A Self-Organizing Connectionist Model of Character Acquisition in Chinese

Hongbing Xing (xinghb@blcu.edu.cn)
Center for Studies of Chinese as a Second Language
Beijing Language and Culture University
Beijing, 100083, China

Hua Shu (shuh@bnu.edu.cn)
Department of Psychology, Beijing Normal University
Beijing, 100875, China

Ping Li (pli@richmond.edu)
Department of Psychology, University of Richmond
Richmond, VA 23713, USA

Abstract

Despite growing interests in the acquisition of Chinese orthography, few studies have modeled the acquisition process using connectionist networks. This study uses a self-organizing connectionist model to simulate children's learning of Chinese characters. There are two major goals of our study: (1) To evaluate the degree to which connectionist models can inform us of the complex structural and processing properties of the Chinese orthography. One of the most difficult tasks in achieving this goal is how to faithfully capture the orthographic similarities of Chinese characters. We derived our character representations on the basis of analyzing a large-scale character database that can be readily mapped to school children's orthographic acquisition. (2) To test the utility of self-organizing neural networks in orthographic acquisition. Most previous connectionist models of orthographic processing have relied on the use of feed-forward networks. Results from our simulations present positive evidence for both of our goals. In particular, we show that our model demonstrates early regularity effects and frequency effects in the acquisition of Chinese characters, matching up with acquisition patterns from empirical research.

Introduction

In recent years there have been growing interests in the psycholinguistic study of orthographic acquisition in Chinese (see Yang & Peng, 1997; Shu & Anderson, 1998; Shu, Anderson, & Wu, 2000). A unique feature of the Chinese orthography is that it uses characters rather than alphabetic letters as the basic writing unit, in square configurations that map mostly onto meaningful morphemes rather than spoken phonemes. Processing or acquisition within this "fractal" organization of characters may differ in important ways from that of English and other alphabetic languages (Shu & Anderson, 1998). There are four major types of Chinese characters: pictographic, referential, associative compounds, and ideophonetic compounds. The last type, also known as the semantic-phonetic compounds or, simply, phonetic compounds, is the most interesting and important. In the *Dictionary of Modern Chinese Frequent Characters* (National Language Commission, 1988), there are 5,631 ideophonetic characters, accounting for 81% of the total 7,000 frequent characters in the dictionary (Li & Kang, 1993). Shu, Chen, Anderson, Wu, and Xuan (in press) collected 2,570 characters listed in the Elementary School Textbooks used in Beijing to establish the "School Chinese Corpus". They categorized and labeled every character in this corpus, on dimensions such as phonetic part, phonetic type, position of the phonetic part in the character, age at which the character is taught, and frequency of the character. Shu et al.'s analyses reveal that most of the Chinese characters taught in elementary schools are ideophonetics, as shown in Table 1.

Table 1 Ratios of ideophonetics in each grade
(Shu, Chen, Anderson, Wu, & Xuan, in press)

Grade	1	2	3	4	5	6	Mean
Ratio	.45	.70	.76	.84	.86	.81	.74

Given the prominence of ideophonetics in Chinese orthography, it is thus important for us to understand the functions of these characters. Ideophonetics consist of two major components: the semantic part (often called a radical) that gives information about the character's meaning, and the phonetic part that gives partial information about the whole character's pronunciation. We say "partial", because the phonetic radical may or may not indicate the true pronunciation of the whole character, in one of three ways: (a) Regular: the whole character is pronounced the same as the phonetic radical in isolation – that is, the same as the phonetic radical when it is being used as a simple character; for example, "清/qing1/ and "青/qing1/". (b) Semi-regular: the whole character is pronounced partly as the phonetic radical, 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 radical (e.g., "猜/cai1/ and "青/qing1/"). These patterns of (ir)regularities in the pronunciations of ideophonetics influence the recognition and processing of Chinese characters, a phenomenon known as the *regularity effect* in the literature.

Previous studies have examined regularity effects in children's acquisition of Chinese characters. Shu, Anderson, and Wu (2000) showed that children display regularity effects when they are required to write down the pronunciations of Chinese characters: they perform better on regular characters (type a discussed above) than on irregular characters (types b and c). When children see unfamiliar characters, they often exploit the pronunciation of the phonetic radical as a possible reading of the whole character, and this ability increases with school grade. 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 interacting with the regularity of characters; for example, Shu and Wu (1996) showed that children in Grade 3 display no regularity effects on characters of low frequency, while children in Grades 4 and 6 do. Finally, Shu, Zhou and Wu (2000) found that young children develop from early on phonological awareness of the structures of characters and the functions of the phonetic and semantic radicals. Some 4th graders already start to acquire the awareness of the consistency of phonetic radicals, and by Grade 6 this awareness becomes more transparent.

The above-mentioned properties of Chinese characters and the acquisition profiles therein lend themselves naturally to connectionist modeling. Given the discrepancy between regular and irregular characters, do we need to assume dual mechanisms to handle the two types of characters in acquisition (as symbolic theorists would like to argue)? Or rather, can we assume a connectionist learning mechanism that can capture the acquisition of both types of characters? Previous research in English has examined these issues in language acquisition and orthographic processing (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989). However, due to the difficulty in representing the complex structure of the Chinese orthography, there has been very little research in this domain in Chinese. In this study, we make an initial attempt to model the acquisition of Chinese orthography, in particular, the regularity effect in acquisition (as reported in empirical studies) with a neural network.

Method

Architecture

Most previous connectionist models of orthographic processing have relied on the use of feed-forward networks, typically with the back-propagation learning algorithm (e.g., Seidenberg & McClelland, 1989). Recently, a number of studies have explored self-organizing neural networks as viable models of language processing and language acquisition (Anderson, 1999; Miikkulainen, 1993, 1997; Li, 1999, 2000). Self-organizing networks are particularly well suited for the study of language acquisition, due to their biological plausibility, unsupervised learning, and the ability to develop semantic structures (Li, 2002).

In this study, we use a self-organizing feature map model developed by Miikkulainen (1997), originally for modeling disordered lexicons (DISLEX). DISLEX relies on principles of self-organization and Hebbian learning. In this model, different feature maps dedicated to different types of linguistic information (orthography, phonology, or semantics) are connected through associative links via Hebbian learning. To model orthographic processing, an input pattern activates a group of units on the orthographic input map, and the resulting bubble of activity propagates through the associative links and causes an activity bubble to form in the other map (semantic or phonological). Fig. 1 presents a diagrammatic sketch of the model's reading process from seeing the orthographic representation of *dog* to the comprehension of the word's meaning.

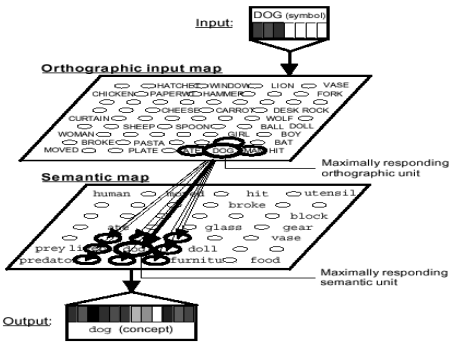

Fig.1 Reading comprehension of *dog* in DISLEX (Miikkulainen, 1997; reproduced with author's permission)

If the direction of the associative propagation is from orthography to semantics, comprehension is modeled, as shown in Fig. 1; production is modeled if it goes from semantics to orthography. In our simulations, we examine the associative propagation from orthography to phonology to model the character naming process in the acquisition of ideophonetics by Chinese-speaking children. At this stage our model has not yet included semantic information.

Input Representation

There are a few general properties of Chinese characters that are important for us to consider for their accurate representations. (1) They are structurally complex.

The basic units of characters are strokes and components. A few simple strokes can make up a component, or a character; (2) They are combinatorially complex. Compound characters have two to over ten components, and these components combine to form a character according different rules in a hierarchically organized structure; (3) The majority of Chinese characters are ideophonetic compounds, as discussed in *Introduction*; and (4) One character corresponds to one monosyllable in spoken language.

Phonological Representation. According to traditions in Chinese linguistics, the monosyllable of each character consists of three parts: initial, final, and tone (see Table 2). Initial is usually a consonant. Final consists of at least the nucleus vowel, sometimes with or without a head vowel or a tail vowel. 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 representation 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 (Table 3). The overall method of representation is similar to PatPho, a phonological representation scheme for English described by Li and MacWhinney (2002).

Table 2 Structure of the syllable and representation

Initial	Final			Tone
	Head Vowel	*Nucleus Vowel*	*Tail Vowel*	
5 dim.	5 dim	5 +5dims	5 dim	5 dim.

With this method we can represent all Chinese monosyllables with tone (a total of 1,335), each of which on a 30-dimensional feature vector. Table 3 lists the articulatory features we used to represent the Chinese phonemes.

*Table 3 Articulatory features on 5 dimensions (D1-D5) for the representation of Chinese phonemes**

D1	vowel	voiced	voiceless				
	0.1	0.75	1.0				
D2	bilabial	Labio-dental	front	central	back	palatal	velar
	0.143	0.286	0.429	0.572	0.715	0.858	1.0
D3	round	nasal	stop	fricative	affricate	retroflex	lateral
	0.143	0.286	0.429	0.572	0.715	0.858	1.0
D4	high	mid	low				
	0.333	0.666	1.0				
D5	front	central	central-back	back			
	0.25	0.5	0.75	1.0			

* Numbers indicate dimensional values for each feature

Orthographic Representation. To accurately represent Chinese orthography in feature vectors has proven a very challenging task, and this may have been the primary reason for the lack of modeling research in this domain, as we pointed out earlier. The only large-scale attempt in this respect was Chen & Peng (1994). They used 30 feature units to represent the various components of about 1,108 Chinese characters. Their representation scheme, however, is still insufficient for our purposes of modeling children's development in character acquisition. To overcome this bottleneck problem, we did a detailed analysis of all the characters in the *UCS Chinese Character Database* (Standards Press, 1994) and examined 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 components, number of strokes, and frequency of usage. The database lists 560 basic components for the 20,902 Chinese characters, including the character's structural features, shape features, position of components, number of component strokes, etc. Most relevant for our study is the information about phonetic radicals in ideophonetic characters. This includes the position of the phonetic radical in the character, whether the position of the radical is fixed, and the relationship between the pronunciation of the phonetic radical and that of the character. Finally, the database contains information about the frequency of each character in elementary school texts, as well as some of the original texts.

On the basis of our analyses of this database, we represented each ideophonetic Chinese character with a 60-unit feature vector, along the dimensions as depicted in Fig.2.

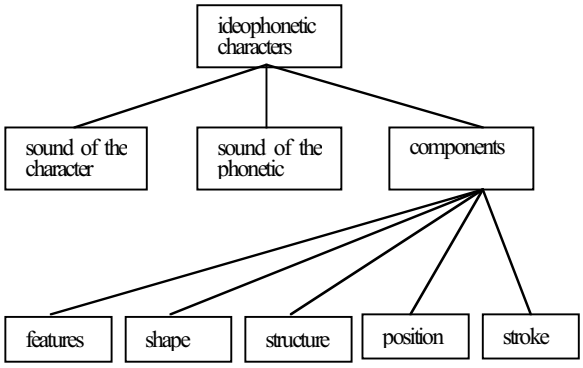

Fig. 2 Orthographic representation of characters

The first 6 units represent the sound of the character, while the second 6 units represent the sound of the phonetic radicals. The purpose of these phonological units is to see how much overlap there is between the

pronunciation of the phonetic radical and that of the whole character. The next 48 units are used to represent component features, shapes, stroke structures, position of radicals, and stroke numbers. For example, component features include single, separate, crossing, and connecting; position of radicals includes top, bottom, left-side, right-side, middle, and inner, etc. The last unit is used for stroke numbers, and to determine the value of this unit, we analyzed the number of strokes of characters in our database. Given that most of the characters are between 1 to 10 strokes, characters with 10 or more strokes are represented as 1.0, and the rest as values (in decrements of .1) corresponding to the number of strokes (i.e., 0.9 for characters with 9 strokes, 0.8 for characters with 8 strokes, and so on).

Materials and Procedure

Materials. The basic training materials consist of groups or families of Chinese characters – characters in the same family have the same phonetic radicals, sometimes including the radical itself as a character. Because we are modeling elementary school children's acquisition of characters, the amount of character families differ for different grades, and the same family may also contain different numbers of family members, according to our analysis of the School Chinese Corpus (Shu et al., in press). We selected families of characters from the elementary textbooks for Grades 1, 3, and 5 as the basic materials in our training. Characters are selected as our input materials (a) if they have been learned in or before this grade, or (b) if the family includes all ideophonetic characters that have been learned before. Table 4 shows the composition of our training materials, based on depth of learning in school grades.

Table 4 Selected characters and family compositions

Grade	Total characters	# of families	Mean members of a family
One	306	214	2.35
Three	305	139	4.33
Five	300	113	5.64

Training. Each batch of characters corresponding to each grade was submitted to the model, trained for 350 epochs for the self-organization of phonological representations and of orthographic representations. Upon training of the network, a phonological representation of a character was inputted to the network, and simultaneously, the orthographic representation of the same character was also presented to the network. By way of 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. As the end of learning, the network should have created compressed new representations in the corresponding maps for all the inputs and linked the phonological representation to its orthographic pattern. All simulations were conducted with the DISLEX simulator (Miikkulainen, 1999).

Testing. Once the network has completed self-organizing on the phonological and orthographic inputs and has learned the associative connections, we tested the model's performance by presenting the network with 16 ideophonetic characters. We inputted the orthographic and phonological patterns of these 16 characters to the trained and well-settled network to test the output pronunciations of the characters in the model (see Fig. 1). No learning takes place at this stage. For each grade, a total of 48 characters was tested in the model, in three batches: 16 high frequency characters, 16 low frequency characters, and 16 new characters that have not be learned by the grade being tested.

Results and Discussion

Table 5 shows 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. These results show that the network achieved an average of 76% accuracy for orthographic representations; for phonological representations, it reached an average of 79% accuracy, and for the associative connections from orthography to phonology it achieved an average of 93% accuracy, a highly successful naming ability.

Table 5 Percent accuracy on orthography, phonology, and associative connections in the model after training

Grade	Ortho. map	Phono. map	Associative ortho.->phono.
One	75.6	78.4	90.5
Three	71.7	76.2	93.7
Five	80.2	82	95

Thus, after training, the model developed clearly structured representations for both the phonological and orthographic input patterns. Fig. 3 shows an example from the orthographic map trained on Grade 5 characters (only a portion of the entire map is shown here due to space limit).

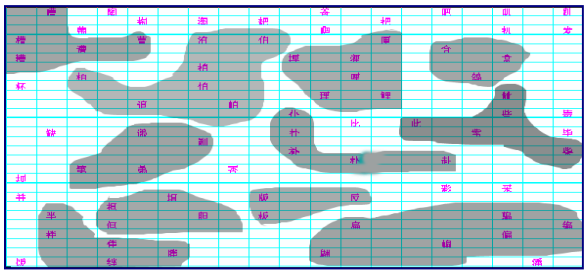

Fig. 3 Orthographic map trained on Grade 5 characters

In Fig. 3, it can be seen that clear families of characters emerged in the map after the network was trained for 300 epochs on the Grade 5 characters. For example, one group of characters on the lower right-hand corner represents the "扁/bian3" family with "骗/pian4" ,"编/bian1", "偏/pian1", "蝙/bian1", "翩/pian", while the other group on the upper right hand of the map represents the "合/he2" family with "盒/he2" ,"鸽/ge1", etc.

To see the model's ability in character naming, we tested the accuracy of its naming of regular and irregular characters for Grades 1, 3, and 5, with regular character being one that has exactly same pronunciation as its phonetic radical (see Introduction). The ratios of naming accuracy are presented in Fig. 4, on which we can make several observations: (a) the model's naming accuracy increases over time for both regular and irregular characters; (b) the difference in naming accuracy between regular and irregular characters also increases across grade; and (c) regularity effect does not exist for Grade 1 but becomes transparent for Grades 3 and 5. These results match up well with the empirical patterns observed by Shu et al (2000).

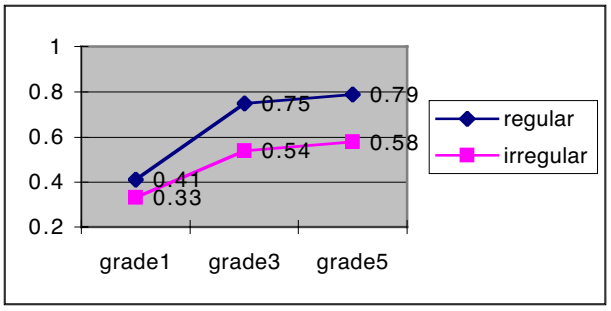

Fig. 4 Naming accuracy for regular and irregular characters

Interestingly, the regularity effect shown in Fig. 4 is modulated by character frequency. Ample empirical evidence suggests that frequency effect interacts with regularity effect in character acquisition (Shu, Anderson, & Wu, 2000). Table 6 shows that regularity effect in the model is only marginal for high frequency characters, but is much clearer for low frequency characters and novel characters (whose frequency is unknown to the network).

Finally, we analyzed the network's error types in naming irregular characters. In naming ideophonetic characters, children as well as our network could use a variety of methods to get at the pronunciation of the irregular character. These methods allow us to discern regularity effects in reading acquisition. There are basically three major methods they could use: (1) reading the irregular character as the pronunciation of its phonetic radical (e.g., "橙/cheng2" as "登/deng1"); (2) reading the character as another character having a similar orthography/radical in the same family (e.g., "蝙/bian1" as "偏/pian1"); and (3) reading the character as other irrelevant characters (e.g., "枞/zong1" as "凯/kai3"). Table 7 shows the ratio of the network's erroneous naming of irregular characters for each grade, as a function of naming methods (M1 = Method (1); M2 = Method (2); and M3 = Method (3)).

Table 6 Interaction between regularity and frequency

Frequency	Regular	Irregular
	.88	.83
Low frequency	.75	.46
New characters	.33	.17

Table 7 Network's naming for irregular characters

Grade	Irregular Character Naming Errors		
	M1	*M2*	*M3*
One	.06	.25	.69
Three	.36	.46	.18
Five	.30	.50	.20

Table 7 shows several interesting patterns. First, for Grade 1 the network's errors are mainly based on Method 3, i.e., reading characters as irrelevant characters. This shows that regularity effect has not played much of a role yet in the naming of irregular characters. For Grades 3 and 5, however, the error types shift more toward Methods 1 and 2, showing that the model is exploring orthographic and phonological similarities of the radical to give possible pronunciations of the irregular character. These developmental patterns of regularity effect are consistent with empirical data from Shu et al (2000), according to which children, although in principle can utilize ideophonetic information early on, show regularity effect only after they have learned a relatively large number of items in the ideophonetic families.

Conclusion

This study uses a self-organizing connectionist model to simulate children's acquisition of Chinese characters. There are two major goals of our study. First, we wanted to see if connectionist models can be applied successfully to model the learning process in the acquisition of Chinese characters, a topic that has not been touched on in the literature. Given the complex structural and processing properties of the Chinese orthography, it is only natural that we examine this

domain with systematically varying modeling parameters. One of the most difficult tasks in achieving this goal is how to faithfully capture the orthographic similarities of Chinese characters, as discussed in *Method*. We derived our character representations on the basis of analyses of a large-scale character database that can be readily mapped to school children's orthographic development.

The second goal of our study is to test the utility of self-organizing neural networks. Most previous connectionist models in this domain have relied on the use of feed-forward networks, typically with the back-propagation learning algorithm. In previous research, Miikkulainen (1993, 1997) explored self-organizing neural networks as plausible models of language and memory processing, and Li (1999, 2000, 2002) showed these networks as viable models of language acquisition. We wanted to see if such models can be used successfully to examine orthographic acquisition. Our initial simulations as presented here seem to provide positive evidence in this respect. In particular, we showed that our self-organizing network demonstrates regularity effect and frequency effect in the acquisition of Chinese characters, and that these results match up with developmental patterns observed in empirical research. In future studies, we will continue these lines of experiments to examine frequency effect, phonological consistency effect, and their interaction with regularity effect in character acquisition, addressing other relevant theoretical issues in connectionist language acquisition.

Acknowledgments

This research was supported by a grant from Natural Science Foundation of China #60083005 to H.S., and in part by an NSF grant #9975249 to P.L. while the first author was visiting the Cognitive Science Lab at the University of Richmond. We would like to thank Risto Miikkulainen for making available the source code of the DISLEX program, and Igor Farkas for helping with the set-up of the simulations.

References

Anderson, B. (1999). Kohonen neural networks and language. *Brain and Language*, *70*, 86-94.

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.

Li, P. (1999). Generalization, representation, and recovery in a self-organizing feature-map model of language acquisition. In M. Hahn & S.C. Stoness (eds.), *Proceedings of the 21st Annual Conference of the Cognitive Science Society* (pp.308-313). Mahwah, NJ: Lawrence Erlbaum

Li, P. (2000). The acquisition of tense-aspect morphology in a self-organizing feature map model. In L. Gleitman & A.K. Joshi (eds.), *Proceedings of the 22nd Annual Conference of the Cognitive Science Society* (pp.304-309). Mahwah, NJ: Lawrence Erlbaum.

Li, P. (2002). Language acquisition in a self-organizing neural network model. In P. Quinlan (ed.), *Connectionist models of development*. Philadelphia and Briton: Psychology Press.

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

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)

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 (1988). *Dictionary of Frequent Characters in Modern Chinese*. Beijing: Yuwen Press.

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., & McClelland, J. (1989). A distributed, developmental model of word recognition and naming. *Phonological 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. (in press). Properties of School Chinese: Implications for learning to read. *Child Development*.

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

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

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.