Advanced Learning Chinese Characters Method Based on the Characteristics of Component and Character Frequency

Chung-Ching Wang (stanleyccwang1987@gmail.com), Ming-Liang Wei(N26011623@mail.ncku.edu.tw)
Department of Electrical Engineering, National Cheng Kung University, Tainan, Taiwan

Yu-Lin Chang (gtyulin@gmail.com), Hsueh-Chih Chen(chcjyh@ntnu.edu.tw)
Department of Educational Psychology and Counselling, National Taiwan Normal University, Taipei, Taiwan

Yi-Ling Chung (lydia193@gmail.com), Jon-Fan Hu(jfhu@mail.ncku.edu.tw)
Department of Psychology, National Cheng Kung University, Tainan, Taiwan

Abstract

Chinese has been recognized as one of most major languages in the world, and it is evident that more and more people are interested in understanding or using Chinese. Thus, developing an efficient approach for learning Chinese characters is considered as an important issue. Certain previous studies have suggested various methods to learning Chinese characters for the purpose of showing students how to read Chinese characters. In Chinese, the components can offer learners phonological and morphological meanings similar to the prefixes and suffixes in English, and character frequency provides learners a character list which can be widely used in daily life. However, very few studies have considered integrating the characteristics of component and character frequency. In this study, we have developed an effective and systematic approach for learning Chinese characters based on both components and character frequency. The purpose of the study is to propose a traditional Chinese character learning metric and to present a method for learning only a few components and then the resulting reading of more high frequency characters made up of these components. Combining components and character frequency advantages, it can present an effective, systematic and rapid mechanism for learning traditional Chinese characters.

Keywords: Learning Chinese; Character frequency; Components.

Background

Chinese is a popular and widely used language in the world, and it is quite difficult and complicated to read and write Chinese characters. In the field of Chinese character recognition research, previous studies have actively involved Chinese character encoding strategies(Hayes, 1988), and character recognition strategies, such as meaning recognition(Everson, 1998), orthographic effect(Lin, 2000), and the reading process(Ke, 1998). These studies have contributed many approaches to improve Chinese character recognition. Also, when learning Chinese characters, one can assume that characters that correctly match phonetic and orthographic patterns are easier to absorb(Ellis & Beaton, 1993). Moreover, numerous studies have offered well developed strategies for learning Chinese characters by using radicals or components. These studies have briefly indicated that the internal component structure of a Chinese character helps learners to clearly remember that character(Taft & Chung, 1999). Since components are the unit of characters and because they can consist of many distinct characters, even the characteristics of a component can provide its meaning, phonological and morphological, because characters have the meaning of their internal components. Consequently, learners are able to readily and rapidly recognize and write characters by using components. For example, 木(pin-yin: mu4, meaning: tree) is a component, and two 木(tree) can compose 林(pin-yin: lin2, meaning: wood) and three 木(tree) can constitute 森(pin-yin: sen1, meaning: forest). Obviously, we are able to readily figure out their meanings due to such processes. Hence, most cognitive strategies frequently use components to teach students of Chinese how to read characters(Shen, 2004). In addition, many studies have extensively exploited strategies related to character frequency because humans are sensitive to the frequencies of events in their daily life and remember these things correctly (Ellis, 2002). Therefore, character frequency can help learners memorize characters correctly and can also contribute to retention. However, only using this strategy to learn Chinese is not suitable. Most Chinese characters characterized by high frequency have intricate construction, such as 謝謝(meaning: thank you) and 對不起(meaning: sorry). Beside, frequency is also an important factor with respect to education. Previous studies have the impact of high frequency characters on learning Chinese characters for beginners(Wang, Hung, Chang, & Chen, 2008), and found the learners could learn approximately 700 high frequency characters with ease. Therefore, the present study presumes that most learners are able to absorb high frequency characters effectively.

A recent study used character network construction to establish an efficient strategy for learning Chinese characters(Yan, Fan, Di, Havlin, & Wu, 2013). This strategy exploits the network of Chinese characters in a hierarchical structure and the weight of the network nodes to develop a Chinese character metric called distributed node weight (DNW). However, the character network of the DNW strategy merely considers associations between characters and character frequency. Actually, a component can be made up of many distinctive characters, but most of these are not high frequency characters. Reviewing the DNW strategy, it is suggested that characters having high frequency that are also composed of many components that constitute high frequency characters should be learned first.

In this study, we have tightly integrated the component and character frequency characteristics in order to develop an ap-

proach for creating a character order for learning traditional Chinese. Different from the above studies, the present approach has effectively exploited the components that compose high frequency characters and the characters having high frequency and high frequency components in order to construct a metric for generating an optimal Chinese character learning order. The proposed approach indicates that if characters are ordered based on the metric, learners will be able to learn simple construction and more significant characters first. As a result, the approach is expected to provide an effective and systematic order for learning Chinese characters. That is, by simply learning a few components, students of Chinese can read more characters with high frequency. The present study has been designed to create a Chinese character learning order designed to enable learners to read character expeditiously and systematically.

Method

Material. In this study, our approach was constructed based on the data components and character frequency data for the collected Chinese characters. We retrieved character data for the format of character pairs and their components to generate component data from Beijing Normal University(Yan et al., 2013), and character frequency from the combined character frequency list of both classical and modern Chinese(Jun Da, 2005).

Component Data. Components are parts that comprise characters and are associated with Chinese character acquisition(Shen & Ke, 2007). They have not only their own definitions, but also their own unique pronunciation and orthography. For the purpose of this study, 3,910 traditional Chinese characters and 310 traditional Chinese components were retrieved. By mapping these characters into their components, we are capable of obtaining information about the associations between characters and components.

Character Frequency Data. Character frequency significantly affects the ability to read and identify characters because learners are sensitive to the frequencies of characters seen on a daily basis (Ellis, 2002). We gathered the frequencies of 3,910 characters from the combined character frequency list of classical and modern Chinese(Jun Da, 2005). From this list, we were able to receive information regarding token frequencies of characters and character frequency.

Metric. The metric is a simple construction considering character components and characters with token frequencies. The method for this metric is to pick components that can compose many characters and to calculate the scores of these components. Furthermore, component scores are regularly used by rigorously selecting characters providing high token frequencies and high component scores. Use of this metric can generate an effective learning character order, and it can help beginners to study characters possessing components with high token frequencies first. The procedure for the approach is calculating the Component Score(COS) and then computing the Character Score(CHS), accordingly obtaining the list of Chinese characters as sorted by Character Score(CHS) from high to low frequency.

Component Score. We developed a metric (eq.1) to calculate the Component Score(COS). Here i represents the component, j represents the character, m represents the largest number of components, which is decomposed by the most complicated character j, and having component i as part of its components, and the largest number is 9. k ranges from 0 (a character consists of a unique component and includes component i) to m(a character consists of m different components and includes component i), f_j represents the token frequency of character j which is at k level. The symbol n represents that all characters at level k consist of n components. An example of a character 非 (pin-yin: fei1, meaning: not) is given in Table 1. At level 1, one character 非, one component and then f_j is 214873, n is 1, k is 0, so the level 1 score is 5.33. At level 2, with nine characters and ten components, $\sum\log(f_j)$ is 32.745, n is 10, k is 1, so the level 2 score is 1.64. At level 3, with 2 characters and 5 components, $\sum\log(f_j)$ is 9.056, n is 5, k is 2, so the level 3 score is 0.264. At level 4, with 2 characters and 5 components, $\sum\log(f_j)$ is 5.185, n is 5, k is 3, so the level 4 score is 0.13. The total character score is 7.364. The purpose of the COS is to pick component i which can comprise high token frequencies of characters with a few components, thus lowering n and leading to a higher score. Moreover, the aim of parameter 2 is to decrease scores that are at a higher level, in that characters at higher levels represent very intricate structures. In this manner, a component with a higher score means that it can consist of high token frequencies and characters with a simple structure and few components.

$$COS(i) = \sum_{k=0}^{m} \frac{\sum\log(f_j)}{n} \times 2^{-k} \tag{1}$$

Character Score. The Character Score(CHS) metric (eq.2) contains the token frequency of character j, COS of components which constitute character j and the number of components. Here, u represents the number of components which comprise character j, and f_j represents the token frequency of character j.

$$CHS(j) = f_j \times \frac{\sum COS(i)}{u} \tag{2}$$

The purpose of the Character Score (CHS) is to pick a character with high token frequency which is composed of components with a high COS score. By calculating the character score (CHS), we are able to clearly determine which characters have high token frequencies and include components with a structure consisting of many characters that also have high token frequencies and simple construction. Therefore, these two metrics, COS and CHS, generate a valid, comprehensive and rapid method for learning Chinese characters.

Table 1: Characters are composed of 非, TF means token frequency.

Characters	TF	*log*(TF)	Components	Level
非	214873	5.332	非	1
誹	1781	3.251	言,非	2
罪	90904	4.959	四,非	2
菲	16263	4.211	艸,非	2
排	63402	4.802	手,非	2
徘	3060	3.486	彳,非	2
啡	5143	3.711	口,非	2
輩	4	0.602	車,非	2
悲	32183	4.508	心,非	2
韭	1647	3.217	一,非	2
靠	46625	4.669	牛,口,非	3
鐕	4	0.602	人,弋,非	3
纖	16	1.204	糸,人,弋,非	4
殲	9562	3.981	歹,人,弋,非	4

Results

Comparing the CHS method with other approaches from a more specific perspective, we evaluated the three learning methods, the CHS method, DNW(Yan et al., 2013) and their learning method of character frequency by using the component cost of learning characters and accumulating character usage frequency. From the component cost of learning characters perspective as shown in Figure 1, we directly observed that the CHS learning order can cost less for components and will also result in the ability to learn more characters. Although the DNW learning order exhibited satisfactory performance, it appears to cost more components than CHS to learn more characters. Additionally, the learning order of usage frequency only considers usage frequency, but not the cost of components. For instance, when learners learn 300 characters, they need to learn 149 components using DNW, but need to learn only 129 components using CHS.

Table 2: Top 10 characters in CHS, DNW, Freqency

Rank	CHS	DNW	Freqency
1	的	人	的
2	人	一	一
3	一	口	不
4	以	的	是
5	來	日	了
6	有	白	人
7	是	土	在
8	他	又	有
9	個	言	我
10	和	勺	他

According to the accumulating character usage frequency perspective as shown in Figure 2, we can see that the CHS

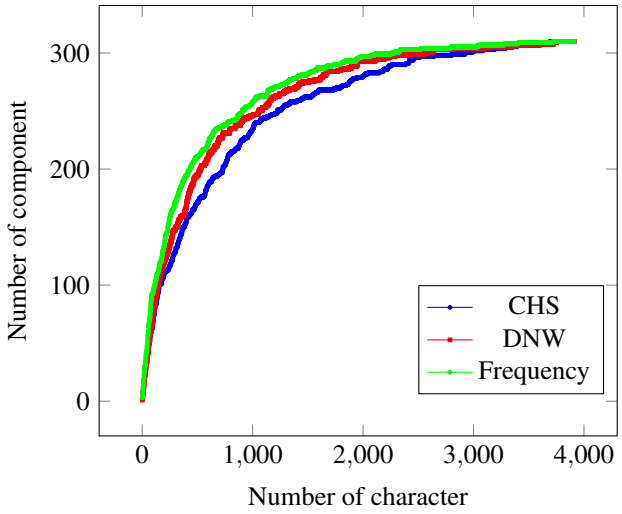

Figure 1: Learning efficiency comparison for different learning methods: character score (CHS), distributed node weight (DNW), and the usage frequency of characters (Frequency). This figure illustrates the component cost for learning characters.

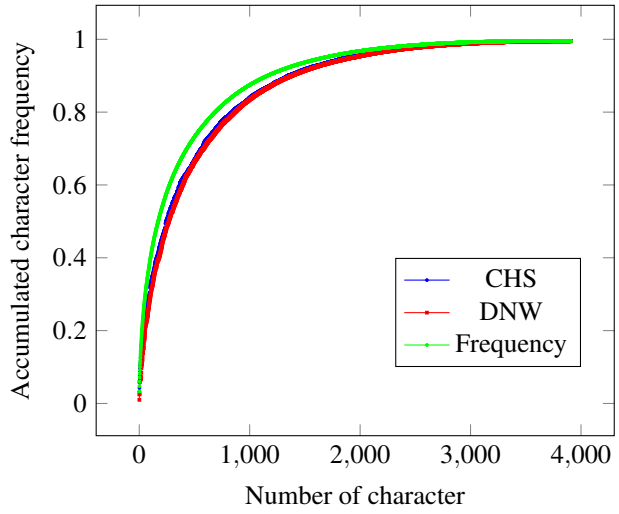

Figure 2: Evaluated three methods (CHS, DNW, Frequency) with accumulated character frequency.

learning order has almost the same excellent performance as that of the DNW strategy, and we can also see that it is even better than DNW. The architecture of the DNW strategy was constructed based on a hierarchical node weighted network, and the DNW score is calculated by the usage frequencies that are the node weights. Therefore, a component including many characters at higher layers and at lower layers may lead to a higher calculated score using the DNW strategy. As an illustrative result, two methods have very close curves. Doubtless, the method of usage frequency has better performance in this measure. Actually, for learning characters made up

of 80% accumulating character usage frequency, learners use 216 components to learn 841 characters in order to achieve 80% accumulating character usage frequency using CHS, but they need to learn 238 components to learn 867 characters to accomplish this task using the DNW strategy.

In addition, we compared three methods by using the results for the top 10 characters, as shown in Table 2. For the CHS method, we can clearly see that there are 5 characters 人(pinyin: ren2, meaning: people), 以(pin-yin: yi3, meaning: use), 來(pin-yin: lai2, meaning: come), 個(pinyin: ge4, meaning: a measure word), 他(pin-yin: ta1, meaning: he) which are composed using人. Moreover,的(pin-yin: de, meaning: of) and 是(pin-yin: shi4, meaning: is) are made up of 日(pin-yin: ri4, meaning: sun). Furthermore, 口(pinyin: kou3, meaning: mouth) composed 個 and 和(pin-yin: he2, meaning: and). In addition, these 10 characters have high usage frequency. However, there are two groups in the DNW strategy. First, 日, 的 and 白(pin-yin: bai2, meaning: white) have the same component, 日. Second, 口 and 言(pinyin: yan2, meaning: speak) have the same component, 口, yet other characters do not group with common components. In the case of the characters in the usage frequency learning order, although these characters exhibit high frequency, they have almost no association with each other. These three different measures clearly and graphically demonstrated that CHS is an effective and systematic learning method.

Discussion and Conclusion

In this study, we developed an effective learning method and demonstrated the advantages of thoroughly integrating component and character frequency, respectively. Advantages of the use of components is that they not only present their meaning but also their morphological information. This can help students learn Chinese characters easily and in greater number. Examples of this would be our top 10 characters 人 and 來; 來 is composed of 人 and 木; the morphological meaning is that two people are walking to a tree, so the meaning of 來 is 'to come'. Also, a component that can be used to compose many characters having high frequency and simple construction should be learned first. Regarding the merits of character frequency, literature has been found that character frequency can provide value with regard to utilization. In the case of learners, learning high frequency characters can help them read characters that are used in the daily life quickly. The present approach of the study also suggests that learning a few components and then reading more high frequency characters can generate an effective, systematic and rapid method for learning traditional Chinese characters. Furthermore, for the purpose of examining the effectiveness of the approach proposed in the present study, we are planning a series of teaching experiments to be conducted in real classrooms for enhancing Chinese learners' reading comprehension in the near future. Nowadays, textbooks for learning traditional Chinese characters are edited by various publishing companies in Taiwan, but the learning content of these textbooks is disorganized and difficult to use due to non-associative arrangements of characters in articles intended for students of Chinese. Consequently, most students feel confused with regard to recognizing most Chinese characters, which leads to inefficient learning. We anticipate that the approach presented in the study can be applied to improve the quality of Chinese textbooks or learning materials for the purpose of optimal learning in a systematic and meaningful manner. Also the approach may be used to design on-line materials and to customize or personalize Chinese character learning lists for learners who already possessed certain limited knowledge of Chinese characters and expect to extend their ability to read Chinese characters.

Acknowledgments

This research is partially supported by the 'Aim for the Top University Project' of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan,R.O.C. and the 'International Research-Intensive Center of Excellence Program' of NTNU and National Science Council, Taiwan, R.O.C. under Grant no. NSC 103-2911-I-003-301.

References

Ellis, N. (2002). Frequency effects in language processing. *Studies in second language acquisition*, *24*(2), 143–188.

Ellis, N., & Beaton, A. (1993). Factors affecting the learning of foreign language vocabulary: Imagery keyword mediators and phonological short-term memory. *The Quarterly Journal of Experimental Psychology*, *46*(3), 533–558.

Everson, M. E. (1998). Word recognition among learners of chinese as a foreign language: Investigating the relationship between naming and knowing. *The Modern Language Journal*, *82*(2), 194–204.

Hayes, E. B. (1988). Encoding strategies used by native and non-native readers of chinese mandarin. *The Modern Language Journal*, *72*(2), 188–195.

Ke, C. (1998). Effects of strategies on the learning of chinese characters among foreign language students. *Journal-Chinese Language Teachers Assocation*, *33*, 93–111.

Lin, Y. (2000). Vocabulary acquisition and learning chinese as a foreign language. *Journal-Chinese Language Teachers Assocation*, *35*(1), 85–108.

Shen, H. H. (2004). Level of cognitive processing: Effects on character learning among non-native learners of chinese as a foreign language. *Language and Education*, *18*(2), 167–182.

Shen, H. H., & Ke, C. (2007). Radical awareness and word acquisition among nonnative learners of chinese. *The Modern Language Journal*, *91*(1), 97–111.

Taft, M., & Chung, K. (1999). Using radicals in teaching chinese characters to second language learners. *Psychologia*, *42*(4), 243–251.

Wang, C. C., Hung, L. Y., Chang, Y. W., & Chen, H. F. (2008). Number of characters school students know from

grade 1 to 9. *Journal of Educational Psychology*, *39*(4), 555–568.

Yan, X.-Y., Fan, Y., Di, Z., Havlin, S., & Wu, J. (2013). Efficient learning strategy of chinese characters based on network approach. *ArXiv Preprint:1303.1599*.