < .01$), with the accuracy for Grade 4 and Grade 6 being higher than that for Grade 2 but there was no difference between Grade 4 and Grade 6.

There were also a number of significant interactions. First, the interaction of character type by frequency was significant ($F(4,108) = 35.35$, $p < .01$). New characters showed the largest consistency effect, as compared with high and

low frequency characters (Fig. 10). Second, the interaction of frequency by grade was also significant ($F(4,54) = 15.09$, $p < .01$). The model in Grade 2 showed a relatively larger frequency effect than in Grades 4 and 6. Third, there was also a significant interaction of character type by grade ($F(4,54) = 12.65$, $p < .01$). For Grade 2 the regular and irregular characters showed larger differences because of the model's poor performance on irregular-inconsistent characters, whereas for Grade 6 relatively larger differences were observed between consistent and inconsistent characters, especially for new characters (Fig. 11).

Finally, a significant three-way interaction of character type x grade x frequency ($F(8,54) = 19.88$, $p < .01$) reveals that, for high frequency characters, there was no consistency effect (i.e., the model performed best for all except I-I characters), for low frequency characters, there was a week consistency effect, and for new characters, the consistency effect increased with grade (performance increased for regular-consistent characters but remained steady for regular-inconsistent characters; see Fig. 12). That naming accuracy increases for consistent but not for inconsistent characters could be explained as follows: the addition of new characters to consistent character families increases the degree of consistency and thus boosts the overall naming

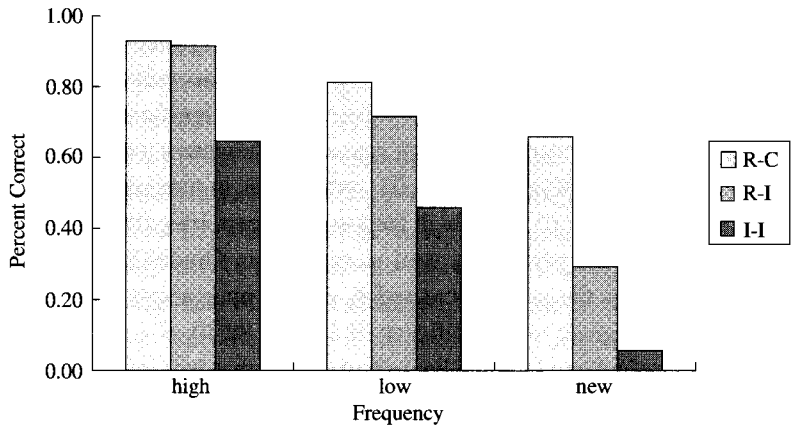

Fig. 10. Interaction between character type and frequency in Simulation 2

Note. R-C = regular consistent; R-I = regular inconsistent; I-I = irregular inconsistent

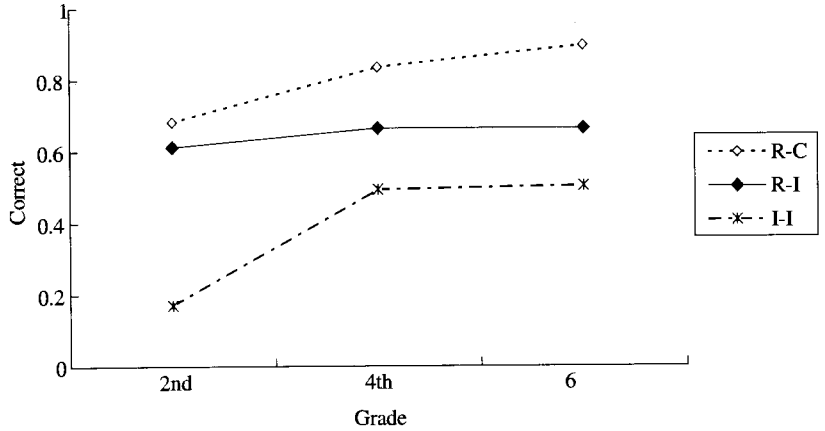

Fig. 11. Interaction between character type and grade in Simulation 2

Note. R-C = regular consistent; R-I = regular inconsistent; I-I = irregular inconsistent

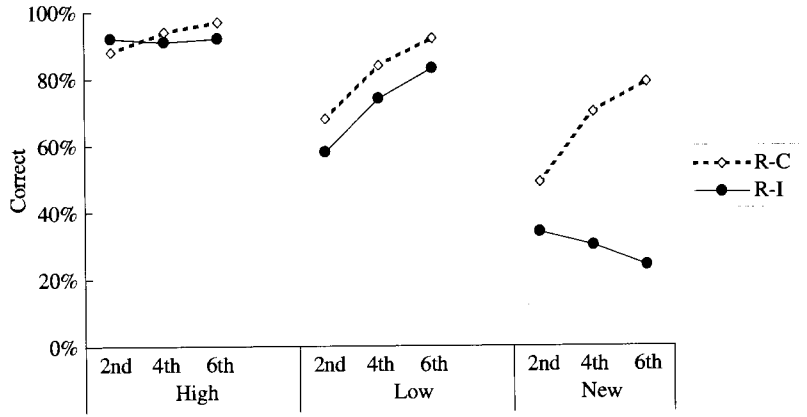

Fig. 12. Three-way interaction of consistency x grade x frequency in Simulation 2

Note. R-C = regular consistent; R-I = regular inconsistent

accuracy, whereas the addition of new characters to inconsistent characters can only deteriorate the overall degree of consistency for given character families. This pattern is consistent with results from empirical studies (e.g., Shu, Zhou,

& Wu, 2000).

To summarize, the model in Simulation 2 displays both regularity effect and consistency effect, and these effects interact with character frequency and grade of learning. The consistency effect increases with grade, especially in the naming of low frequency and new characters. These results are consistent with the empirical results in showing that the consistency effect, compared with the regularity effect, develops more slowly, perhaps not until when the learner is in fourth grade (Tzeng, Zhang, Hung & Lee, 1995; Shu & Wu, in press). Thus, in general, we see that consistency effects increase with grade while regularity effects decrease with grade (Shu, Zhou & Wu, 2000; Yang & Peng, 1997). The model also shows the interaction between character type and frequency, in that the differences among three types of characters are larger for low-frequency characters than for high-frequency characters. This interaction is also consistent with empirical data from adults (Hue, 1992) and from children (Yang & Peng, 1997), according to which the naming latencies of different type of characters (regular-consistent, regular-inconsistent, irregular-inconsistent and non-phonograms) differ more for low-frequency characters than for high-frequency characters.

We should note that our modeling results tend to show earlier effects of both regularity and consistency (grades 2 to 4) than what empirical data suggest for children. This discrepancy may be related to the fact that there are many (about 55% in first grade) non-phonograms (e.g., pictographs and ideographs that bear no phonological cues to pronunciations) in the textbook materials for children to learn in lower grades. In our model, however, we used only phonogram characters for simulation (see earlier discussion of rationale). Thus, in the realistic learning situation, children will have a more difficult time than our model to learn the relationships between orthography and phonology, and hence acquire regularity and consistency a bit later than does the model.

General Discussion

A major debate on word reading and orthographic processing in the past decades centers on whether skilled readers use dual routes to pronounce written words, a phonological or sublexical route for regular words and a lexical route for exception words. While research by Coltheart and colleagues

support the dual-route theory of reading, connectionist models of reading by Seidenberg and colleagues argue that a single mechanism can account for the reading of regular words, exception words, and pronounceable nonwords, in both normal and impaired reading situations. Connectionist models have successfully captured the effects of word frequency, regularity and consistency in spelling-sound correspondence, and the interaction among these variables in the on-line processing of written word recognition (see earlier discussion).

Chinese is a non-alphabetic language, and the characters that form the basic units of word reading are logograms. Because the orthographic shapes of characters map mostly to meaning rather than to sound, many believe that the learning of Chinese characters involves mainly rote learning (i.e., character by character). Empirical research in the last twenty years indicates, however, that readers use the regularity and consistency information in the mapping between orthography and phonology, and that these variables interact with character frequency and experience of learning. Moreover, these variables also play important roles in word reading and recognition in adults, and they affect the process of character acquisition in children. Thus, although the orthographic structure of Chinese is fundamentally different from that of English or other Western languages, word reading in Chinese is also a systematic and dynamic process.

In this study, we set out to capture a natural character corpus as learned by children in elementary school, with particular reference to the roles that character regularity, consistency, frequency, and their interactions play in children's acquisition process. From our corpus analysis, we show that there are many inherent statistical properties of the characters that children learn as they progress through the elementary school years. These statistical properties promote children's awareness to regularity, consistency, and frequency effects. In the connectionist models, we further examined how the learner, a self-organizing neural network, extracts these properties in the process of character naming.

On the basis of character's visual properties from a large-scale character database, we have developed representation schemes to faithfully capture the orthographic similarities of Chinese characters and to serve as input to our model. Our model successfully captures various effects of character properties, and the complex interactions between them and their interactions with age of

learning. In particular, consistent with empirical evidence, our simulations indicate that regularity effects are more transparent for low-frequency characters than for high-frequency characters, and is more pronounced for lower grades than for higher grades. The analysis of the network's naming errors shows that the model shifts from using no regularity information in naming novel characters to using such information at a later stage. Our results also show clear consistency effects, which interact with character frequency and grade of learning, and in contrast to regularity effects, consistency effects increase with the learning grade, especially in the naming of low frequency and new characters. These simulated patterns of regularity, consistency, frequency and their interactions with age/grade match with the predictions based on our corpus analyses of school textbooks, and with available empirical evidence on children's acquisition of characters during the elementary school years.

With respect to regularity effects, our model indicates a general regularity effect from Grade 2 on, but detailed analyses of low-frequency characters and the naming strategies show, consistent with Shu et al.'s (2000) argument, that systematic knowledge of character regularities (and the application of it to new characters) takes time to develop. Shu et al. (2000) argued that although school children can in principle utilize phonetic information early on, they display regularity effect only after they have learned a relatively large number of items in the phonogram families (around grade 3 or 4). At this point, it is not yet clear what number would constitute a large enough number, for example, whether children at Grade 2 have learned enough characters for them to develop awareness to character regularity, or whether this ability has to wait until later grades. Future simulations with better control of the *quantity* of characters in learning are needed to address this issue. With respect to consistency effects, our model indicates that character consistency depends highly on character frequency: consistency effects are much less clear for high frequency characters than for low-frequency and new characters. Empirical data from children and adults both suggest that naming latencies of different type of characters (regular-consistent, regular-inconsistent, irregular-inconsistent) do not differ for high frequency characters, but differ for low frequency characters.

Note that our results present only a general (averaged) picture of children's development of sensitivity to regularity and consistency of characters.

Empirical research shows that children may be divided into "good" or "poor" readers, depending on a number of variables in the developmental process. Such individual variations obviously would affect how early a given child will show effects of regularity and consistency. There is an emerging literature that indicates, for example, that children's knowledge or awareness of the phonological structure of characters contributes significantly to their reading abilities (Leong, in press; McBride-Chang & Zhong, in press; Shu & Wu, in press; Siok & Fletcher, 2001), which, among other variables (e.g., visual skills), would account for individual variations in character acquisition. Future research should consider parameters that lead to such individual variations in the model.

Our corpus analyses and connectionist modeling studies represent an initial attempt in the systematic investigation of children's acquisition of Chinese characters. The computational properties as implemented in self-organizing neural networks (e.g., DISLEX) have allowed us to model the classical effects of word reading and naming in an orthographically different language. Our analyses and models often suggest more detailed patterns of interaction than what is currently available in the empirical data. For example, empirical research has yet to provide us with a detailed picture of the various interactions among character types, frequency, and age in acquisition by elementary school children. Thus, our study could serve to inspire more empirical studies, against which detailed modeling results can be compared.

References

Beijing Academy of Educational Sciences (1998). *Liunianzhi xiaoxue shiyong keben* (Elementary school textbooks for first through sixth grades). Beijing: Beijing Press.

Chen, Y., & Peng, D. (1994). A connectionist model of recognition and naming of Chinese characters. In H-W. Chang, J-T. Huang, C-W Hue, & O. Tzeng (eds.), *Advances in the study of Chinese language processing* (Vol.1, pp. 211-240). Taipei: National Taiwan University Press.

Coltheart, M. (1978). Lexical access in simple reading tasks. In G. Underwood (Ed.), *Strategies of information processing* (pp. 151-216). New York: Academic Press.

Fang, S. P., Horng, R. Y., & Tzeng, O. J. L. (1986). Consistency effects in the

Chinese character and pseudo-character naming tasks. In H. S. R. Kao and R. Hoosain (eds.), *Linguistics, psychology, and the Chinese language* (pp.11-21). Center of Asian Studies, University of Hong Kong.

Glushko, R. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, 5, 674-691.

Goodman, J.C., Dale, P.S., & Ping, L. (2002). *The Relationship between parental frequency and order of acquisition in lexical development.* Poster presented at the Society for Research in Child Language Disorders, Madison, WI, April.

Hanyu Da Zidian Commission (1990). *Hanyu Da Zidian* (A comprehensive dictionary of Chinese). Sichuan, China: Sichuan Dictionary Publisher.

Harm, M.W. & Seidenberg, M.S. (1999) Phonology, reading acquisition, and dyslexia: Insights from Connectionist Models. *Psychological Review*, 106, 491-528.

Harm, M.W. and Seidenberg, M.S. (in press). Computing the meanings of words in reading: Divisino of labor between visual and phonological access. *Psychological Review.*

Hebb, D. (1949). *The organization of behavior.* New York: Wiley.

Hernandez, A., Li, P., & MacWhinney, B. (2005). The emergence of competing modules in bilingualism. *Trends in Cognitive Sciences*, 9, 220-225.

Ho, C. S., & Bryant, P. (1997). Learning to read Chinese beyond the logographic phase. *Reading Research Quarterly*, 32, 276-289.

Hu, C. F., & Catts, H. W. (1998). The role of phonological processing in early reading ability: What we can learn from Chinese. *Scientific Studies of Reading*, 2, 55-79.

Hue, C. W. (1992). Recognition processing in character naming. In H. C. Chen & O. J. L. Tzeng (Eds.) *Language Processing in Chinese*. North-Holland: Elsevier Science Publisher.

Kohonen, T. (1989). *Self-organization and associative memory*. Heidelberg: Springer-Verlag.

Kohonen, T. (1995). *Self-organizing maps*. Heidelberg: Springer-Verlag.

Leong, C.K. (in press). Making explicit children's implicit epilanguage in learning to read Chinese. In P. Li, L.H. Tan, E. Bates, & O. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol 1: Chinese). Cambridge, UK: Cambridge University Press.

Li, P. (2003). Language acquisition in a self-organizing neural network model. In P. Quinlan (ed.), *Connectionist models of development: Developmental processes in real and artificial neural networks* (pp.115-149). Hove & Briton: Psychology Press.

Li, P., & Farkas, I. (2002). A self-organizing connectionist model of bilingual processing. In R. Heredia & J. Altarriba (eds.), *Bilingual sentence processing* (pp.59-85). North-Holland: Elsevier Science Publisher.

Li, P., Farkas, I., & MacWhinney, B. (2004). Early lexical acquisition in a self-organizing neural network. *Neural Networks*, 17, 1345-1362.

Li, P., & MacWhinney, B. (2002). *PatPho:* A phonological pattern generator for neural networks. *Behavior Research Methods, Instruments, and Computers*, 34, 408-415.

Li, Y. & Kang, J. S. (1993). Analysis of phonetics of the ideophonetic characters in Modern Chinese. In Y. Chen (ed.), *Information analysis of usage of characters in Modern Chinese* (pp. 84-98). Shanghai: Shanghai Education Publisher. (in Chinese)

MacWhinney, B. (2001). Emergence from what? *Journal of Child Language*, 28, 726-732.

McBride-Chang, C., & Zhong, Y. (in press). Emergent literacy skills in Chinese. In P. Li, L.H. Tan, E. Bates, & O. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. 1: Chinese). Cambridge, UK: Cambridge University Press.

Miikkulainen, R. (1993). *Subsymbolic natural language processing: An integrated model of scripts, lexicon, and memory.* Cambridge, MA: MIT Press.

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

Miikkulainen, R. (1999). The DISLEX simulator (new version). (Available on-line at http://www.cs.utexas.edu/users/nn/pages/software/).

National Language Commission of China (1989). *Xiandai Hanyu tongyong zibiao* (Dictionary of frequent characters in Modern Chinese). Beijing: Language Press.

Peng, D., & Jiang, H. (in press). Naming of Chinese phonograms: from cognitive science to cognitive neuroscience. In P. Li, L.H. Tan, E. Bates, & O. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. 1: Chinese). Cambridge, UK: Cambridge University Press.

Perfetti, C.A; & Liu, Y. (in press). Reading Chinese characters: Orthography, phonology, meaning and the lexical constituency model. In P. Li, L.H. Tan, E. Bates, & O. Tzeng (Eds.), *Handbook of East Asian Psycholinguistics* (Vol. 1: Chinese). Cambridge, UK: Cambridge University Press.

Perfetti, C. A., Liu, Y., & Tan, L. H. (2002). How the mind can meet the brain in reading: A comparative writing systems approach. In H. S. R. Kao, C. K. Leong, & D.-G. Gao (Eds.), *Cognitive neuroscience studies of the*

Chinese language (pp. 36-60). Hong Kong University of Press.

Perfetti, C.A. & Tan, L.H. (1998). The time course of graphic, phonological, and semantic activation in Chinese character identification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24*, 101-118.

Perfetti, C. A. & Zhang, S. (1991). Phonological processes in reading Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 17*, 633-643.

Plaut, D., McClelland, J., Seidenberg, M., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, 103, 56-115.

Seidenberg, M. S. (1985). The time course of phonological activation in two writing system. *Cognition*, 19, 1-30.

Seidenberg, M., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, 96, 523-568.

Shu, H., & Anderson, R. C. (1998). Learning to read Chinese: The development of metalinguistic awareness. In J. Wang, A. W. Inhoff, H.-C. Chen (eds.). *Reading Chinese script: A cognitive analysis* (pp. 1-18). Mahwah, NJ: Lawrence Erlbaum.

Shu, H., Anderson, R. C., & Wu, N. (2000). Phonetic awareness: Knowledge of orthography-phonology relationships in the character acquisition by Chinese children. *Journal of Educational Psychology*, 92, 56-62.

Shu, H., Chen, X., Anderson, R. C., Wu, N., & Xuan, Y. (2003). Properties of School Chinese: Implications for learning to read. *Child Development, 74*, 27-47.

Shu, H. & Wu, N. (in press). Growth of orthography-phonology knowledge in the Chinese writing system. In P. Li, L.H. Tan, E. Bates & O.J.L. Tzeng (eds.) *Handbook of East Asian Psycholinguistics* (Vol.1: Chinese). Cambridge, UK: Cambridge University Press.

Shu, H. & Zhang, H.C. (1987). The process of pronouncing the Chinese character by adult skilled reader. *Acta Psychologica Sinica* 19, 282-290. (In Chinese)

Shu, H., Zhou, X., & Wu, N. (2000). Utilizing phonological cues in Chinese characters: A developmental study. *Acta Psychologica Sinica, 32*, 164-169. (In Chinese)

Siok, W., & Fletcher, P. (2001). The role of phonological awareness and visual-orthographic skills in Chinese reading acquisition. *Developmental Psychology, 37*, 886-899.

Standards Press of China (1994). *Information Technology - UCS: Universal Multiple-Octet Coded Character Set* (Part 1: Architecture and Basic Multilingual Plane). Beijing.

Sun, M.S. (1998). The Chinese written language corpus. Beijing, China: Tsinghua University. (Available on-line at: http://www.lits.tsinghua.edu.cn/ainlp/source1.htm).

Taft, M. (in press). Processing of characters by native Chinese readers. In P. Li, L.H. Tan, E. Bates & O.J.L. Tzeng (eds.) *Handbook of East Asian Psycholinguistics* (Vol.1: Chinese). Cambridge, UK: Cambridge University Press.

Tzeng, O.J. L., Zhang, H. L., Hung, D. L., & Lee, W. L. (1995). Learning to be a conspirator: A tale of becoming a good Chinese reader. In B. de Gelder and J. Morais (eds.), *Speech and reading: A comparative approach*. Lawrence Erlbaum.

Wu, N., Zhou, X., & Shu, H. (1999). Sublexical processing in reading Chinese: A development study. *Language and Cognitive Processes,14*, 503-524.

Xing, H.B., Shu, H. & Li, P.(2002). A self-organizing connectionist model of character acquisition in Chinese In W.D. Gray & C.D. Schunn (Eds.), *Proceedings of the Twenty-fourth Annual Conference of the Cognitive Science Society* (pp. 950-955). Mahwah, NJ: Lawrence Erlbaum.

Yang H, & Peng, D. L. (1997). How are Chinese characters represented by children? The regularity and consistency effects in naming. In H. C. Chen (ed.). *The cognitive processing of Chinese and related Asian Languages*. Hong Kong: The Chinese University Press.

Yang, H., Peng, D.L., Perfetti, C. & Tan, L.H. (2000). 'Phonological activation and representation of Chinese character (1): The phonology of Chinese characters and their sub-character units', *Acta Psychological Sinica*, 32, 144-151.

Zhou, X., & Marslen-Wilson, W. (1999). The nature of sublexical processing in reading Chinese characters. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 25, 819-837.

Acknowledgments

This research was supported by a grant from the Natural Science Foundation of China (#60083005, #30070259) to H.S., and in part by an NSF grant (BCS-9975249) to P.L. while the first author was visiting the Cognitive Science Laboratory at the University of Richmond. Preparation of the article was also made possible by an NSF grant (BCS-0131829) to P.L. We would like to thank Risto Miikkulainen for making available the source code of the DISLEX program, Igor Farkas for helping with the set-up of the simulations, and Jianfeng Yang for assisting in manuscript preparation. Please address correspondence to: Ping Li, Department of Psychology, University of Richmond, Richmond, VA 23173, USA. Email: pli@richmond.edu.

Appendix: Representation of Chinese Orthography in Feature Vectors

In this part, we first describe the method for representing the 560 basic radicals in our model, and then describe how we set up the orthographic representations of whole characters (see Figure 6 for a general structure of the representational scheme).

Representation of radicals

Forty-four units are used to set up the representation of radicals. Six important properties are included in the representation: structure of radical, shape of radical, number of strokes, stroke relation in radical, position of radical, and shape of stroke.

Structure of radical. Because many radicals can stand as independent characters, the structures of radicals are similar to the structures of characters. The most common structures of radicals according to our database analysis are: unique, left-right or left-middle-right, top-bottom, top-middle-bottom, surrounding, and frame. We use 6 values to present the structure of radical.

Shape of radicals. Since the visual characteristics of radicals are very complex, we divided the possible features into 36 groups, for which only visual features, but not structural features, were considered. For each of the 36 groups, we further represented their within-group differences by ten values. For example, all radicals visually similar to 口, such as "口", "囗", "田", "日", "曰", "中", "四", are classified in one group, but they also differ slightly from one another. In Appendix Table II an example is given for one of the 36 groups. Within the group small differences between shapes are differentiated by the 10 values.

Number of strokes. One unit in our representation scheme is used for stroke numbers. To determine the value of this dimension, we analyzed the number of

Appendix Table I. The units and values for the representation of radicals

	Structure of radical	Shape of radicals	Number of strokes	Stroke relations in radicals	Position of radicals	Shape of stroke
Number of units	1	36	1	1	1	4
Number of values	6	36×10	10	6	9	4×5

Appendix Table II. Representation of orthographic properties of radicals (dimensions and values)

Structure of radicals	Single	Left-right; left-middle-right	Top-bottom	Top-middle-bottom	Surrounding	Frame								
	1.000	0.833	0.667	0.500	0.333	0.167								
Stroke relation in radical	Single	Crossing	Separate	Connecting	Crossed-connecting	Crossed-separate								
	0.167	0.333	0.500	0.667	0.833	1.000								
Position of radicals	Character	Stroke	Right	Bottom	Corner	Outline	Left	Top	Body	Middle	Inside			
	0.091	0.182	0.273	0.364	0.455	0.545	0.636	0.727	0.818	0.909	1			
Shape of radicals (36 units)	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	Ten				
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0				
Shape of stroke (4 units)	Horizontal	Vertical	Slanted	Pointed	Crooked									
	0.2	0.4	0.6	0.8	1.0									
Number of stroke	One	Two	Three	Four	Five	Six	Seven	Eight	Nine	> Ten				
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0				

strokes of characters in the UCS database. Given that about 2/3 of the Chinese characters are within 10 strokes, characters with 10 or more strokes are represented as 1.0, and the rest as values (in decrements of .1) that correspond to the number of strokes (i.e., 0.9 for characters with 9 strokes, 0.8 for characters with 8 strokes, and so on).

Stroke relations in radicals. There are six basic relations among strokes in radicals: single strokes, crossing (two strokes are crossed in a point), separate (two strokes are not connected), connecting (two strokes are connected in a point), crossed-connecting (one end of the stroke is crossed and another end is connected; for example, the second stroke of 土), and crossed-separate (one end of the stroke is crossed and another end is separate, for example, the third stroke of 东). Six values are used to represent the stroke relation in radicals.

Position of radicals. Radicals usually have a fixed or a frequent position within a character. For example, "亻"usually occurs on the left position of a character, and "刂"on the right of a character. We represented the position of a radical in a character by the following values: the radical on the left position of a character(e.g. "亻", "人字旁"), on the right position (e.g. "刂", "立刀旁"), at the top position (e.g. "𡈼", "祭字头"), at the bottom position of (e.g. "廿", "弄字底"), at the middle position (e.g. "灬", "变字腰"), on the outskirts (e.g. "冂", "同字框"), inside of a character(e.g. "夫", "庚字心"), at a corner of a character (e.g. "夕", "餐字角"), a single stroke (e.g. "一"), and the radical as an independent character (e.g. "日"). Ten values are used to represent the position of radicals.

Shape of strokes. There are four units for representing the shape of the first, second, third and last strokes in a radical. Five values are used for the shape of each stroke, according to commonly used categories in the teaching of Chinese characters. These are: horizontal, vertical, slanted, pointed, crooked.

Representation of whole characters Four important properties are included in the representation of whole characters: structure of character, ordering of radicals, split structure of radicals, number of radicals. We use the slot-based representation for combining basic radicals into characters. A total of 322 units (see Appendix Table III) is used for the orthographic representation of characters, including the units representing basic radicals.

Structure of character. The overall orthographic structures for Chinese characters can be divided into 13 categories at least, according to our analysis

Appendix Table III. The 322 units in the orthographic representation of Chinese characters

	Number of radicals	Structure of whole word	Split structure 1	Radical 1	Split structure 2	Radical 2	Split structure 3	Radical 3	Split structure 4	Radical 4	Split structure 5	Radical 5	Split structure 6	Radical 6	Radical 7
Unit	1	7	1	44	1	44	1	44	1	44	1	44	1	44	44
Value	7	34	6	411	6	411	6	411	6	411	6	411	6	411	411

of the *UCS Chinese Character Database*. However, the major six categories according to our analyses are: unique characters without apparent structure, left-right or left-middle-right (character made of left radical plus right radical, sometimes with a third radical in the middle), top-bottom (character made of upper radical and lower radical), top-middle-bottom (character made of three radicals of the order), surrounding (character made of a surrounding frame), and frame (Shu, Chen, et al., 2003; Standards Press, 1994). Six units, with 34 values, are used to represent the structure of character. For example, there are 7 values for top-bottom characters. The character "蓓" and "贺" are both top-bottom characters. However, given that the bottom of "蓓" is a left-right structure while the top of "贺" is a left-right structure, they have different values. The former is 0.429, and the later is 0.875.

Ordering of radicals. Once the structure, position, and shape of radicals are numerically coded, each radical of a character needs to be arranged in the serial order in the vector representation. The same ordering problem for letters of words occurs in English (the so-called "dispersion problem"; Plaut et al., 1996). If all radicals are simply assembled in the order in which they appear in the character, we could end up with very different representations for characters sharing the same radicals in similar position, in much the same way as we would with *spot* and *pot* if we simply code the order of letters as they

come into the representation (see Plaut et al., 1996). To preserve the structural similarities of radicals in characters, we used a slot-based or template-based representation for radicals. It works in much the same way as we would assign an initial consonant slot for /s/ in *spot*, but leave that slot empty for *pot*, such that we can align the corresponding letters or phonemes in the two words in their corresponding slots (see Li & MacWhinney, 2002, for a detailed treatment). For each character, then, we arrange the radicals to appear in a total of seven slots (given that in our corpus the maximum number of radicals is seven). In such a representation, for example, for the character "别", radical 1 is "口", radical 2 is "力", and radical 7 is "刂" For the character "利" radical 1 is "禾" and radical 7 is "刂" The two characters thus have more similarities in the representation than if they were arranged serially from one to the next. Characters that have fewer radicals are arranged so that the first radical fills the first slot, the second radical the fourth slot, and the third radical the seventh slot, thus covering the overall slot template evenly. Table IV gives examples along with their slot assignments in the template.

Split structure of radicals. One unit, with 6 values, is used to represent the split structure of radical: single, left-right (left-middle-right), top-bottom, top-middle-bottom, surrounding, frame. For example, the character "蓓" is first split into two radicals: the top radical "艹" is a single-structure radical. This is the first level of split structure, valued 1.000. The bottom radical "蓓" is a left-right structure radical, valued .833. According to ordering of radicals, it is in split structure 5. The radical "音" can be further divided into "立" and "口". It

Appendix Table IV. Examples for slot assignment in the template

Example	Structure	Radical 1	Radical 2	Radical 3	Radical 4	Radical 5	Radical 6	Radical 7
为	Single				为			
副	Left-right	一	口	田				刂
斑	Left-mid-right	王			文			王
盒	Up-down	人	一	口				
曼	Up-middle-down	日			冖			又
蓓	Up-down	艹				亻	立	口

is a top-bottom structure radical, valued .667. It is thus in split structure 6. If we combine the radical and character levels, the representation of "艹" is in Radical 1, the representation of "亻" is in Radical 5, the representation of "立" is in Radical 6, and the representation of "口" is in Radical 7, with 44 units in all of the radical representations.

Number of radicals. A given character can consist of up to 13 different radicals, according to our analysis of the database. However, most of the characters in the school corpus have three radicals, and the most complex ones have seven radicals. Seven values are used to represent the number of radicals.

Representation of phonology of radicals

One aspect of our orthographic representation scheme as depicted in Figure 6 is that it includes phonological information of the character and its phonetic component. The purpose of these phonological units is to see how much overlap there is between the pronunciation of the phonetic and that of the whole character. It would seem strange to mix phonological information with orthographic information in one representation, but this arrangement was taken because of the consideration of the unique features of Chinese characters and the empirical evidence for the role of phonology in the processing of Chinese orthography. Many studies report that phonograms are automatically decomposed into phonetic and semantic radicals during lexical access, and the sound information of the phonetic component as well as the character is also strongly activated (Taft, in press; Yang, Peng, Perfetti & Tan, 2000; Zhou & Marslen-Wilson, 1999; Wu, Zhou & Shu, 1999). These studies suggest that the orthography of a phonogram is not simply a graphic symbol of the whole character, but contains sublexical information, such as sound information of a phonetic. In our model, thus, sound information of whole characters and their phonetics is included into the orthographic representation of phonograms. Earlier models (e.g., Chen & Peng, 1994; Perfetti & Liu, in press) did not include this aspect in their representations and therefore failed to capture critical aspects of character processing.